Mαthematical Mαrketing

Charles F. Hofacker

New South
Network Services

Mathematical Marketing Version NSNS.2007.08.26

6) **Attribution Rights.** You must retain, in the Source Code of any Derivative Works that You create, all copyright, patent or trademark notices from the Source Code of the Original Work, as well as any notices of licensing and any descriptive text identified therein as an "Attribution Notice." You must cause the Source Code for any Derivative Works that You create to carry a prominent Attribution Notice reasonably calculated to inform recipients that You have modified the Original Work.

7) **Warranty and Disclaimer of Warranty.** Licensor warrants that the copyright in and to the Original Work is owned by the Licensor or that the Original Work is distributed by Licensor under a valid current license from the copyright owner. Except as expressly stated in the immediately proceeding sentence, the Original Work is provided under this License on an "AS IS" BASIS and WITHOUT WARRANTY, either express or implied, including, without limitation, the warranties of NON-INFRINGEMENT, MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY OF THE ORIGINAL WORK IS WITH YOU. This DISCLAIMER OF WARRANTY constitutes an essential part of this License. No license to Original Work is granted hereunder except under this disclaimer.

8) **Limitation of Liability.** Under no circumstances and under no legal theory, whether in tort (including negligence), contract, or otherwise, shall the Licensor be liable to any person for any direct, indirect, special, incidental, or consequential damages of any character arising as a result of this License or the use of the Original Work including, without limitation, damages for loss of goodwill, work stoppage, computer failure or malfunction, or any and all other commercial damages or losses. This limitation of liability shall not apply to liability for death or personal injury resulting from Licensor's negligence to the extent applicable law prohibits such limitation. Some jurisdictions do not allow the exclusion or limitation of incidental or consequential damages, so this exclusion and limitation may not apply to You.

9) **Acceptance and Termination.** If You distribute copies of the Original Work or a Derivative Work, You must make a reasonable effort under the circumstances to obtain the express and volitional assent of recipients to the terms of this License. Nothing else but this License (or another written agreement between Licensor and You) grants You permission to create Derivative Works based upon the Original Work or to exercise any of the rights granted in Sections 1 herein, and any attempt to do so except under the terms of this License (or another written agreement between Licensor and You) is expressly prohibited by U.S. copyright law, the equivalent laws of other countries, and by international treaty. Therefore, by exercising any of the rights granted to You in Sections 1 herein, You indicate Your acceptance of this License and all of its terms and conditions. This License shall terminate immediately and you may no longer exercise any of the rights granted to You by this License upon Your failure to honor the proviso in Section 1(c) herein.

10) **Mutual Termination for Patent Action.** This License shall terminate automatically and You may no longer exercise any of the rights granted to You by this License if You file a lawsuit in any court alleging that any OSI Certified open source software that is licensed under any license containing this "Mutual Termination for Patent Action" clause infringes any patent claims that are essential to use that software.

11) **Jurisdiction, Venue and Governing Law.** Any action or suit relating to this License may be brought only in the courts of a jurisdiction wherein the Licensor resides or in which Licensor conducts its primary business, and under the laws of that jurisdiction excluding its conflict-of-law provisions. The application of the United Nations Convention on Contracts for the International Sale of Goods is expressly excluded. Any use of the Original Work outside the scope of this License or after its termination shall be subject to the requirements and penalties of the U.S.

Copyright Act, 17 U.S.C. § 101 et seq., the equivalent laws of other countries, and international treaty. This section shall survive the termination of this License.

12) **Attorneys Fees.** In any action to enforce the terms of this License or seeking damages relating thereto, the prevailing party shall be entitled to recover its costs and expenses, including, without limitation, reasonable attorneys' fees and costs incurred in connection with such action, including any appeal of such action. This section shall survive the termination of this License.

13) **Miscellaneous.** This License represents the complete agreement concerning the subject matter hereof. If any provision of this License is held to be unenforceable, such provision shall be reformed only to the extent necessary to make it enforceable.

14) **Definition of "You" in This License.** "You" throughout this License, whether in upper or lower case, means an individual or a legal entity exercising rights under, and complying with all of the terms of, this License. For legal entities, "You" includes any entity that controls, is controlled by, or is under common control with you. For purposes of this definition, "control" means (i) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (ii) ownership of fifty percent (50%) or more of the outstanding shares, or (iii) beneficial ownership of such entity.

15) **Right to Use.** You may use the Original Work in all ways not otherwise restricted or conditioned by this License or by law, and Licensor promises not to interfere with or be responsible for such uses by You.

License Addendum. Names of contributors must be prominently featured in any use of this or derivative works.

To Linda Susan Vaughn,
La Dolcissima

Table of Contents

Charles Hofacker
College of Business
Florida State University
Tallahassee, FL 32306-1110
+1 850 644 7864
chofack@cob.fsu.edu

Preface to Version NSNS.2007.08.26

Over the years I have taught two different doctoral level quantitative methods courses in Marketing at Florida State University. The nature of these two courses changed somewhat from time to time depending on the exact offering of courses by other departments such as the Statistics Department and the Educational Measurement Department, as well as the needs of each particular cohort of students. The upshot was that I eventually ended up with a set of notes covering a good number of the most important mathematical tools used in marketing; perhaps enough material for two and a half semesters worth of classes. This manuscript is the result of transcribing most of those notes into book form.

There are generally two sorts of textbooks available for a Marketing Ph D quantitative methods course. The first sort is generally user-friendly, but with little or no actual mathematics. This kind of book generally treats each presented statistical procedure as a black box. The student is taught what type of stuff to put into the box, what type of stuff is likely to come out of the box, and how to write it up. The second sort of book is journal article quality "raw material"; very technical and prone to assuming that the student has had numerous recent courses in mathematical statistics. Diving into the second type of text can be daunting. What's more, since the tradition in mathematical marketing borrows from psychology, economics and management science, no single technical book tends to cover everything and a thorough coverage of the field requires that the student jump from one set of notational conventions to another. Yet if you pick up any issue of *Marketing Science* or the *Journal of Marketing Research*, the author is liable to assume you are familiar with anything from repeated measures ANOVA to efficient parameter estimation in econometric models. This book is my attempt to create a text designed to allow students to follow the mathematical reasoning used in *Marketing Science*, *Journal of Marketing Research* and other quantitatively oriented journals. As such, it does not hide the mathematics, but it does not assume that the student already has an extensive background in mathematics.

One unusual feature of this book is the license agreement, which is inspired by open source software. Although software development is a different kind of intellectual activity from book writing, the two are perhaps similar enough that methods that have worked in the former domain will also work in the latter. We shall see, but one way or the other I believe it will be interesting experiment.

It occurs to me that a fairly esoteric course like PhD quantitative methods is an ideal laboratory for an open source book. I envision a community of adopters around the world adding to this version, improving it, adding teaching notes, exercises, data sets, more references, improved slides, even new chapters. There are a variety of chapters that might be useful: Bayesian techniques and Proportional Hazards models are two examples. In addition, there are many topics that could be added to individual chapters.

Prerequisite Structure

Chapter	Chapter Prereq	Calculus	Distributions	Eigen structure	Esimation, ML
1					
2	1				
3	1				
4	1				
5	1, 2	3.1, 3.2, 3.3	4.1, 4.2		
6	5				
7	6				
8	7	3.4		3.5 - 3.8	
9	5		4.3		3.9, 3.10
10	9				
11	9			3.5 - 3.8	
12	5, 6.8				3.9, 3.10
13	12.1 - 12.4				
14	7				3.9
15	5				3.9, 3.10
16	5				
17	6			3.5 - 3.8	
18	5				

Much of what Ph D students learn about substantive issues in marketing can be studied in any order. It is, however, in the nature of technical topics that the sequence of study must be more carefully arranged. With that in mind, the above is a suggested prerequisite sequence for each chapter. The prerequisites are repeated in the beginning of each chapter.

You will note that the Chapter on OLS, Chapter 5, is needed for almost everything else in the book. One could follow a sequence of Chapters 1 and 2, Sections 3.1, 3.2, 3.3, 4.1 and 4.2, Chapter 5 and then do almost any subset of chapters as a specialized course. For example, a course in LISREL modeling would go from Chapter 5 to Sections 3.9, 3.10 and 4.3 into Chapters 9 and 10. A course oriented more towards choice modeling would start out the same up to Chapter 5 to but then cover Sections 3.9, 3.10 and 6.8 followed by Chapters 12 and 13. A course oriented more towards econometrics might start with the sequence leading up to Chapter 5 and then do Chapter 6, Sections 3.5 - 3.8 on eigenstructrure and 16, 17and 18. Many other possibilities exist.

Acknowledgments

I now officially dedicate this version of the book to the following: my wife and some other people and stuff.

The list of other people begins with statistics professors I have had, whose opinions and ways of explanation are mixed into the chapters that follow: Don Butler, Sam Pineau, Tom Wickens, Eric Holman, Art Woodward, Peter Bentler, Andrew Comrey, Bengt Muthèn and Ed Leamer. It extends to Sandeep Krisnamurthy for bringing open source to the attention of academic marketing.

E addesso, dei ringraziamenti in Italiano. Innanzi tutto a Sandro Castaldo e l'Università Bocconi. Alla Roma per vincere lo scudetto nel 2001. Per la cucina Italiana. Per quella strada a lungo del crinale vicino a Fiesole. Alla gente che stanno ripristinando le colonne del Tempio di Adriano (oggi la Borsa di Roma). Per Salvatore Zappalà perchè si fermò ad aiutarmi un pomeriggio autunale nel 2001. A Luca Debenedittis per essere il più bravo insegnante d'Italiano che abbia mai conosciuto. A Rafaele Desantis per aspettarmi dopo le discese. Per il suono che fanno la ali degli uccelli. Per la luna vista di giorno. La colonna Sonora del film La Finestra Di Fronte. Alle bici in titanio.

Spero di non confondere il mondo troppo.

Charles F Hofacker
August 28, 2007

Section I: Mathematical Fundamentals

Chapter 1: Linear Algebra

1.1 *Introduction to Vector and Matrix Notation*

Much of the mathematical reasoning in all of the sciences that pertain to humans is linear in nature, and linear equations can be greatly condensed by matrix notation and matrix algebra. In fact, were it not for matrix notation, some equations could fill entire pages and defy our understanding. The first step in creating easier-to-grasp linear equations is to define the *vector*. A vector is defined as an ordered set of numbers. Vectors are classified as either *row vectors* or *column vectors*. Note that a vector with one element is called a *scalar*. Here are two examples. The vector **a** is a column vector with m elements,

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_m \end{bmatrix},$$

and the vector **b** is a row vector with q elements:

$$\mathbf{b} = [b_1 b_2 \cdots b_q].$$

You should notice that in this text vectors are generally represented with lower case letters in bold.

There are a variety of ways that we can operate on vectors, but one of the simplest is the *transpose operator*, which, when applied to a vector, turns a row into a column and vice versa. For example,

$$\mathbf{a}' = [a_1 \quad a_2 \quad \cdots \quad a_m].$$

By convention, in this book, a vector with a transpose will generally imply that we are dealing with a row. The implication is that by default, all vectors are columns.

A *matrix* is defined as a collection of vectors, for example

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

$$= \{x_{ij}\}.$$

In this text, matrices are typically represented with an upper case bold letter.

The square brackets are used to list all of the elements of a matrix while the curly brackets are sometimes used to show a typical element of the matrix and thereby symbolize the entire matrix in that manner. Note that the first subscript of **X** indexes the row, while the second indexes columns. Matrices are characterized by their *order*, that is to say, the number of rows and columns that they have. The above matrix **X** is of order n by m, sometimes written $n \cdot m$. From time to time we may see a matrix like **X** written with its order like so: ${}_n\mathbf{X}_m$. It is semantically appropriate to say that a row vector is a matrix of but one row, and a column vector is a matrix of one column. Of course, a scalar can be thought of as the special case when we have a 1 by 1 matrix.

At times it will prove useful to keep track of the individual vectors that comprise a matrix. Suppose, for example that we defined each of the rows of **X** as

$$\begin{aligned}
\mathbf{x}'_{1\cdot} &= [x_{11} \quad x_{12} \quad \cdots \quad x_{1m}] \\
\mathbf{x}'_{2\cdot} &= [x_{21} \quad x_{22} \quad \cdots \quad x_{2m}] \\
\cdots &= \quad \cdots \\
\mathbf{x}'_{n\cdot} &= [x_{n1} \quad x_{n2} \quad \cdots \quad x_{nm}]
\end{aligned}$$

and then defined each column of **X**:

$$\mathbf{x}_{\cdot 1} = \begin{bmatrix} x_{11} \\ x_{21} \\ \cdots \\ x_{n1} \end{bmatrix}, \mathbf{x}_{\cdot 2} = \begin{bmatrix} x_{12} \\ x_{22} \\ \cdots \\ x_{n2} \end{bmatrix}, \cdots, \mathbf{x}_{\cdot m} = \begin{bmatrix} x_{1m} \\ x_{2m} \\ \cdots \\ x_{nm} \end{bmatrix} \tag{1.1}$$

so that **X** could be represented as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}'_{1\cdot} \\ \mathbf{x}'_{2\cdot} \\ \cdots \\ \mathbf{x}'_{n\cdot} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{\cdot 1} & \mathbf{x}_{\cdot 2} & \cdots & \mathbf{x}_{\cdot m} \end{bmatrix}. \tag{1.2}$$

In this context, the dot is known as a *subscript reduction operator* since it allows us to aggregate over the subscript replaced by the dot. So for example, the dot in $\mathbf{x}'_{i\cdot}$ summarizes all of the columns in the ith row of **X**.

Every so often a matrix will have exactly as many rows as columns, in which case it is a *square matrix*. Many matrices of importance in statistics are in fact square.

1.2 *The First Steps Towards an Algebra for Matrices*

One of the first steps we need to make to create an algebra for matrices is to define equality. We now do so defining two matrices

$$\mathbf{A} = \mathbf{B} \text{ iff } a_{ij} = b_{ij} \text{ for all i, j.} \tag{1.3}$$

Every element of **A** and **B** needs to be identical. For this to be possible, obviously both **A** and **B** must have the same order!

Just as one can transpose a vector, a matrix can be transposed as well. *Matrix transposition* takes all rows into columns and vice versa. For example,

$$\begin{bmatrix} 3 & 2 \\ 4 & 1 \\ 4 & 5 \end{bmatrix}' = \begin{bmatrix} 3 & 4 & 4 \\ 2 & 1 & 5 \end{bmatrix}.$$

Bringing our old friend **X** back, we could say that

$$\mathbf{X}' = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{x}'_{\cdot 1} \\ \mathbf{x}'_{\cdot 2} \\ \cdots \\ \mathbf{x}'_{\cdot m} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{1\cdot} & \mathbf{x}_{2\cdot} & \cdots & \mathbf{x}_{n\cdot} \end{bmatrix}.$$

We might add that

$$\left(\mathbf{X}'\right)' = \mathbf{X}. \tag{1.4}$$

A square matrix **S** is called *symmetric* if

$$\mathbf{S} = \mathbf{S}'. \tag{1.5}$$

Of course, a scalar, being a 1 by 1 matrix, is always symmetric.

Now we are ready to define *matrix addition*. For two matrices **A** and **B** of the same order, their sum is defined as the addition of each corresponding element as in

$$\begin{gathered} \mathbf{C} = \mathbf{A} + \mathbf{B} \\ \{c_{ij}\} = a_{ij} + b_{ij}\ . \end{gathered} \tag{1.6}$$

That is to say, we take each element of **A** and **B** and add them to produce the corresponding element of the sum. Here it must be emphasized that matrix addition is only possible if the components are *conformable for addition*. In order to be conformable for addition, they must have the same number of rows and columns.

It is possible to multiply a scalar times a matrix. This is called, appropriately enough, *scalar multiplication*. If c is a scalar, we could have

$$c\mathbf{A} = \mathbf{B}.$$

For example we might have

$$c\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} = \begin{bmatrix} c \cdot a_{11} & c \cdot a_{12} \\ c \cdot a_{21} & c \cdot a_{22} \\ c \cdot a_{31} & c \cdot a_{32} \end{bmatrix}.$$

Assuming that c_1 and c_2 are scalars, we can outline some properties of scalar multiplication:

Associative: $c_1(c_2\,\mathbf{A}) = (c_1 c_2)\,\mathbf{A}$ (1.7)

Distributive: $(c_1 + c_2)\,\mathbf{A} = c_1\,\mathbf{A} + c_2\,\mathbf{A}$ (1.8)

Now that we have defined matrix addition and scalar multiplication, we can define *matrix subtraction* as

$$\mathbf{A} - \mathbf{B} = \mathbf{A} + (-1)\mathbf{B}\,.$$

There are a few special matrices that will be of use later that have particular names. For example, an n by m matrix filled with zeroes is called a *null matrix*,

$$_{n}\mathbf{0}_{m} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{1.9}$$

and an n · m matrix of ones is called a *unit matrix*:

$$_{n}\mathbf{1}_{m} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}. \tag{1.10}$$

We have already seen that a matrix that is equal to its transpose ($\mathbf{S} = \mathbf{S}'$) is referred to as symmetric. A *diagonal matrix*, such as **D**, is a special case of a symmetric matrix such that

$$\mathbf{D} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & d_{mm} \end{bmatrix}. \tag{1.11}$$

i. e. the matrix consists of zeroes in all of the *off-diagonal* positions. In contrast, the *diagonal* positions hold elements for which the subscripts are identical.

A special case of a diagonal matrix is called a *scalar matrix*, a typical example of which appears below:

$$\begin{bmatrix} c & 0 & \cdots & 0 \\ 0 & c & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & c \end{bmatrix}. \tag{1.12}$$

And finally, a special type of scalar matrix is called the *identity matrix*. As we will soon see, the identity matrix serves as the identity element of matrix multiplication. For now, note that we generally use the symbol **I** to refer to such a matrix:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Having defined the identity matrix, we can think of a scalar matrix as being expressible as c**I** where c is a scalar.

We can now define some properties of matrix addition.

Commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$ (1.13)

Associative: $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$ (1.14)

Identity: $\mathbf{A} + \mathbf{0} = \mathbf{A}$ (1.15)

Note that in the definitions above we have assumed that all matrices are conformable for addition.

At this point we are ever closer to having all of the tools we need to create an algebra with vectors and matrices. We are only missing a way to multiply vectors and matrices. We now turn to that task. Assume we have a 1 by m row vector, **a'**, and an m by 1 column vector, **b**. In that case, we can have

$$\mathbf{a}'\mathbf{b} = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_m \end{bmatrix}$$

$$= a_1 b_1 + a_2 b_2 + \cdots + a_m b_m \tag{1.16}$$

$$= \sum_{i=1}^{m} a_i b_i .$$

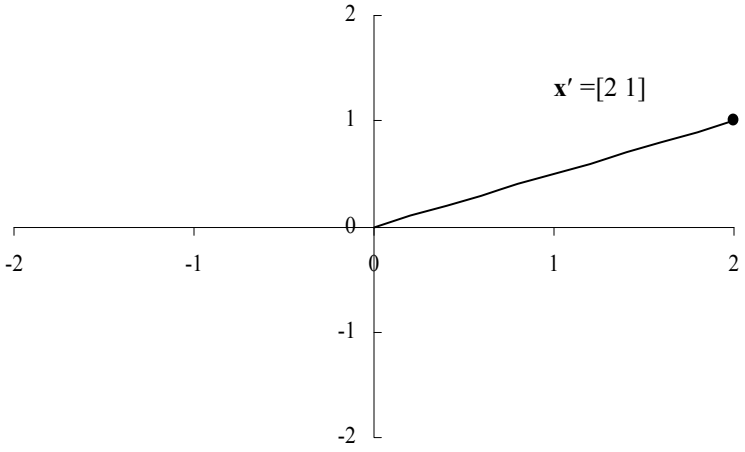
This operation is called taking a *linear combination*, but it is also known as the *scalar product*, the *inner product*, and the *dot product*. This is an extremely useful operation and a way to express a linear function with a very dense notation. For example, to sum the elements of a vector, we need only write

$$\mathbf{1}'\mathbf{a} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_m \end{bmatrix} = \sum_{i}^{n} a_i .$$

When a linear combination of two non-null vectors equals zero, we say that they are *orthogonal* as $\mathbf{x}'$ and **y** below:

$$\mathbf{x}'\mathbf{y} = 0 . \tag{1.17}$$

Geometrically, this is equivalent to saying that they are at right angles in a space with as many axes as there are elements in the vector. Assume for example that we have a 2 element vector. This can be interpreted as a point, or a vector with a terminus, in a plane (a two space). Consider the graph below:



Note that the vector $\mathbf{x}' = \begin{bmatrix} 2 & 1 \end{bmatrix}$ is represented in the graph. Can you picture an orthogonal vector? The *length of a vector* is given by $\sqrt{\mathbf{x}'\mathbf{x}} = \sum x_i^2$.

1.3 *Matrix Multiplication*

The main difference between scalar and matrix multiplication, a difference that can really throw off students, is that the commutative property does not apply in matrix multiplication. In general, $\mathbf{AB} \neq \mathbf{BA}$, but what's more, **BA** may not even be possible. We shall see why in a second. For now, note that in the product **AB**, where **A** is m · n and **B** is n · p, we would call **A** the *premultiplying* matrix and **B** the *postmultiplying* matrix. Each row of **A** is combined with each column of **B** in vector multiplication. An element of the product matrix, c_{ij}, is produced from the ith row of **A** and the jth column of **B**. In other words,

$$c_{ij} = \mathbf{a}'_{i\cdot}\mathbf{b}_{\cdot j} = \sum_{k}^{n} a_{ik} b_{kj} \tag{1.18}$$

The first thing we should note here is that the row vectors of **A** must be of the same order as the column vectors of **B**, in our case of order n. If not, **A** and **B** would not be *conformable for multiplication.* We could diagram things like this:

$$_{m}\mathbf{C}_{p} = {}_{m}\mathbf{A}_{n}\,{}_{n}\mathbf{B}_{p}$$

Here the new matrix **C** takes on the number of rows of **A** and the number of columns of **B**. The number of columns of **A** must match the number of rows of **B**. OK, now let's look at a quick example. Say

$$\mathbf{C} = \begin{bmatrix} -1 & 3 & 2 \\ 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 1 & 4 \\ 1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} (-1)2 + (3)1 + (2)1 & (-1)3 + (3)4 + (2)2 \\ (2)2 + (0)1 + (1)1 & (2)3 + (0)4 + (1)2 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 13 \\ 5 & 8 \end{bmatrix}.$$

A particular triple product, with a premultiplying row vector, a square matrix, and a postmultiplying column vector, is known as a *bilinear form*:

$$_1\mathbf{c}_1 = {}_1\mathbf{a}'_m \; {}_m\mathbf{B}_m \; {}_m\mathbf{d}_1 \tag{1.19}$$

A very important special case of the bilinear form is the *quadratic form*, in which the vectors **a** and **d** above are the same:

$$_1\mathbf{c}_1 = {}_1\mathbf{a}'_m \; {}_m\mathbf{B}_m \; {}_m\mathbf{a}_1 \tag{1.20}$$

The quadratic form is widely used because it represents the variance of a linear transformation.

For completion, we now present a *vector outer product*, in which an m by 1 vector, say **a**, is postmultiplied by a row vector, $\mathbf{b}'$:

$$_m\mathbf{C}_n = {}_m\mathbf{a}_1 \; {}_1\mathbf{b}'_n$$

$$= \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_m \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \tag{1.21}$$

$$= \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \cdots & \cdots & \cdots & \cdots \\ a_mb_1 & a_mb_2 & \cdots & a_mb_n \end{bmatrix}$$

The matrix **C** has m · n elements, but yet it was created from only m + n elements. Obviously, some elements in **C** must be redundant in some way. It is possible to have a matrix outer product as well - for example a 4 by 2 multiplied by a 2 by 4 would also be considered an outer product.

1.4 *Partitioned Matrices*

It is sometimes desirable to keep track of parts of matrices other than either individual rows or columns as we did with the dot subscript reduction operator. For example, lets say that the matrix **A**, which is m by p consists of two partitions, $\mathbf{A_1}$ which is m by p_1 and $\mathbf{A_2}$ which is m by p_2, where $p_1 + p_2 = p$. Thus both $\mathbf{A_1}$ and $\mathbf{A_2}$ have the same number of rows and when stacked horizontally, as they will be below, their columns

will add up to the number of columns of **A**. Then lets say we have the matrix **B**, which is of the order p by r, has two partitions $\mathbf{B_1}$ and $\mathbf{B_2}$ with $\mathbf{B_1}$ being p_1 by r and $\mathbf{B_2}$ being p_2 by r. The partitions $\mathbf{B}_1$ and $\mathbf{B}_2$ both have the same number of columns, namely r, so that when they are stacked vertically they match perfectly and their rows add up to the number of rows in **B**. In that case,

$$\mathbf{AB} = \left[\mathbf{A}_1 \,\vdots\, \mathbf{A}_2\right]\begin{bmatrix}\mathbf{B}_1 \\ \cdots \\ \mathbf{B}_2\end{bmatrix} = \mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_2\mathbf{B}_2 \tag{1.22}$$

We note that the product $\mathbf{A}_1\mathbf{B}_1$ and the product $\mathbf{A}_2\mathbf{B}_2$ are both conformable with order of m by r, precisely the order of **AB**.

1.5 *Cross-Product Matrices*

The cross product matrix is one of the most useful and common matrices in statistics. Assume we have a sample of n cases and that we have m variables. We define x_{ij} as the observation on consumer i (or store i or competitor i or segment i, etc.) with variable or measurement j. We can say that $\mathbf{x}'_{i\cdot}$ is a $1 \cdot m$ row vector that contains all of the measurements on case i and that $\mathbf{x}_{\cdot j}$ is the $n \cdot 1$ column vector containing all cases' measurements on variable j. The matrix **X** can then be expressed as a partitioned matrix, either as a series of row vectors, one per case, or as a series of columns, one per variable:

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}'_{1\cdot} \\ \mathbf{x}'_{2\cdot} \\ \cdots \\ \mathbf{x}'_{n\cdot}\end{bmatrix} = \begin{bmatrix}\mathbf{x}_{\cdot 1} & \mathbf{x}_{\cdot 2} & \cdots & \mathbf{x}_{\cdot m}\end{bmatrix} \tag{1.23}$$

What happens when we transpose X? All the rows become columns and all the columns become rows, as we can see below:

$$\mathbf{X}' = \begin{bmatrix}\mathbf{x}_{1\cdot} & \mathbf{x}_{2\cdot} & \cdots & \mathbf{x}_{n\cdot}\end{bmatrix} = \begin{bmatrix}\mathbf{x}'_{\cdot 1} \\ \mathbf{x}'_{\cdot 2} \\ \cdots \\ \mathbf{x}'_{\cdot m}\end{bmatrix} \tag{1.25}$$

In the right piece, a typical row would be $\mathbf{x}'_{\cdot j}$ which holds the data on variable j, but now in row format. This row has n columns. In the left piece, $\mathbf{x}_{i\cdot}$ is an m by 1 column holding all of the variables for case i. Now we have two possible ways to express the cross product, $\mathbf{X}'\mathbf{X}$. In the first approach, we show the columns of **X** which are now the rows of **X'**:

$$\mathbf{B} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{x}'_{\cdot 1} \\ \mathbf{x}'_{\cdot 2} \\ \cdots \\ \mathbf{x}'_{\cdot m} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{\cdot 1} & \mathbf{x}_{\cdot 2} & \cdots & \mathbf{x}_{\cdot m} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{x}'_{\cdot 1}\mathbf{x}_{\cdot 1} & \mathbf{x}'_{\cdot 1}\mathbf{x}_{\cdot 2} & \cdots & \mathbf{x}'_{\cdot 1}\mathbf{x}_{\cdot m} \\ \mathbf{x}'_{\cdot 2}\mathbf{x}_{\cdot 1} & \mathbf{x}'_{\cdot 2}\mathbf{x}_{\cdot 2} & \cdots & \mathbf{x}'_{\cdot 2}\mathbf{x}_{\cdot m} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{x}'_{\cdot m}\mathbf{x}_{\cdot 1} & \mathbf{x}'_{\cdot m}\mathbf{x}_{\cdot 2} & \cdots & \mathbf{x}'_{\cdot m}\mathbf{x}_{\cdot m} \end{bmatrix} \tag{1.26}$$

$$= \{\mathbf{x}'_{\cdot j}\mathbf{x}_{\cdot k}\} = \{b_{jk}\}$$

The above method of describing $\mathbf{X}'\mathbf{X}$ shows each element of the m by m matrix being created, one at a time. Each element of **X'X** is comprised of an inner product created by multiplying two n element vectors together. But now lets keep track of the rows of **X**, which are columns of **X'** which is just the opposite of what we did above. In this case, we have

$$\mathbf{B} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{x}_{1\cdot} & \mathbf{x}_{2\cdot} & \cdots & \mathbf{x}_{n\cdot} \end{bmatrix} \begin{bmatrix} \mathbf{x}'_{1\cdot} \\ \mathbf{x}'_{2\cdot} \\ \cdots \\ \mathbf{x}'_{n\cdot} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{x}_{1\cdot}\mathbf{x}'_{1\cdot} + \mathbf{x}_{2\cdot}\mathbf{x}'_{2\cdot} + \cdots + \mathbf{x}_{n\cdot}\mathbf{x}'_{n\cdot} \end{bmatrix} \tag{1.27}$$

$$= \sum_{i}^{n} \mathbf{x}_{i\cdot}\mathbf{x}'_{i\cdot}$$

and the m · m outer products, $\mathbf{x}_{i\cdot}\mathbf{x}'_{i\cdot}$, are summed across all n cases to build up the cross product matrix, **B**.

1.6 *Properties of Matrix Multiplication*

In what follows, c is a scalar, and **A**, **B**, **C**, **D**, **E** are matrices. Note that we are assuming in all instances below that the matrices are conformable for multiplication.

Commutative: $$c\mathbf{A} = \mathbf{A}c \tag{1.28}$$

Associative: $$\mathbf{A}(c\mathbf{B}) = (c\mathbf{A})\mathbf{B} = c(\mathbf{AB}) \tag{1.29}$$

Looking at the above associative property for scalar multiplication, we can say that a scalar can *pass through* a matrix or a parenthesis.

Associative: $$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) \tag{1.30}$$

Right Distributive: $$\mathbf{A}[\mathbf{B} + \mathbf{C}] = \mathbf{AB} + \mathbf{AC} \tag{1.31}$$

Left Distributive: $$[\mathbf{B} + \mathbf{C}]\mathbf{A} = \mathbf{BA} + \mathbf{CA} \tag{1.32}$$

It is important to note here that unlike scalar algebra, we must distinguish between the left and right distributive properties. Again, note that these properties only hold when the symbols represent matrices that are conformable to the operations used in the equation.

From Equation (1.31) and (1.32) we can deduce the following

$$(\mathbf{A} + \mathbf{B})'(\mathbf{A} + \mathbf{B}) = \mathbf{A}'\mathbf{A} + \mathbf{A}'\mathbf{B} + \mathbf{B}'\mathbf{A} + \mathbf{B}'\mathbf{B} \; . \tag{1.33}$$

To multiply out an equation like Equation (1.33), students sometimes remember the mnemonic FOIL = first, outside, inside, last, which gives the sequence of terms to be multiplied.

Transpose of a Product: $$[\mathbf{AB}]' = \mathbf{B}'\mathbf{A}' \tag{1.34}$$

In words, the above theorem states that the transpose of a product is the product of the transposes in reverse order. And finally, the *identity element of matrix multiplication* is the previously defined matrix **I**:

Identity: $$\mathbf{IA} = \mathbf{AI} = \mathbf{A} \tag{1.35}$$

1.7 *The Trace of a Square Matrix*

With a square matrix, from time to time we will have occasion to add up the diagonal elements, a sum known as *the trace of a matrix*. For example for the p by p matrix **S**, the trace of **S** is defined as

$$\text{Tr } \mathbf{S} = \sum_i s_{ii} \; . \tag{1.36}$$

A scalar is equal to its own trace. We can also say that with conformable matrices **A** and **B**, such that **AB** and **BA** both exist, it can be shown that the

$$\text{Tr}[\mathbf{AB}] = \text{Tr}[\mathbf{BA}] \; . \tag{1.37}$$

The theorem is applicable if both **A** and **B** are square, or if **A** is m · n and **B** is n · m.

1.8 *The Determinant of a Matrix*

While a square matrix of order m contains m^2 elements, one way to summarize all these numbers with one quantity is the *determinant*. The determinant has a key role in solving systems of linear equations. Consider the following two equations in two unknowns, x_1 and x_2.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= y_1 \\ a_{21}x_1 + a_{22}x_2 &= y_2 \end{aligned}$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$\mathbf{Ax} = \mathbf{y}$$

In a little while we will solve for the unknowns in the vector **x** using matrix notation. But for now, sticking with scalars, we can solve this using the following formula for x_1:

$$x_1 = \frac{y_1 a_{22} - y_2 a_{12}}{a_{11} a_{22} - a_{12} a_{21}} \tag{1.38}$$

The denominator of this formula is the determinant of the 2 by 2 matrix **A**. The determinant of a square matrix like **A** is usually written |**A**|. Being in the denominator, the system cannot be solved when the determinant is zero. Whether the determinant is zero depends on how much information is in **A**. If rows or columns are redundant, then |**A**| = 0 and there is no unique solution to the system of equations.

The determinant of a scalar is simply that scalar. Rules for determining the determinant of 3 by 3 and larger matrices can be found in Bock (1975, p. 62), Johnson and Wichern (2002, pp. 94-5) and other books on the linear model.

1.9 *The Inverse of a Matrix*

In scalar algebra we implicitly take the inverse to solve multiplication problems. If our system above was one equation in one unknown, it would be

$$ax = y$$

$$a^{-1}ax = a^{-1}y$$

$$1x = a^{-1}y$$

$$x = a^{-1}y$$

With a system of equations, the analog of $a^{-1} = 1/a$ is the inverse of a matrix, $\mathbf{A}^{-1}$.

$$\mathbf{Ax} = \mathbf{y}$$

$$\mathbf{A^{-1}Ax} = \mathbf{A^{-1}y}$$

$$\mathbf{Ix} = \mathbf{A^{-1}y}$$

To solve the system, you must find a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. You can only do so when $|\mathbf{A}| \neq 0$. In fact, we have now just officially defined the inverse of a matrix. The inverse of a square matrix **A** is simply that matrix, which when pre- or post-multiplied by **A**, yields the identity matrix, i. e. $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. One property of inverses is that the inverse of a product is equal to the product of the inverses in reverse order:

Inverse of a Product: $$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{1.39}$$

For proof, consider that

$$\mathbf{B}^{-1}\mathbf{A}^{-1}\mathbf{AB} = \mathbf{B}^{-1}(\mathbf{A}^{-1}\mathbf{A})\mathbf{B}$$

$$= \mathbf{B}^{-1}\mathbf{IB}$$

$$= \mathbf{I}$$

The inverse of the transpose of a square matrix is equal to the transpose of the inverse of that matrix. In other words, if $\mathbf{A}^{-1}$ is the inverse of **A**, then

$$\mathbf{A}^{-1\prime}\mathbf{A}' = \mathbf{I}\,. \tag{1.40}$$

1.10 *Kronecker Product*

The Kronecker Product with operator $\otimes$, is defined as

$$_{mp}\mathbf{C}_{nq} = {}_{m}\mathbf{A}_{n} \otimes {}_{p}\mathbf{B}_{q} = \{a_{ij}\mathbf{B}\}\,. \tag{1.41}$$

For example,

$$\begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}\mathbf{B} \\ \underline{\quad\quad} \\ a_{21}\mathbf{B} \end{bmatrix}.$$

References

R. Darrell Bock (1975) *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill.

Green, Paul E. (1976) *Mathematical Tools for Applied Multivariate Analysis.* New York: Academic

Johnson, Richard A. and Dean W. Wichern (2002) *Applied Multivariate Statistical Analysis, Fifth Edition*. Upper Saddle River, NJ: Prentice-Hall.

Chapter 2: Descriptive Statistics

Prerequisite: Chapter 1

2.1 *Review of Univariate Statistics*

The central tendency of a more or less symmetric distribution of a set of interval, or higher, scaled scores, is often summarized by the *arithmetic mean*, which is defined as

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i \,. \tag{2.1}$$

We can use the mean to create a *deviation score*,

$$d_i = x_i - \bar{x}, \tag{2.2}$$

so named because it quantifies the deviation of the score from the mean.

Deviation is often measured by squaring, since it equates negative and positive deviations. The sum of squared deviations, usually just called the *sum of squares*, is given by

$$\begin{aligned} a &= \sum_{i}^{n} (x_i - \bar{x})^2 \text{ or} \\ &= \sum_{i}^{n} d_i^2 \,. \end{aligned} \tag{2.3}$$

Another method of calculating the sum of squares was frequently used during the era that preceded computers when students would work with calculating machines,

$$a = \sum_{i}^{n} x_i^2 - \frac{\left(\sum_{i}^{n} x_i\right)^2}{n}. \tag{2.4}$$

Regardless whether one uses Equation (2.3) or Equation (2.4), the amount of deviation that exists around the mean in a set of scores can be averaged using the *standard deviation*, or its square, the *variance*. The variance is just

$$s^2 = \frac{1}{n-1} a$$

with s being the positive square root of s^2.

We can take the deviation scores and standardize them, creating, well; *standardized scores*:

$$z_i = \frac{x_i - \bar{x}}{s} = \frac{d_i}{s} \,. \tag{2.5}$$

Next, we define a very important concept, that of the *covariance* of two variables, in this case x and y. The covariance between x and y may be written Cov(x, y). We have

$$s_{xy} = \frac{1}{n-1}\left[\sum_i^n x_i y_{i-} \frac{\left(\sum_i^n x_i\right)\left(\sum_i^n y_i\right)}{n}\right]$$

$$= \frac{1}{n-1}\sum_i^n d_{x_i} d_{y_i} \,, \tag{2.6}$$

where the d_{x_i} are the deviation scores for the x variable, and the d_{y_i} are defined analogously for y. Note that with a little semantic gamesmanship, we can say that the variance is the covariance of a variable with itself. The product $d_{x_i} d_{y_i}$ is usually called a *cross product*.

2.2 *Matrix Expressions for Descriptive Statistics*

In this section we will return to our data matrix, **X**, with n observations and m variables,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

$$= \{x_{ij}\}.$$

We now define the *mean vector* $\bar{\mathbf{x}}$, such that

$$\bar{\mathbf{x}}' = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_m \end{bmatrix}$$

$$= \frac{1}{n}\,{}_1\mathbf{1}'_n\,\mathbf{X} \tag{2.7}$$

$$= \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}.$$

You might note that here we are beginning to see some of the advantages of matrix notation. For example, look at the second line of the above equation. The piece **1'X** expresses the operation of adding each of the columns of the **X** matrix and putting them in a row vector. How many more symbols would it take to express this using scalar notation using the summation operator Σ?

The mean vector can then be used to create the deviation score matrix, as below.

$$\mathbf{D} = \mathbf{X} - {}_{n}\mathbf{1}_{1}\overline{\mathbf{X}}'$$

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} = \mathbf{X} - \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} \begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \end{bmatrix} \tag{2.8}$$

$$= \mathbf{X} - \begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \\ \cdots & \cdots & \cdots & \cdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_m \end{bmatrix}$$

We would say of the **D** matrix that it is *column-centered*, as we have used the column means to center each column around zero.

Now lets reconsider the matrix **X'X**. This matrix is known as the *raw*, or *uncorrected*, *sum of squares and cross products matrix*. Often the latter part of this name is abbreviated *SSCP*. We will use the symbol **B** for the raw SSCP matrix:

$$\mathbf{B} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum x_{i1}x_{i1} & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{im} \\ \sum x_{i2}x_{i1} & \sum x_{i2}x_{i2} & \cdots & \sum x_{i2}x_{im} \\ \cdots & \cdots & \cdots & \cdots \\ \sum x_{im}x_{i1} & \sum x_{im}x_{i2} & \cdots & \sum x_{im}x_{im} \end{bmatrix}. \tag{2.9}$$

In addition, we have seen this matrix expressed row by row and column by column in Equations (1.26) and (1.27). The uncorrected SSCP matrix can be corrected for the mean of each variable in **X**. Of course, it is then called the *corrected SSCP* matrix at that point:

$$\mathbf{A} = \mathbf{D}'\mathbf{D} \tag{2.10}$$

$$\mathbf{A} = \mathbf{B} - \frac{1}{n}\begin{bmatrix} \left(\sum x_{i1}\right)^2 & \left(\sum x_{i1}\right)\left(\sum x_{i2}\right) & \cdots & \left(\sum x_{i1}\right)\left(\sum x_{im}\right) \\ \left(\sum x_{i2}\right)\left(\sum x_{i1}\right) & \left(\sum x_{i2}\right)^2 & \cdots & \left(\sum x_{i2}\right)\left(\sum x_{im}\right) \\ \cdots & \cdots & \cdots & \cdots \\ \left(\sum x_{im}\right)\left(\sum x_{i1}\right) & \left(\sum x_{im}\right)\left(\sum x_{i2}\right) & \cdots & \left(\sum x_{im}\right)^2 \end{bmatrix} \tag{2.11}$$

Note that Equation (2.10) is analogous to the classic statement of the sum of squares in Equation (2.3) while the second version in Equation (2.11) resembles the hand calculator formula found in Equation (2.4). The correction for the mean in the formula for the corrected SSCP matrix **A** can be expressed in a variety of other ways:

$$\mathbf{A} = \mathbf{B} - \frac{1}{n}\left({}_{m}\mathbf{X}'_{n\,n}\mathbf{1}_{1}\right)\left({}_{m}\mathbf{X}'_{n\,n}\mathbf{1}_{1}\right)'$$

$$= \mathbf{B} - \frac{1}{n}\left(\mathbf{X}'\mathbf{1}\right)\left(\mathbf{1}'\mathbf{X}\right)$$

$$= \mathbf{B} - \frac{1}{n}\mathbf{X}'(\mathbf{1}\mathbf{1}')\mathbf{X}$$

$$= \mathbf{B} - \frac{1}{n}\mathbf{X}'\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}\mathbf{X}$$

$$= \mathbf{B} - \bar{\mathbf{x}}(\mathbf{1}'\mathbf{X}).$$

Now, we come to one of the most important matrices in all of statistics, namely the *variance-covariance matrix*, often just called the *variance matrix*. It is created by multiplying the scalar 1/(n-1) times **A**, i. e.

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A} \tag{2.12}$$

This is the unbiased formula for **S**. From time to time we might have occasion to see the maximum likelihood formula which uses n instead of n - 1. The covariance matrix is a symmetric matrix, square, with as many rows (and columns) as there are variables. We can think of it as summarizing the relationships between the variables. As such, we must remember that the covariance between variable 1 and variable 2 is the same as the covariance between variable 2 and variable 1. The matrix **S** has $m(m+1)/2$ unique elements and $m(m-1)/2$ unique off-diagonal elements (of course there are m diagonal elements). We should also point out that $m(m-1)/2$ is the number of m things taken two at a time.

Previously we had mean-centered **X** using its column means to create the matrix **D** of deviation scores. Now we will further standardize our variables by creating Z scores. Define $\boldsymbol{\Delta}$ as the matrix consisting of diagonal elements of **S**. We define the function *Diag*(·) for this purpose:

$$\begin{aligned} \boldsymbol{\Delta} &= \text{Diag}(\mathbf{S}) \\ &= \begin{bmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & s_m^2 \end{bmatrix} \end{aligned} \tag{2.13}$$

Next, we need to invert the $\boldsymbol{\Delta}$ matrix, and take the square root of the diagonal elements. We can use the following notation in this case:

$$\boldsymbol{\Delta}^{-1/2} = \begin{bmatrix} 1/\sqrt{s_1^2} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{s_2^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1/\sqrt{s_m^2} \end{bmatrix} \tag{2.14}$$

The notion of taking the square root does not exactly generalize to matrices [see Equation (3.38)]. However, with a diagonal matrix, one can create a unique square root by taking the square roots of all the diagonal elements. With non-diagonal matrices there is no unique way to decompose a matrix into two identical components. In any case, the matrix $\boldsymbol{\Delta}^{-1/2}$ will now prove useful to us in creating Z scores. When you postmultiply a matrix by a
diagonal matrix, you operate on the columns of the premultiplying matrix. That is what we will do to **D**:

$$\mathbf{Z} = \mathbf{D}\boldsymbol{\Delta}^{-1/2}$$

$$= \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} \begin{bmatrix} 1/\sqrt{s_1^2} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{s_2^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1/\sqrt{s_m^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{d_{11}}{\sqrt{s_1^2}} & \frac{d_{12}}{\sqrt{s_2^2}} & \cdots & \frac{d_{1m}}{\sqrt{s_m^2}} \\ \frac{d_{21}}{\sqrt{s_1^2}} & \frac{d_{22}}{\sqrt{s_2^2}} & \cdots & \frac{d_{2m}}{\sqrt{s_m^2}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{d_{n1}}{\sqrt{s_1^2}} & \frac{d_{n2}}{\sqrt{s_2^2}} & \cdots & \frac{d_{nm}}{\sqrt{s_m^2}} \end{bmatrix} \tag{2.15}$$

which creates a matrix full of z scores. Note that just as postmultiplication by a diagonal matrix operates on the columns of the premultiplying matrix, premultiplying by a diagonal matrix operates on the rows of the postmultiplying matrix.

Now we are ready to create the matrix of correlations, **R**. The correlation matrix is the covariance matrix of the z scores,

$$\mathbf{R} = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z}$$

$$= \boldsymbol{\Delta}^{-1/2}\mathbf{S}\boldsymbol{\Delta}^{-1/2} \tag{2.16}$$

$$= \begin{bmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{bmatrix}$$

Since the correlation of x and y is the same as the correlation between y and x, **R**, like **S**, is a symmetric matrix. As such we will have occasion to write it like

$$\mathbf{R} = \begin{bmatrix} 1 & & & \\ r_{21} & 1 & - & \\ \cdots & \cdots & & \\ r_{m1} & r_{m2} & \cdots & 1 \end{bmatrix}$$

leaving off the upper triangular part. We can also do this for **S**.

Chapter 3: Calculus Tools

Prerequisite: Chapter 1

3.1 *Logarithms and Exponents*

By definition, the log function to the base b is the function such that $c = \log_b a$ if $b^c = a$. It is a very useful function in statistical reasoning, since it takes multiplication into addition as we will see in Equation (3.1). We generally use the notation *log* to imply a base of 10, i. e. $\log a = \log_{10} a$ and we use the notation *ln* to imply a base of Euler's e (2.7182812…), that is $\ln a = \log_e a$. Some rules of logarithms follow:

$$\ln ab = \ln a + \ln b \tag{3.1}$$

$$\ln \frac{a}{b} = \ln a - \ln b \tag{3.2}$$

$$\ln a^b = b \ln a \tag{3.3}$$

$$\ln e^a = a \tag{3.4}$$

$$\ln e = 1 \tag{3.5}$$

$$\ln 1 = 0 \tag{3.6}$$

As for exponents, we have the following rules:

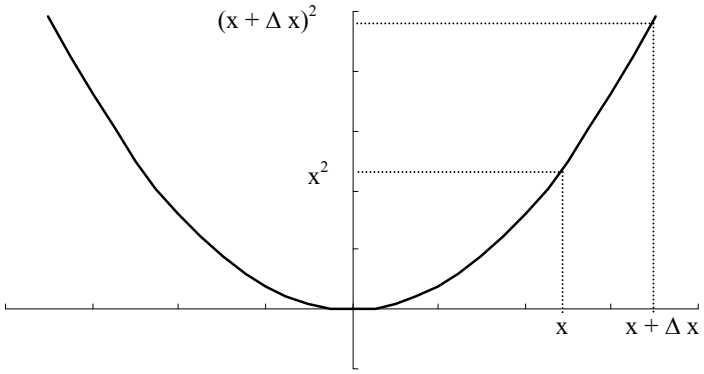
$$a^b \cdot a^c = a^{b+c} \tag{3.7}$$

$$a^{1/2} = \sqrt{a} \tag{3.8}$$

From a purely typographical point of view, it is sometimes more convenient to use the notation $\exp(a) = e^a$.

3.2 *A Review of Scalar Calculus*

Consider the problem of calculating the slope of $f(x) = x^2$. Unlike the equation for a line, the slope of the f(x) function changes depending on the value of x. However, at a small enough segment of the function, fitting a straight line would be reasonable. A picture of the situation is given below:



The slope is composed of the amount of change in the y axis (the rise) divided by the change in the x axis (the run). The fraction looks like

$$\text{slope} = \frac{(x+\Delta x)^2 - x^2}{(x+\Delta x) - x}$$

$$= \frac{x^2 + 2x\cdot\Delta x + (\Delta x)^2 - x^2}{\Delta x}$$

$$= 2x + \Delta x.$$

As we reduce Δx smaller and smaller, making a closer approximation to the slope, it converges on the value 2x. The derivative is the slope of a function at a point. There are two notations in common use. Thus we could write $dx^2/dx = 2x$ or $f'(x) = 2x$. In this book we will generally stick to the first way of writing the derivative.

More generally, for a function consisting of the power of a variable,

$$\frac{dx^m}{dx} = m\cdot x^{m-1}. \tag{3.9}$$

For the function f(x) = c where c is a constant, we would have

$$d(c)/dx = 0 \tag{3.10}$$

and for f(x) = cx,

$$d(cx)/dx = c. \tag{3.11}$$

The derivative of a sum is equal to the sum of the derivatives as we now see:

$$\frac{d[f(x)+g(x)]}{dx} = \frac{df(x)}{dx} + \frac{dg(x)}{dx}. \tag{3.12}$$

The exponential function to the base e has the interesting property that

$$\frac{de^x}{dx} = e^x \tag{3.13}$$

And we finish up this review by noting that for compound functions, such as g[f(x)], we can employ the *chain rule* which states that

$$\frac{dg[f(x)]}{dx} = \frac{dg[f(x)]}{df(x)} \cdot \frac{df(x)}{dx}. \tag{3.14}$$

Now, taking the chain rule into account we can state

$$\frac{de^{f(x)}}{dx} = e^{f(x)} \cdot \frac{df(x)}{dx}. \tag{3.15}$$

3.3 *The Scalar Function of a Vector*

We can now define the derivative of a function with respect to a whole vector of "independent" variables, $\partial f(\mathbf{x}')/\partial \mathbf{x}'$. Note that the function of the vector, f(**x**'), is a scalar. To begin, we will start with the constant function, that is, f(**x**') = c where c is a constant (scalar). The derivative of this function with respect to the row vector **x**' is itself a row vector with the same order as **x**'. That is because we need a derivative of the function with respect to each element of the vector. This vector derivative is called a *partial derivative* which means that as we take the derivative of the function with respect x_i, each of the other elements of **x** are treated as constants.

$$\begin{aligned}\frac{\partial c}{\partial \mathbf{x}'} &= \left[\frac{\partial c}{\partial x_1} \quad \frac{\partial c}{\partial x_2} \quad \cdots \quad \frac{\partial c}{\partial x_m}\right] \\ &= [0 \quad 0 \quad \cdots \quad 0].\end{aligned} \tag{3.16}$$

The derivative of the function with respect to x_i is 0, and i runs from 1 to m. Thus a vector derivative is created. For the linear combination **a'x** we have

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}'} = \left[\frac{\partial \mathbf{a}'\mathbf{x}}{\partial x_1} \quad \frac{\partial \mathbf{a}'\mathbf{x}}{\partial x_2} \quad \cdots \quad \frac{\partial \mathbf{a}'\mathbf{x}}{\partial x_m}\right]$$

$$= \left[\frac{\partial}{\partial x_1}[a_1x_1 + a_2x_2 + \cdots + a_mx_m] \quad \frac{\partial}{\partial x_2}[a_1x_1 + a_2x_2 + \cdots + a_mx_m] \quad \cdots \quad \frac{\partial}{\partial x_m}[a_1x_1 + a_2x_2 + \cdots + a_mx_m]\right]$$

$$= [a_1 \quad a_2 \quad \cdots \quad a_m] = \mathbf{a}'. \tag{3.17}$$

Another important result is the derivative of a quadratic form [Equation (1.20)]. In the equation below, we assume that **A** is a symmetric m · m matrix so that

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}'} = 2\mathbf{x}'\mathbf{A}' \tag{3.18}$$

with $\mathbf{A}' = \mathbf{A}$.

We now state the rule that the derivative of the transpose is equal to the transpose of the derivative, that is

$$\frac{\partial \mathrm{f}}{\partial \mathbf{x}'} = \left[\frac{\partial \mathrm{f}}{\partial \mathbf{x}}\right]' \text{ and}$$
$$\frac{\partial \mathrm{f}}{\partial \mathbf{x}} = \left[\frac{\partial \mathrm{f}}{\partial \mathbf{x}'}\right]' \tag{3.19}$$

From time to time we will need to use the *second order derivative* of a scalar function. It may be the case that the $\partial f/\partial x_i$ changes as a function of x_j, for example. The slope of the $\partial f/\partial x_i$ with respect to x_j, in other words the derivative of the derivative, is written as

$$\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right) = \frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

There are many uses of this second order derivative including nonlinear estimation [Section (3.9)], Maximum Likelihood parameter estimation [Section (3.10)], as well as determining whether, when the first order derivative is 0, we are at a maximum or minimum.

3.4 *Derivative of Multiple Functions with Respect to a Vector*

Suppose we have the linear system,

$$_{\mathrm{n}}\mathbf{y}_{1} = {}_{\mathrm{n}}\mathbf{A}_{\mathrm{m}\,\mathrm{m}}\,\mathbf{x}_{1}\,.$$

Now

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}'} = \frac{\partial}{\partial \mathbf{x}'}\begin{bmatrix}\mathbf{a}'_{1\cdot}\\ \mathbf{a}'_{2\cdot}\\ \cdots\\ \mathbf{a}'_{\mathrm{n}\cdot}\end{bmatrix}\mathbf{x}$$

$$= \begin{bmatrix}\frac{\partial \mathbf{a}'_{1\cdot}}{\partial \mathbf{x}'}\\ \frac{\partial \mathbf{a}'_{2\cdot}}{\partial \mathbf{x}'}\\ \cdots\\ \frac{\partial \mathbf{a}'_{\mathrm{n}\cdot}}{\partial \mathbf{x}'}\end{bmatrix} = \begin{bmatrix}\mathbf{a}'_{1\cdot}\\ \mathbf{a}'_{2\cdot}\\ \cdots\\ \mathbf{a}'_{\mathrm{n}\cdot}\end{bmatrix} = \mathbf{A}$$

To summarize,

$$\frac{\partial_n \mathbf{y}_1}{\partial_1 \mathbf{x}'_m} = \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}'} = {}_n\mathbf{A}_m \tag{3.20}$$

with each of the n rows of $\partial\mathbf{y}/\partial\mathbf{x}'$ a different function of $\mathbf{x}'$, $y_1, y_2, \ldots, y_n$ and each of the m columns of $\partial\mathbf{y}/\partial\mathbf{x}'$ referring to a different independent variable: $x_1, x_2, \ldots, x_m$. In other words, element i, j of $\partial\mathbf{y}/\partial\mathbf{x}'$ is of $\partial y_i/\partial x_j = a_{ij}$.

Of course given Equation (3.19),

$$\frac{\partial \mathbf{y}'}{\partial \mathbf{x}} = \left[\frac{\partial \mathbf{y}}{\partial \mathbf{x}'}\right]' = {}_m\mathbf{A}'_n \, .$$

3.5 *Eigen Structure for Symmetric Matrices*

Consider the p by 1 random vector $\mathbf{y}$, consisting of p observations taken on a randomly chosen case. The covariance matrix $\mathbf{S}$, which is the covariance matrix for p variables [that is, $V(\mathbf{y}) = \mathbf{S}$], is a symmetric matrix. I wish to create a linear combination

$$u = \mathbf{x}'\mathbf{y}, \tag{3.21}$$

such that q = V(u) is maximized. In this way I can replace the p elements of $\mathbf{y}$ with a single number that behaves as much as possible like the original p values. The problem can be written as

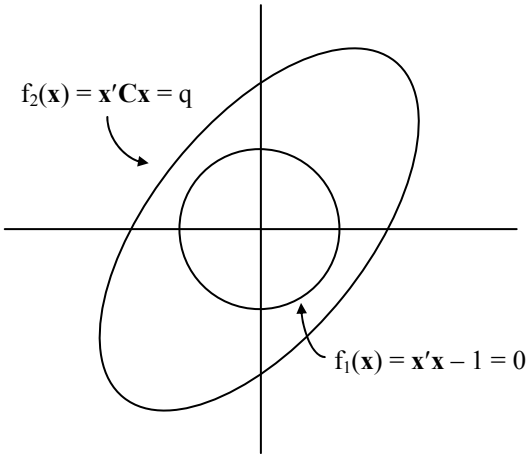
$$\frac{\text{Max } q = \mathbf{x}'\mathbf{S}\mathbf{x}}{\mathbf{x}} \, . \tag{3.22}$$

The notation here can be read, "Max q over all values of $\mathbf{x}$." One easy way to do this would be to pick $\mathbf{x} = [\infty \; \infty \; \cdots \; \infty]'$ but this would be completely uninteresting. Instead we will normalize $\mathbf{x}$, or constrain it so that we do not fall into a solution with a series of infinities. The reasoning behind how we maximize a function under constraints was introduced into mathematics by Lagrange. We can arbitrarily fix

$$\sum_j^p x_j^2 = \mathbf{x}'\mathbf{x} = 1 \text{ or set}$$

$$\mathbf{x}'\mathbf{x} - 1 = 0 \, . \tag{3.23}$$

This will allow us to focus on the pattern in the $\mathbf{x}$ vector that allows us to extract the maximum variance from $\mathbf{S}$. Geometrically, we can represent the situation as in the graph below:



Rather than trying to maximize $f_2(\mathbf{x})$, we will maximize $f_2(\mathbf{x})$ subject to $f_1(\mathbf{x})$. This is equivalent to maximizing $f_2(\mathbf{x})$ - $f_1(\mathbf{x})$, or finding the *principal axis* of the ellipse in the figure. The problem can now be written as

$$\underset{\mathbf{x}}{\text{Max}}\left[f_2(\mathbf{x}) - \lambda f_1(\mathbf{x})\right] = \underset{\mathbf{x}}{\text{Max}}\left[\mathbf{x}'\mathbf{S}\mathbf{x} - \lambda(\mathbf{x}'\mathbf{x} - 1)\right] \tag{3.24}$$

Note the sudden and mysterious appearance of the scalar λ! This creature is called a *Lagrange multiplier*. But where did it come from? Indeed. In defense of this equation, note that $f_2(\mathbf{x}) = \mathbf{x}'\mathbf{x}$ - 1 = 0. The scalar λ does not change the equation one iota, or better; one lambda. The function $f_2(\mathbf{x})$, as well as λ, are doomed to vanish. In short, λ is a mathematical throw-away. Using the rule for the derivative of a quadratic form [Equation (3.18)], along with some help from Equation (3.19), we see that

$$\frac{\partial \mathbf{x}'\mathbf{S}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{S}\mathbf{x}$$

and that

$$\frac{\partial \lambda(\mathbf{x}'\mathbf{x} - 1)}{\partial \mathbf{x}} = 2\lambda\mathbf{I}\mathbf{x}\,. \tag{3.25}$$

In that case, to maximize (3.24) we set

$$\frac{\partial}{\partial \mathbf{x}}\left[\mathbf{x}'\mathbf{S}\mathbf{x} - \lambda(\mathbf{x}'\mathbf{x} - 1)\right] = \mathbf{0}$$

$$2\mathbf{S}\mathbf{x} - 2\lambda\mathbf{I}\mathbf{x} = \mathbf{0}\,. \tag{3.26}$$

We can simplify further as below,

$$\mathbf{S}\mathbf{x} = \lambda\mathbf{x}, \tag{3.27}$$

where λ is now "acting like" **S**. Putting λ in Equation (3.24) is certainly legal since $\mathbf{x}'\mathbf{x}$ - 1 will be zero anyway. But what is it doing still hanging around in Equation (3.27)? We promised it would

go away, didn't we? What is λ anyway? Before we can answer this we need to return to Equation (3.27), were we had

$$\mathbf{Sx} = \lambda\mathbf{x}$$

which when premultiplied by $\mathbf{x}'$ leads to

$$\mathbf{x}'\mathbf{Sx} = \mathbf{x}'\lambda\mathbf{x}.$$

By the rules of scalar multiplication [in particular Equation (1.28)], and by the fact that $\mathbf{x}'\mathbf{x} = 1$ we have

$$\mathbf{x}'\lambda\mathbf{x} = \mathbf{x}'\mathbf{x}\lambda = \lambda$$

so that we can conclude

$$\mathbf{x}'\mathbf{Sx} = \lambda. \tag{3.28}$$

At this point the reader will recognize the formula for the variance of a linear combination, Equation 4.9. The value λ is called an *eigenvalue* of the matrix **S**. It is the maximum value, q, of the variance of $u = \mathbf{x}'\mathbf{y}$ which was our original motivation for this problem way back in Equation (3.21). The vector **x** chosen to maximize this variance is called an *eigenvector* of **S**.

3.6 *A Small Example Calculating the Eigenvalue and Eigenvector*

We will now return to Equation (3.26), which although it looked like

$$2\mathbf{Sx} - 2\lambda\mathbf{Ix} = \mathbf{0},$$

we can multiply by 1/2 to create

$$\mathbf{Sx} - \lambda\mathbf{Ix} = \mathbf{0} \quad \text{or}$$

$$[\mathbf{S} - \lambda\mathbf{I}]\,\mathbf{x} = \mathbf{0}. \tag{3.29}$$

Equation (3.29) can be solved trivially, as

$$[\mathbf{S} - \lambda\mathbf{I}]^{-1}[\mathbf{S} - \lambda\mathbf{I}]\,\mathbf{x} = [\mathbf{S} - \lambda\mathbf{I}]^{-1}\mathbf{0}$$

$$\mathbf{x} = \mathbf{0},$$

but such a solution would not be useful at all to us and in fact would not give us what we are looking for, namely, the linear combination $u = \mathbf{x}'\mathbf{y}$ such that $V(u) = \mathbf{x}'\mathbf{Sx}$ is as large as possible. To avoid falling into this trivial solution we must somehow pick λ such that

$$|\mathbf{S} - \lambda\mathbf{I}| = 0$$

which in turn implies that $[\mathbf{S} - \lambda\mathbf{I}]^{-1}$ does not exist (see Section 1.8). If $[\mathbf{S} - \lambda\mathbf{I}]^{-1}$ does not exist, we are not stuck with $\mathbf{x} = \mathbf{0}$, the trivial solution. Below, we can see how this works with a 2×2 example, lets say

$$\mathbf{S} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

In that case we have

$$|\mathbf{S} - \lambda\mathbf{I}| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0.$$

Recalling the equation for the determinant of a 2 × 2 matrix [from the denominator of Equation (1.38)], we have

$$(2 - \lambda)^2 - 1^2 = 0$$

which as a quadratic equation has two roots, i. e.

$$4 - 2\lambda - 2\lambda + \lambda^2 - 1 = 0$$

$$\lambda^2 - 4\lambda + 3 = 0$$

$$(\lambda - 3)(\lambda - 1) = 0$$

where the roots are $\lambda_1 = 3$ and $\lambda_2 = 1$. The first eigenvalue represents the maximum variance while the second represents the maximum variance that can be found after the first linear combination has been extracted. It is also true that the last eigenvalue represents the minimum amount of variance that can be extracted by a linear combination. We can now substitute λ_1 back into Equation (3.29) in order to solve for the first eigenvector. Calling this first eigenvector $\mathbf{x}_{\cdot 1}$, we have

$$[\mathbf{S} - \lambda\mathbf{I}]\,\mathbf{x}_{\cdot 1} = \mathbf{0}$$

$$\begin{bmatrix} 2-3 & 1 \\ 1 & 2-3 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

so that $-x_{11} + x_{21} = 0$ and $x_{11} - x_{21} = 0$. It is obvious then that $x_{11} = x_{21}$. Taken together with the restriction that $\mathbf{x}'\mathbf{x} = 1$ that we imposed in Equation (3.23), we have

$$\mathrm{x}_{\cdot 1} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

3.7 *Some Properties of Eigenstructure*

Before proceeding, it will be useful to take each of the eigenvectors, $\mathbf{x}_{\cdot 1}$, $\mathbf{x}_{\cdot 2}$, $\cdots$, $\mathbf{x}_{\cdot p}$, and place them as columns into the matrix $\mathbf{X}$. We also take the eigenvalues, λ_1, λ_2, $\cdots$, λ_p and put them on the diagonal of the matrix $\mathbf{L}$. The eigenvalues in $\mathbf{L}$ summarize a variety of properties of the original matrix $\mathbf{S}$. For example:

$$\mathrm{Tr}(\mathbf{S}) = \sum_{i}^{p} \lambda_i = \mathrm{tr}(\mathbf{\Lambda}) \tag{3.30}$$

$$|\mathbf{S}| = \prod_{i}^{p} \lambda_i \tag{3.31}$$

The *rank of a square matrix* $\mathbf{S}$ is given by the number of eigenvalues > 0. In other words, the rank of a square matrix is given by the number of non-null eigenvectors. We say that a square matrix is of *full rank* if one cannot pick a non-null vector $\mathbf{x}$ such that $\mathbf{x}'\mathbf{S}\mathbf{x} = 0$. We can see then from Equation (3.31) that if no eigenvalue is zero, the determinant, $|\mathbf{S}|$, will be non-zero and it will be possible to find $\mathbf{S}^{-1}$.

For each eigenvector-eigenvalue combination i, we have

$$\mathbf{S}\mathbf{x}_{\cdot i} = \mathbf{x}_{\cdot i}\lambda_i$$

so that if we premultiply by $\mathbf{x}'_{\cdot j}$ we have

$$\mathbf{x}'_{\cdot j}\mathbf{S}\mathbf{x} = \mathbf{x}'_{\cdot j}\mathbf{x}_{\cdot i}\boldsymbol{\lambda}_i \,.$$

Making the same argument for the eigenvalue and eigenvector j, we have

$$\mathbf{S}\mathbf{x}_{\cdot j} = \mathbf{x}_{\cdot j}\lambda_j$$

but now premultiplying by $\mathbf{x}'_{\cdot i}$

$$\mathbf{x}'_{\cdot i}\mathbf{S}\mathbf{x} = \mathbf{x}'_{\cdot i}\mathbf{x}_{\cdot j}\lambda_j \,.$$

Clearly it has to be the case that

$$\mathbf{x}'_{\cdot j}\mathbf{S}\mathbf{x}_{\cdot i} = \mathbf{x}'_{\cdot i}\mathbf{S}\mathbf{x}_{\cdot j}$$

in which case,

$$\mathbf{x}'_{\cdot j}\mathbf{x}_{\cdot i}\lambda_i = \mathbf{x}'_{\cdot i}\mathbf{x}_{\cdot j}\lambda_j \,.$$

But for that to happen, it must be true that

$$\mathbf{x}'_{\cdot j}\mathbf{x}_{\cdot i} = 0. \tag{3.32}$$

In other words, each pair of eigenvectors is orthogonal. When you add the standardizing constraint, Equation (3.23), we can say that

$$\mathbf{X}'\mathbf{X} = \mathbf{I} . \tag{3.33}$$

The **X** matrix, as can be seen above, acts as its own inverse. Any matrix **X** for which $\mathbf{X}'\mathbf{X} = \mathbf{X}\mathbf{X}' = \mathbf{I}$ is called *orthonormal*.

Here are some more properties of the eigenvalues and eigenvectors. From Equation (3.27) we can make the simultaneous statement about each eigenvalue-eigenvector below,

$$\mathbf{S}\mathbf{X} = \mathbf{X}\mathbf{L}. \tag{3.34}$$

Premultiplying by $\mathbf{X}'$ leads to

$$\mathbf{X}'\mathbf{S}\mathbf{X} = \mathbf{L} . \tag{3.35}$$

Or, starting again from Equation (3.34) but postmultiplying by $\mathbf{X}'$ this time leads to

$$\mathbf{S} = \mathbf{X}\mathbf{L}\mathbf{X}' . \tag{3.36}$$

$$= \mathbf{X}\mathbf{L}^{1/2}\,\mathbf{L}^{1/2}\mathbf{X}' \tag{3.37}$$

where the "square root" of the matrix **L** is clearly defined as $\{\lambda_i^{1/2}\}$, that is having the square root of each of the λ_i on the diagonal [c.f. Equation (2.14) and the discussion thereof]. Now if we define

$$\mathbf{B} = \mathbf{X}\mathbf{L}^{1/2}$$

We can say that

$$\mathbf{S} = \mathbf{B}\mathbf{B}' \tag{3.38}$$

which provides a "square root" like effect, even if the square root of a non-diagonal matrix cannot be uniquely defined. That this equation is not unique can be shown simply by defining the orthonormal matrix **J**, i. e. $\mathbf{J}'\mathbf{J} = \mathbf{J}\mathbf{J}' = \mathbf{I}$. Now if $\mathbf{B}^* = \mathbf{B}\mathbf{J}$ then

$$\mathbf{S} = \mathbf{B}^{*\prime}\,\mathbf{B}^* = \mathbf{B}\mathbf{J}\mathbf{J}'\mathbf{B} = \mathbf{B}\mathbf{B}'.$$

In factor analysis we seek a **B** matrix corresponding to a hypothesis about latent variables. In Cholesky factorization, we produce a lower triangular **B** matrix. In finding the eigenstructure of the **S** matrix, the columns of the **B** matrix produced in Equation 3.38) maximize the variance of the extracted components.

But the eigenstructure of **S** captures even more of the properties of **S**. For example, if $\mathbf{S}^{-1}$ exists,

$$\mathbf{S}^{-1} = \mathbf{X}\mathbf{L}^{-1}\mathbf{X}' . \tag{3.39}$$

In addition, if $\mathbf{A} = c\mathbf{S}$ where c is a scalar, then

$$\mathbf{A} = \mathbf{X}c\mathbf{L}\mathbf{X}' , \tag{3.40}$$

and if $\mathbf{A} = \mathbf{S} + c\mathbf{I}$ then

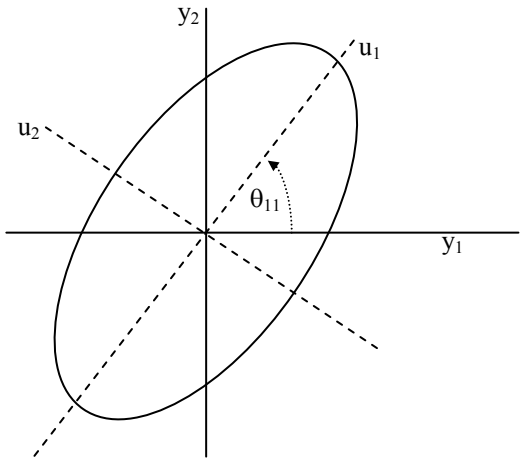
$$\mathbf{A} = \mathbf{X}[\mathbf{L} + c\mathbf{I}]\mathbf{X}' . \tag{3.41}$$

3.8 *Some Geometric Aspects of Eigenstructure*

Since $\mathbf{X}'\mathbf{X} = \mathbf{I}$, $\mathbf{X}$ can be thought of as a rigid, or *angle-preserving transformation* of a coordinate space. The original vector $\mathbf{y}$ is transformed to $\mathbf{u}$ by $\mathbf{X}$ as in

$$\mathbf{u} = \mathbf{X}'y.$$

Here we have repeated Equation (3.21), except the transformation occurs for each eigenvector, not just the first one. Alternatively, instead of thinking of $\mathbf{y}$ as moving to $\mathbf{u}$, we can think of this as the axes of the space moving. A picture of this is now shown:



The angle between an old axis, y_i, and a new axis, u_j, is notated θ_{ij}. We note then for the two dimensional example given above, we have for X

$$\mathbf{X} = \begin{bmatrix} \cos\theta_{11} & \cos\theta_{12} \\ -\cos\theta_{21} & \cos\theta_{22} \end{bmatrix}.$$

The angles θ_{ij} are determined by the direction of the principle axis of the ellipsoid $\mathbf{x}'\mathbf{S}\mathbf{x} = \boldsymbol{\lambda}$.

3.9 *Linear and Nonlinear Parameter Estimation*

In almost all cases that we have in mathematical reasoning in marketing, there are some aspects of our model that we know, for example there might be the value π, and there are some values that we do not know and that therefore have to be estimated from the sample at hand. For example, in the linear model, $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$, the $\mathbf{X}$ matrix is known, but the $\boldsymbol{\beta}$ vector contains a set of regression slopes that need to be estimated from the sample. The topic of linear estimation is investigated in depth in Chapter 5. For now, we note that we create an objective function, that when optimized, will lead us to good estimates for this unknown parameter vector. For example, we might pick the sum of squares of deviations between predicted data and actual data. In that case we would have

$$f = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

$$= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

as our objective function. The goal then is to pick values in the $\boldsymbol{\beta}$ vector so as to make f as small as possible. According to the calculus, this can be done by determining the derivative of f with respect to $\boldsymbol{\beta}$, and setting it equal to zero as in

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

The derivative $\partial f/\partial \boldsymbol{\beta}$ is a linear equation and happens to contain solely elements of the $\mathbf{X}$ matrix, the $\mathbf{y}$ vector and $\boldsymbol{\beta}$ in various combinations. When we set it equal to zero, we can solve for $\boldsymbol{\beta}$ and end up with things on the right hand side that are known, namely $\mathbf{X}$ and $\mathbf{y}$. This allows us to derive a *closed form* or *analytical solution* for $\boldsymbol{\beta}$ that we call $\hat{\boldsymbol{\beta}}$,

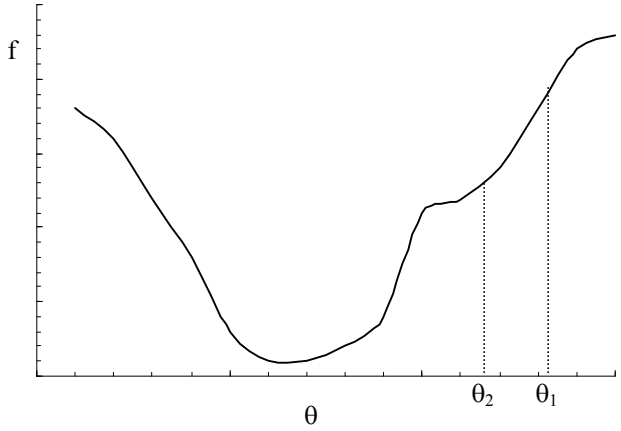
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The term closed-form means that we can use algebraic analysis to find the values for the unknowns. In short, we end up being able to solve for the unknowns. In other cases, our objective function, or its derivative, might be more complex. In those cases we cannot just solve for the unknown parameters using algebra. This often happens when we are trying to model choice probabilities, or market shares, which; since they are bounded by 0 and 1; logically cannot be represented linearly. When this happens we have to use *non-linear optimization*. Non-linear optimization involves the following steps.

1. We take a stab at the unknowns, inventing *starting values* for them and loading them into a vector. Lets call that vector $\boldsymbol{\theta}$.

2. We assess the derivative of the objective function at the current values in $\boldsymbol{\theta}$. If the derivative is not zero, we modify $\boldsymbol{\theta}$ by moving it in the direction in which the derivative is getting closer to $\mathbf{0}$, the null vector. We keep repeating this step until the derivative arrives at the null vector.

How do we know in which direction to move $\boldsymbol{\theta}$? First we will look at a geometric picture, then we will use symbols to make the argument. Lets assume that instead of an entire vector of unknowns, we have a single unknown; the scalar θ. We have started out with an estimate of θ at θ_1.



We are trying to move our estimate towards the bottom of the function. This is logically analogous to a parachutist who lands on the side of the hills surrounding a valley and who wants to find the bottom of the valley in the dead of night. How does he or she know which way to move? By feeling with your foot, you can figure out which way is down. The derivate $\partial f/\partial\theta_1$ gives us the slope of the function that relates θ to f, evaluated at θ_1. It lets us know which way is down. If the derivative is negative, we need to move to our right on the graph, because that is the direction in which f is less. On the other hand, if the derivative is positive, as it would be at position θ_1, we need to move to our left. In more formal terms, in nonlinear optimization we could calculate the next estimate of θ using the formula

$$\theta_{i+1} = \theta_i - \delta\frac{\partial f}{\partial\theta_i}$$

where δ is the step size. Sometimes we use the derivative of the derivative (the second order derivative) to fine-tune the step size. The step size can be important because we want to make sure we end up at the *global minimum* of f, not a *local minimum*. It also can help when you have good, rational, starting values for the first step that are close to their true values. Good start values and a good choice for step size can also make the search go faster, something that is still important even in these days of cheap computing power. In any case, non-linear optimization algorithms stop when the derivative gets close enough to zero, or in other words, when the difference between successive estimates of the unknowns does not change any more. Its important to understand that typically, there are more than one unknown parameters estimated at the same time. Thus the parameters and their derivatives are in vector form.

Nonlinear estimation is used in many branches of statistics and is needed in almost every chapter except for 5, 6, 7 and 8.

3.10 *Maximum Likelihood Parameter Estimation*

Rather than minimize the sum of squared errors, a different philosophy would have us maximize the likelihood of the sample. In general, the probability that our model is correct is proportional to the probability of the data given the model. In *Maximum Likelihood* (ML), we pick parameter estimates such that the probability of the data is as high as possible. Of course, it only makes sense that we would want to maximize the probability of observing the data that we actually did observe.

We can illustrate this using μ, the population mean. Suppose that we had a sample of three people, with scores of 4, 6 and 8. What would the probability be of observing this sample if the true population value of μ was 249? Pretty low, right? What would the probability of the sample be if μ was equal to 6? Certainly it would be quite a bit higher. The likelihood principle tells us to pick that estimate for μ that maximizes the probability of the sample. Of course to do this, we need to make an assumption about the probability distribution of the observations that comprise the sample.

To make the discussion more general, consider a set of observations $y_1, y_2, \cdots, y_n$. Lets say further that we have a model and that the unknown parameters of the model are in the vector $\boldsymbol{\theta}$. According to the model, the likelihood of observation i is $\Pr(y_i \mid \boldsymbol{\theta})$. Assuming independent sample units, i. e. no data point is influenced by any other, the likelihood function according to the model is

$$\ell_0 = \prod_i^n \Pr(y_i \mid \boldsymbol{\theta}) . \tag{3.42}$$

In these cases we also tend to have a version of the $\Pr(y_i)$ that does not depend on $\boldsymbol{\theta}$. The likelihood of the sample under this alternative may be called ℓ_A. It turns out that under very general conditions, $-2\ln(\ell_0/\ell_A)$ is distributed according to the Chi Square distribution, i. e.

$$\hat{\chi}^2 = -2\ln(\ell_0/\ell_A). \tag{3.43}$$

The minus sign in front of the expression for Chi Square means that we can switch from maximizing ℓ_0 to minimizing Chi Square. Minimization is always a safer bet where computers are concerned since a number too large to be processed causes far more of a problem than a number that is too close to zero (the square in Chi Square implies that it is non-negative). What's more, this allows us to test our model against the general alternative hypothesis using the χ^2 distribution. The degrees of freedom of the Chi Square are equal to the difference between the number of data points that we are using; in this case n, and the number of unknown elements in $\boldsymbol{\theta}$.

Here, it could be added that in some cases, such as linear regression, maximum likelihood estimates have a closed form and can be estimated using the formula for $\hat{\boldsymbol{\beta}}$ given in the previous section. In other words, $\hat{\boldsymbol{\beta}}$ does not just minimize the sum of squared errors, it also maximizes the likelihood function. In other cases, we don't get that sort of break and nonlinear optimization must be used.

Maximum likelihood comes with variances and covariances of the parameter vector "built-in". The matrix of the second order derivatives, known as the *Hessian*, contains the elements:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial \hat{\chi}^2}{\partial^2 \theta_1} & \frac{\partial \hat{\chi}^2}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial \hat{\chi}^2}{\partial \theta_1 \partial \theta_q} \\ \frac{\partial \hat{\chi}^2}{\partial \theta_2 \partial \theta_1} & \frac{\partial \hat{\chi}^2}{\partial^2 \theta_2} & \cdots & \frac{\partial \hat{\chi}^2}{\partial \theta_2 \partial \theta_q} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial \hat{\chi}^2}{\partial \theta_q \partial \theta_1} & \frac{\partial \hat{\chi}^2}{\partial \theta_q \partial \theta_2} & \cdots & \frac{\partial \hat{\chi}^2}{\partial^2 \theta_q} \end{bmatrix}. \tag{3.44}$$

Elements of the above matrix, $h_{ij} = \frac{\partial \hat{\chi}^2}{\partial \theta_i \partial \theta_j}$, consist of the derivative of the derivative of $\hat{\chi}^2$ with respect to θ_i, with respect to θ_j. In other words,

$$h_{ij} = \frac{\partial}{\partial \theta_i}\left[\frac{\partial \hat{\chi}^2}{\partial \theta_j}\right]. \tag{3.45}$$

Here we are treating $\frac{\partial \hat{\chi}^2}{\partial \theta_j}$ as a function of θ_j, and taking <u>its</u> derivative with respect to θ_j.

The covariance matrix of $\boldsymbol{\theta}$ is given by

$$V(\boldsymbol{\theta}) = [-E(\mathbf{H})]^{-1} \tag{3.46}$$

with the term in the square brackets, -E(**H**), minus the expectation of the Hessian, called the *information matrix*.

Whenever possible, marketing scientists prefer to work with maximum likelihood estimators given that they have very desirable properties. In addition to knowing the variance matrix of your estimator, if $\hat{\theta}$ is the maximum likelihood estimator of θ then $f(\hat{\theta})$ estimates $f(\theta)$ (for more detail see Johnson and Wichern, 2002, p. 170). You can estimate $\hat{\theta}$ and then apply the function f. More importantly, if you can derive or create a maximum likelihood estimator in a certain situation, that estimator is guaranteed to be consistent, asymptotically normally distributed and asymptotically efficient (a proof of this appears in Theil 1971, pp. 392-7). The phrase asymptotically efficient implies that no other estimator can have a lower variance.

References

Johnson, Richard A. and Dean W. Wichern (2002) *Applied Multivariate Statistical Analysis, Fifth Edition*. Upper Saddle River, NJ: Prentice-Hall.

Theil, Henri (1971) *Principles of Econometrics*. New York: John Wiley.

Chapter 4: Distributions

Prerequisite: Chapter 1

4.1 *The Algebra of Expectations and Variances*

In this section we will make use of the following symbols:

${}_n\mathbf{a}_1$ is a random variable
${}_n\mathbf{b}_1$ is a random variable
${}_n\mathbf{c}_1$ is a constant vector
${}_m\mathbf{D}_n$ is a constant matrix, and
${}_n\mathbf{F}_m$ is a constant matrix.

Now we define the *expectation of a continuous random variable*, such that

$$E(a_i) = \int_{-\infty}^{\infty} f(a_i)\, a_i da_i, \tag{4.1}$$

where $f(a_i)$ is the density of the probability distribution of a_i. Given that $f(a_i)$ is a density function, it must therefore be the case that

$$E(a_i) = \int_{-\infty}^{\infty} f(a_i) da_i = 1.$$

Often in this book, $f(a_i)$ will be taken to be normal, but not always. In fact, in some instances, a_i will be discrete rather than continuous. In that case,

$$E(a_i) = \sum_{j}^{J} Pr(a_i = j) \cdot j \tag{4.2}$$

where there are J discrete possible outcomes for a_i. We call $E(\cdot)$ the *expectation operator*. Regardless as to whether **a** and **b** are normal, the following set of theorems apply. First, we note that the expectation of a constant is simply that constant itself:

$$E(\mathbf{c}) = \mathbf{c}. \tag{4.3}$$

The expectation of a sum is equal to the sum of the expectations:

$$E(\mathbf{a} + \mathbf{b}) = E(\mathbf{a}) + E(\mathbf{b}). \tag{4.4}$$

The expectation of a linear combination comes in two flavors; one for premultiplication and one for postmultiplication:

$$E(\mathbf{Da}) = \mathbf{D}E(\mathbf{a}). \tag{4.5}$$

$$E(\mathbf{a}'\mathbf{F}) = E(\mathbf{a}')\mathbf{F}. \tag{4.6}$$

You can see from the above two equations that a constant matrix can pass through the expectation operator, which often simplifies our algebra greatly. All of these theorems will be important in enabling statistical inference and in trying to understand the average of various quantities.

We now define the *variance operator*, V(·), such that

$$V(\mathbf{a}) = E\left\{[\mathbf{a} - E(\mathbf{a})][\mathbf{a} - E(\mathbf{a})]'\right\}. \tag{4.7}$$

We could note here that if E(**a**) = **0**, that is if **a** is mean centered, the variance of **a** simplifies to E(**aa**').

Whether **a** is mean centered or not we also have the following theorems:

$$V(\mathbf{a} + \mathbf{c}) = V(\mathbf{a}). \tag{4.8}$$

Equation (4.8) shows that the addition (or subtraction) of a constant vector does not modify the variance of the original random vector. That fact will prove useful to us quite often in the chapters to come. But now it is time to look at what is arguably the most important theorem of the book. At least it is safe to say that it is the most referenced equation in the book:

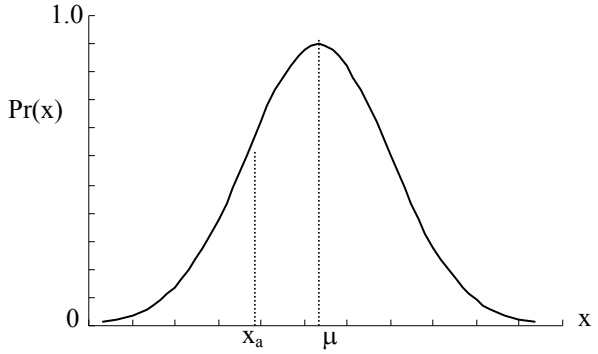
$$V(\mathbf{Da}) = \mathbf{D}V(\mathbf{a})\mathbf{D}' \tag{4.9}$$

$$V(\mathbf{a}'\mathbf{F}) = \mathbf{F}'V(\mathbf{a})\mathbf{F} \tag{4.10}$$

Equation (4.9), that shows that the variance of a linear combination is a quadratic form based on that linear combination, will be extremely useful to us, again and again in this book.

4.2 *The Normal Distribution*

The normal distribution is widely used in both statistical reasoning and in modeling marketing processes. It is so widely used that a short-hand notation exists to state that the variable x is normally distributed with mean μ and variance σ^2: $x \sim N(\mu, \sigma^2)$. We will start out by discussing the *density function* of the normal distribution even though the distribution function is somewhat more fundamental (it is, after all, called the normal distribution) and in fact the density is derived from the distribution function rather than vice versa. In any case, the density gives the probability that a variable takes on a particular value. We plot this probability as a function of the value:



The equation that sketches out the bell shaped curve in the figure is

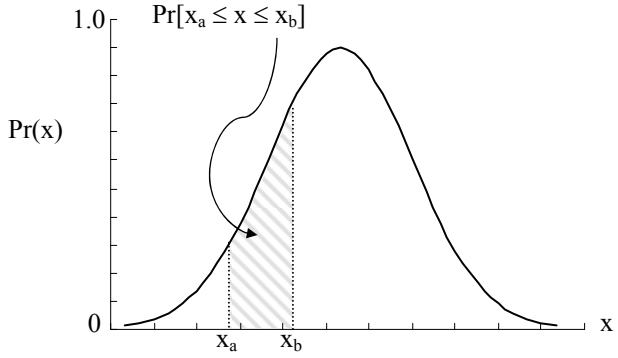
$$\Pr(x) \equiv \Pr(x = x_a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x_a - \mu)^2}{2\sigma^2}\right]. \tag{4.11}$$

Most of the "action" takes place in the exponent [and here we remind you that $\exp(x) = e^x$]. In fact, the constant $1/\sqrt{2\pi}\sigma$ is needed solely to make sure that the total probability under the curve equals one, or in other words, that the function integrates to 1. You might also note that the σ is not under the radical sign. Alternatively you can include a σ^2 under the radical. When we standardize such that $\mu = 0$ and $\sigma^2 = 1$ we generally rename x_a to z_a and then

$$\Pr(z) \equiv \Pr(z = z_a) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z_a^2}{2}\right] = \phi(z_a) \quad . \tag{4.12}$$

Note that $\phi(\cdot)$ is a very widely used notational convention to refer to the standard normal density function. This will show up in many places in the chapters to follow.

In statistical reasoning, we are often interested in the probability that a normal variable falls between two particular values, say x_a and x_b. We can picture this situation as below:



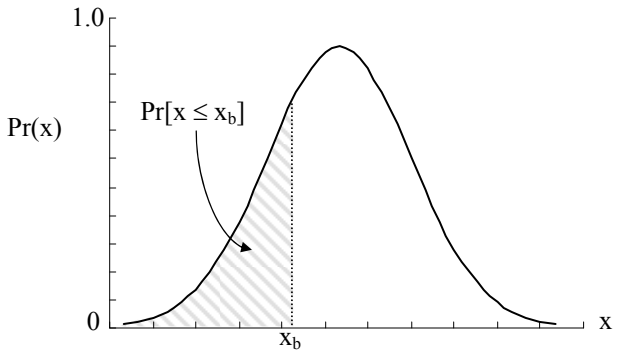
We can derive the probability by integrating the area under the curve from x_a to x_b. There is no analytic answer – that is to say no equation will allow you to calculate the exact value – so the only way you can do it is by a brute force computer program that creates a series of tiny rectangles between x_a and x_b. If the bases of these rectangles become sufficiently small, even though the top of the function is obviously not flat, we can approximate this probability to an arbitrary precision by adding up the areas of these rectangles. We write this area using the integral symbol as below:

$$\Pr[x_a \le x \le x_b] = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_a}^{x_b} \exp\left[\frac{-(x_a - \mu)^2}{2\sigma^2}\right] dx.$$

We can standardize, using the calculus change-of-variables technique, and then move the constant under the integral, all of which yields the same probability as above. This is shown next:

$$\Pr[z_a \le z \le z_b] = \int_{z_a}^{z_b} \phi(z)\, dz \, .$$

We are now ready to define the *normal distribution function*, which means the probability that x is less than or equal to some value, like x_b. This is pictured below:



Here, to calculate this probability, we must integrate the left tail of the distribution, starting at $-\infty$ at ending up at x_b. This will give us the probability that a normal variate x is less than x_b:

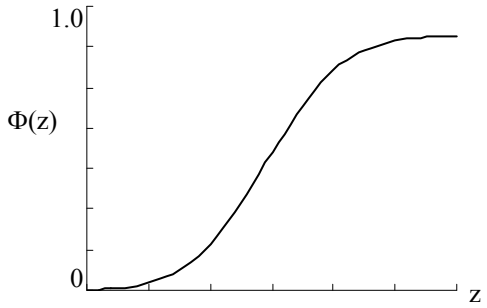
$$\Pr[x \le x_b] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x_b} \exp\left[\frac{-(x_a - \mu)^2}{2\sigma^2}\right] dx \quad \text{or} \tag{4.13}$$

$$= \int_{-\infty}^{z_b} \phi(z)\, dz = \Phi(z_b)\,. \tag{4.14}$$

Note the notation $\Phi(z_b)$ implies the probability that $z \le z_b$ The symbol Φ is an uppercase phi while ϕ is the lowercase version of that Greek letter. It is traditional to use a lower case letter for a function, while the integral of that function is signified with the upper case version of that letter. Note also that

$$\frac{\partial \Phi(z)}{\partial z} = \phi(z)\,. \tag{4.15}$$

A graphical representation of $\Phi(z)$ is show below:



The curve pictured above is often called an *ogive*.

In many cases, for example cases having to do with choice probabilities in Chapter 12, we wish to know that probability that a random variate is greater than 0:

$$\Pr(x \geq 0) = \Phi(\mu / \sigma) \equiv \Phi\left[E(x) \Big/ \sqrt{V(x)}\right] . \tag{4.16}$$

4.3 *The Multivariate Normal Distribution*

For purposes of comparison, let us take the normal distribution as presented in the previous section,

$$\Pr(x) \equiv \Pr(x = x_a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x_a - \mu)^2}{2\sigma^2}\right]$$

and rewrite it a little bit. For one thing, $\sqrt{a} = a^{1/2}$. In that case, rewriting the above gives us

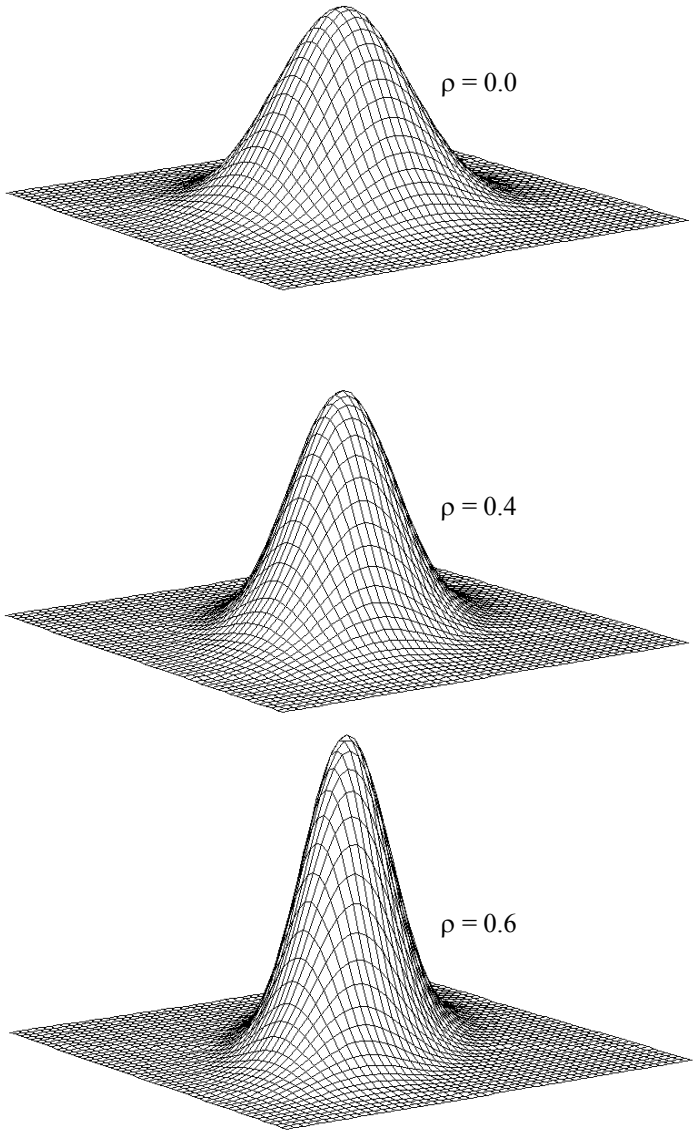
$$\Pr(x = x_a) = \frac{1}{(2\pi)^{1/2}\left(\sigma^2\right)^{1/2}} \exp\left[\frac{-(x_a - \mu)^2}{2\sigma^2}\right].$$

Now lets say we have a column vector of p variables, **x**, and that **x** follows the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ (which is also p by 1), and variance matrix $\boldsymbol{\Sigma}$ (which is a symmetric p by p matrix). In that case, the probability that the random vector **x** takes on the set of m values that we will call $\mathbf{x}_a$ is given by

$$\Pr(\mathbf{x} = \mathbf{x}_a) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left[(\mathbf{x}_a - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_a - \boldsymbol{\mu}) / 2\right]. \tag{4.17}$$

We would ordinarily use a short-hand notation for Equation (4.17), saying that $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Making some analogies, in the univariate expression σ^2 appears in the denominator (of the exponent) while in the multivariate case we have $\boldsymbol{\Sigma}^{-1}$ filling the same role. You might also notice that in the fraction before the exponent, we see σ in the univariate case, but $\mid \boldsymbol{\Sigma} \mid^{1/2}$ shows up in the multivariate case, the square root of the determinant of the variance matrix. In the univariate case there is the square root of 2π, in the multivariate we see the $(p/2)^{th}$ root of 2π. A picture of the bivariate normal density function appears below for three different values of the correlation $\rho = \sigma_{12} / \sigma_1 \sigma_2$.



4.4 *Chi Square*

We have already seen that the scalar y, where $y \sim N(\mu, \sigma^2)$, can be converted to a z score, $z \sim N(0, 1)$ where $z = \frac{y - \mu}{\sigma}$. If I square that z score I end up with a chi square variate with one degree of freedom, i. e.

$$z^2 = \chi_1^2 .$$

More generally, if I have a vector $\mathbf{y} = [y_1 \quad y_2 \quad \cdots \quad y_n]'$ and if $\mathbf{y}$ is normally distributed with mean vector

$$\boldsymbol{\mu} = \begin{bmatrix} \mu \\ \mu \\ \cdots \\ \mu \end{bmatrix}$$

and variance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

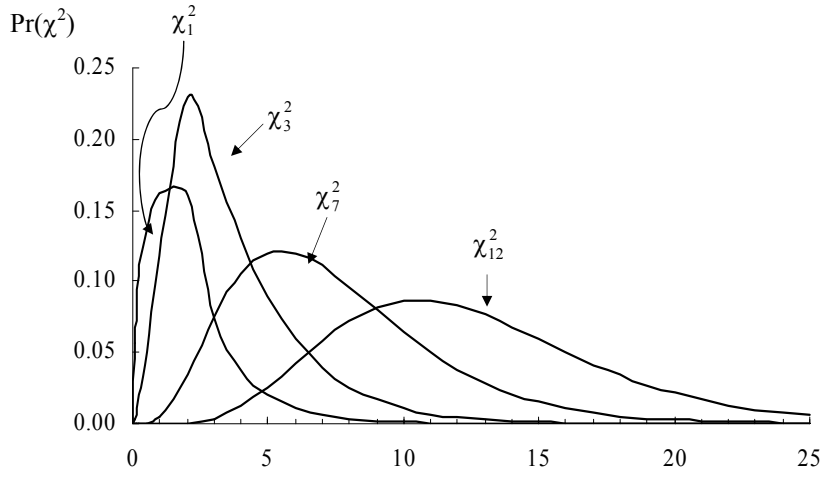
we of course say that $y \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. Converting each of the y_i to z scores, that is

$$z_i = \frac{y_i - \mu}{\sigma}$$

for all i, 1, 2, ···, n; we have $\mathbf{z} = [z_1 \quad z_2 \quad \cdots \quad z_n]'$. We can say that the vector $\mathbf{z} \sim N(0, \mathbf{I})$. In that case,

$$\mathbf{z}'\mathbf{z} = \sum_i^n z_i^2 \sim \chi_n^2.$$

The Chi Square density function is approximated in the following figure, using several different degrees of freedom to illustrate the shape.



With small degrees of freedom, the distribution looks like a normal for which the left tail has been folded over the right. This is more or less what happens when we square something - we fold the negative half over the positive. With larger degrees of freedom, the Chi Square begins to resemble the normal again, and in fact, as can be seen in the graph, the similarity is already quite striking at 12 degrees of freedom. This similarity is virtually complete by 30 degrees of freedom.

4.5 *Cochran's Theorem*

For any n · 1 vector $\mathbf{z} \sim N(0, \mathbf{I})$ and for any set of n · n matrices $\mathbf{A}_i$ where $\sum_i^n \mathbf{A}_i = \mathbf{I}$, then

$$\sum_i^n \mathbf{z}'\mathbf{A}_i\mathbf{z} = \mathbf{z}'\mathbf{z} \tag{4.18}$$

which, as we have just seen, is distributed as χ^2_n. Further, if the rank (see Section 3.7) of $\mathbf{A}_i$ is r_i we can say that

$$\sum_i^n r_i = n \quad \text{and} \tag{4.19}$$

$$\mathbf{z}'\mathbf{A}_i\mathbf{z} \sim \chi^2_{r_i}. \tag{4.20}$$

Each quadratic form $\mathbf{z}'\mathbf{A}_i\mathbf{z}$ is an independent Chi Square. The sum of independent Chi Square values is also a Chi Square variable with degrees of freedom equal to the sum of the component's degrees of freedom. This allows us to test nested models, such as those found in Chapters 9 and 10 as well as Chapters 12 and 13. In addition, multiple degree of freedom hypothesis testing for the linear model is based on this theorem as well. Defining $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{M} = \mathbf{I} - \mathbf{P}$, then since

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{I}\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{y} + \mathbf{y}'\mathbf{M}\mathbf{y},$$

we have met the requirements of Cochran's Theorem and we can form an F ratio using the two components, $\mathbf{y}'\mathbf{P}\mathbf{y}$ and $\mathbf{y}'\mathbf{M}\mathbf{y}$. In addition, the component $\mathbf{y}'\mathbf{P}\mathbf{y}$ can be further partitioned using the hypothesis matrix **A** or restricted models.

4.6 *Student's* t*-Statistic*

Like the normal distribution, the Chi Square is derived with a known value of σ. The formula for Chi Square on n degrees of freedom is

$$\chi^2_n = \sum_i^n \frac{(y_i - \mu)^2}{\sigma^2} = \sum_i^n \frac{[(y_i - \bar{y}) + (\bar{y} - \mu)]^2}{\sigma^2}. \tag{4.21}$$

You will note in the numerator of the right hand piece, a $\bar{y}$ has been added and subtracted. Now we will square the numerator of that right hand piece which yields

$$\chi^2_n = \frac{1}{\sigma^2}\sum_i^n \frac{(y_i - \bar{y})^2 + 2y_i\bar{y} - 2y_i\mu - 2\bar{y}^2 + 2\bar{y}\mu + \bar{y}^2 - 2\bar{y}\mu + \mu^2}{\sigma^2}. \tag{4.22}$$

At this time, we can modify Equation (4.22) by distributing the Σ addition operator, canceling some terms, and taking advantage of the fact that

$$\sum_{i}^{n} y_i = n\bar{y}.$$

Doing so, we find that

$$\chi_n^2 = \sum_{i}^{n} \frac{(y_i - \bar{y})^2}{\sigma^2} + \frac{n(\bar{y} - \mu)^2}{\sigma^2}. \tag{4.23}$$

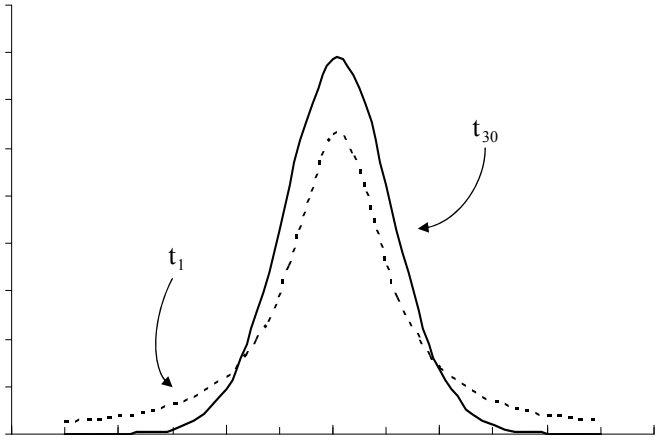
You might note that at this point Equation (4.23) shows the decomposition of an n degree of freedom Chi Square into two components which Cochran's Theorem shows us are both themselves distributed as Chi Square. But the numerator of the summation on the right hand side, that is $\sum_{i}^{n} (y - \bar{y})^2$, is the corrected sum of squares and as such it is equivalent to $(n - 1)s^2$. Rewriting both components slightly we have

$$\chi_n^2 = \frac{(n-1)s^2}{\sigma^2} + \frac{(\bar{y} - \mu)^2}{\sigma^2/n} \tag{4.24}$$

which leaves us with two Chi Squares. The one on the right is a z-score squared and has one degree of freedom. The reader might recognize it as a z score for the arithmetic mean, $\bar{y}$. The Chi Square on the left has n - 1 degrees of freedom. At this point, to get the unknown value σ^2 to vanish we need only create a ratio. In fact, to form a *t*-statistic, we do just that. In addition, we divide by the n - 1 degrees of freedom in order to make the *t* easier to tabulate:

$$\begin{aligned} t &= \sqrt{\frac{(\bar{y} - \mu)^2}{\sigma^2/n} \Big/ \frac{(n-1)s^2}{\sigma^2(n-1)}} \\ &= \frac{\bar{y} - \mu}{s/\sqrt{n}} \end{aligned} \tag{4.25}$$

The more degrees of freedom a *t* distribution has, the more it resembles the normal. The resemblance is well on its way by the time you reach 30 degrees of freedom. Below you can see a graph that compares the approximate density functions for *t* with 1 and with 30 df.



The 1 df function has much more weight in the tails, as it must be more conservative.

4.7 *The F Distribution*

With the F statistic, a ratio is also formed. However, in the case of the F, we do not take the square root, and the numerator χ^2 is not restricted to one degree of freedom:

$$F_{r_1,r_2} = \frac{\chi^2_{r_1} / r_1}{\chi^2_{r_2} / r_2} \quad .$$

Section II: The General Linear Model

Chapter 5: Ordinary Least Squares

Prerequisite: Chapters 1, 2, Sections 3.1, 3.2, 3.3, 4.1, 4.2

5.1 *The Regression Model*

The linear algebra that we covered in Chapter 1 will now be put to use in explaining the variance among observations on a dependent variable, placed in the vector **y**. For each of these observations y_i, we posit the following model:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ik^*}\beta_{k^*} + e_i. \tag{5.1}$$

Economists have traditionally referred to Equation (5.1) as ordinary least squares, while other fields sometime use the expression *regression*, *or least squares regression*. Whatever we choose to call it, putting this equation in matrix terms, we have

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k^*} \\ 1 & x_{21} & \cdots & x_{2k^*} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{nk^*} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_{k^*} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{5.2}$$

The number of columns of the **X** matrix is k = k* + 1. If you wish, you can think of **X** as containing k* "real" independent variables, plus there is one additional independent variable that is nothing more than a series of 1's.

The mechanism of prediction is a linear combination of independent variable values, with coefficients known as β's. The prediction for y_i, in other words $E(y_i)$, is traditionally notated with a hat as below:

$$E(y_i) \equiv \hat{y}_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ik^*}\beta_{k^*}$$

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}. \tag{5.3}$$

Each $\hat{y}_i$ is formed as the linear combination $\mathbf{x}_{i\cdot}'\boldsymbol{\beta}$, with the dot defined as in Equation (1.2).

The difference between $\hat{\mathbf{y}}$ and **y** is the error, that is $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ as $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$. The error vector is a key input in ordinary least squares. Assumptions about the nature of the error are largely responsible for our ability to make inferences from and about the model. To start, we assume that $E(\mathbf{e}) = \mathbf{0}$ where both **e** and **0** are n by 1 columns. Note that this is an assumption that does not restrict us in any way. If $E(\mathbf{e}) \neq \mathbf{0}$, the difference would simply be absorbed in the y-intercept, β_0.

5.2 *Least Squares Estimation*

One of the most important themes in this book is the notion of *estimation*. In our model, the values in the **y** vector and the **X** matrix are known. They are data. The values in the **β** vector, on

the other hand, have a different status. These are unknown and hence reflect ignorance about the theoretical situation at hand. These must be estimated in some way from the sample. How do we go about doing this? In Section 5.4 we cover the maximum likelihood approach to estimating regression parameters. Maximum likelihood is also discussed in Section 3.10. For now, we will be using *the least squares principle*. This is the idea that the sum of the squared errors of prediction of the model, the e_i, should be as small as possible. We can think about this as a *loss function*. As values of y_i and $\hat{y}_i$ increasingly diverge, the square of their difference explodes and observation i figures more and more in the solution for the unknown parameters.

The loss function f is minimized over all possible (combinations of) values in the $\boldsymbol{\beta}$ vector: $\min_{\boldsymbol{\beta}} f$ where f is defined as

$$f = \mathbf{e}'\mathbf{e} = \sum_i^n e_i^2 = \sum_i^n (y_i - \hat{y}_i)^2$$

$$= (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

$$= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$= \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

Note that f is a scalar and so are all four components of the last equation above. Components 2 and 3 are actually identical. (Can you explain why? Hint: Look at Equation (1.5) and the discussion thereof.) We can simplify by combining those two pieces as below:

$$= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \tag{5.4}$$

The minimum possible value of f occurs where $\frac{\partial f}{\partial \boldsymbol{\beta}} = \mathbf{0}$, that is to say, when the partial derivatives of f with respect to each of the elements in $\boldsymbol{\beta}$ are all zero. In this case, the null vector on the right hand side is k by 1, that is, it has k elements, all zeroes. As we learned in Equation (3.12), the derivative of a sum is equal to the sum of the derivatives, so we can analyze our f function one piece at a time. The value of $\partial \mathbf{y}'\mathbf{y}/\partial \boldsymbol{\beta}$ is just a k by 1 null vector since $\mathbf{y}'\mathbf{y}$ is a constant with respect to $\boldsymbol{\beta}$. The derivative $\frac{\partial}{\partial \boldsymbol{\beta}}\left[-2\mathbf{y}'\mathbf{X}\boldsymbol{\beta}\right]$ can be determined from two rules for derivatives covered in Chapter 3, namely the derivative of a linear combination

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}'} = \mathbf{a}'$$

from Equation (3.17) and the derivative of a transpose

$$\frac{\partial f}{\partial \mathbf{x}} = \left[\frac{\partial f}{\partial \mathbf{x}'}\right]'$$

from Equation (3.19).

In this case the role of "**a**" above is being played by -2**y′X** and the role of **x** is being played by **β**:

$$\frac{\partial}{\partial \boldsymbol{\beta}}\left[-2\mathbf{y}'\mathbf{X}\boldsymbol{\beta}\right] = -2\mathbf{X}'\mathbf{y}\,.$$

As for piece number 3, **βX′Xβ** is a quadratic form and we have seen a derivative rule for that also, in Equation (3.18). Using that rule we would have

$$\frac{\partial\, \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\,.$$

Finally, adding all of the pieces together, each being k by 1, we have

$$\frac{\partial \mathrm{f}}{\partial \boldsymbol{\beta}} = 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y} = \mathbf{0}\,. \tag{5.5}$$

We are at an extreme point where any derivative $\partial f(x)/\partial x = 0$. At the minimum, in our case we then have

$$2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y} = \mathbf{0} \tag{5.6}$$

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \tag{5.7}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\,. \tag{5.8}$$

The k equations described in Equation (5.7) are sometimes called the *normal equations*. The last line gives us what we need, a statistical formula we can use to estimate the unknown parameters.

It has to be admitted at this point that a hat somehow snuck onto the **β** vector just in time to show up in the last equation above, Equation (5.8). That is a philosophical matter that has to do with the fact that up to this point, we have had only a theory about how we might go about estimating the parameter matrix **β** in our model. The last equation above, however, gives us a formula we can actually use with a sample of data. Unlike **β**, $\hat{\boldsymbol{\beta}}$ can actually be held in one's hand. It is one of a possible infinite number of ways we could estimate **β**. The hat tells us that it is just one statistic from a sample that might be proposed to estimate the unknown population parameter.

Is the formula any good? We know that it minimizes f. That means that there is no other formula that could give us a smaller sum of squared errors for our model. Perhaps some idea of the efficacy of this formula can be had by thinking about its expectation. So what about the expectation of $\hat{\boldsymbol{\beta}}$? What does that look like?

$$
\begin{aligned}
E(\boldsymbol{\beta}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{X}\boldsymbol{\beta}] \\
&= \boldsymbol{\beta}
\end{aligned}
\tag{5.9}
$$

Here we have relied on the identity $\hat{\mathbf{y}} \equiv E(\mathbf{y})$ going from the second to the third line above. Also, we passed $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ through the expectation operator, something that is certainly legal and in fact was talked about in Equation (4.5). However, applying Theorem (4.5) in that way means that we are treating the $\mathbf{X}$ matrix as constant. Strictly speaking, the fact that $\mathbf{X}$ is fixed implies we cannot generalize beyond the values in $\mathbf{X}$ that we have observed. The good news in the last line above is that the expectation of $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{\beta}$, which certainly appears to be a good sign. However, it actually turns out that this is not strictly necessary. There are other properties that are more important. We turn now to those.

5.3 *What Do We Mean by a Good Statistic?*

A good estimator, like our vector $\hat{\boldsymbol{\beta}}$, should have four properties. We have already talked about one of them: unbiasedness:

Unbiased $$E(\hat{\beta}_i) = \beta_i. \tag{5.10}$$

Consistent $$\Pr(\hat{\beta}_i - \beta_i \leq \varepsilon) \rightarrow 1 \text{ as } n \rightarrow \infty. \tag{5.11}$$

The above expression is sometimes written using the notation *Plim*, which stands for Probability limit. In that case, Equation (5.11) boils down to

$$\text{Plim}\,\hat{\beta}_i = \beta_i.$$

In effect what is going on with consistency is that as $n \rightarrow \infty$, $\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}$. Unbiasedness turns out to not be as important as consistency. Even if the average estimator is not equal to the parameter, if we can show that it gets closer and closer as the sample size increases, this is fine. Conversely, if the average estimator is equal to the parameter, but increasing the sample size doesn't get you any closer to that truth, that would not be good. Now, another characteristic of a good estimator is that it is

Sufficient $$\Pr(\mathbf{y} \mid \hat{\boldsymbol{\beta}}) \text{ does not depend on } \boldsymbol{\beta} \tag{5.12}$$

Sufficiency implies that the formula for the estimator has wrung out all of the information in the sample that there is about the parameter. Finally, efficiency is very important and forms the basis for reasoning about the population based on the sample:

Efficient $\quad V(\hat{\boldsymbol{\beta}}) \equiv E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$ is smaller than other estimators (5.13)

To show that a statistic is efficient, you need to derive its variance, and the variance is invariably needed for hypothesis testing and confidence intervals. If this variance is large, you will not be able to reject even really bad hypotheses.

As we saw above in Equation (5.9), unbiasedness can be demonstrated without any distributional assumptions about the data. You will note that not a word has been mentioned – up to this point - as to whether anything here is normally distributed or not. Some of these other properties require distributional assumptions to prove. In our model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, the $\mathbf{e}$ vector will play an important role in these assumptions. Both $\mathbf{X}$ and $\boldsymbol{\beta}$ contain fixed values; the former being simply data and the latter; by assumption a set of constant values true of the population as a whole. The only input that varies randomly is $\mathbf{e}$. From this point forward in this chapter we will assume that

$$ {}_{n}\mathbf{e}_{1} \sim N({}_{n}\mathbf{0}_{1},\ {}_{n}\boldsymbol{\Sigma}_{n}) . \tag{5.14} $$

This notation (see Section 4.2 for a review) tells us that the n by 1 error vector $\mathbf{e}$ is normally distributed with a mean equal to the null vector, and with a variance matrix $\boldsymbol{\Sigma}$. Since $\mathbf{e}$ is n by 1, its mean must be n by 1, and the variances and covariances among the n elements of $\mathbf{e}$ can be arrayed in an n by n symmetric matrix.

Given the assumption above, and our model, we can deduce [from Equations (4.4) and (4.8)] about the y vector that

$$ {}_{n}\mathbf{y}_{1} \sim N({}_{n}\mathbf{X}\boldsymbol{\beta}_{1},\ {}_{n}\boldsymbol{\Sigma}_{n}). \tag{5.15} $$

Now we are ready to add an important set of assumptions, often called the *Gauss-Markov assumptions*. These deal with the form of the n · n error variance-covariance matrix, $\boldsymbol{\Sigma}$. We assume that

$$ \boldsymbol{\Sigma} = \sigma^2\mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}, \tag{5.16} $$

which is really two assumptions. For one, each e_i value has the same variance, namely σ^2. For another, each pair of errors, e_i and e_j (for which $i \neq j$), is independent. In other words, all of the covariances are zero. Since $\mathbf{e}$ is normal, this series of assumptions is often called *NIID*, that is to say we are asserting that $\mathbf{e}$ is normally, identically and independently distributed.

5.4 *Maximum Likelihood Estimation of Regression Parameters*

Lets review for a moment the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ with $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. *Maximum Likelihood* (*ML*) estimation begins by looking at the probability of observing a particular observation, y_i. The formula for the normal density function, given in Equation (4.11), tells us that

$$\Pr(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(y_i - \mathbf{x}_{i\cdot}'\boldsymbol{\beta})^2 / 2\sigma^2\right] \tag{5.17}$$

where $\mathbf{x}_{i\cdot}'$ is the ith row of $\mathbf{X}$, i. e. the row needed to calculate $\hat{y}_i$ as below,

$$\hat{y}_i = \begin{bmatrix} 1 & x_{i1} & \cdots & x_{ik^*} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_{k^*} \end{bmatrix}.$$

The part of the normal density that appears as an exponent (to e) is basically the negative one half of a z-score squared, that is $-\frac{1}{2}z^2$. The role of "μ" in $z = \frac{y-\mu}{\sigma}$ is being played by $E(y_i) \equiv \hat{y}_i = x_{i\cdot}'\boldsymbol{\beta}$.

Now that we have figured out the probability of an individual observation, the next step in the reasoning behind ML is to calculate the probability of the whole sample. Since we assume that we have independent observations, that means we can simply multiply out the probabilities of all of the individual observations as is done below,

$$\ell = \Pr(\mathbf{y}) = \prod_i^n \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-(y_i - x_{i\cdot}'\boldsymbol{\beta})^2 / 2\sigma^2\right]$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\sum_i^n (y_i - x_{i\cdot}'\boldsymbol{\beta})^2 / 2\sigma^2\right] \tag{5.18}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) / 2\sigma^2\right].$$

How did we get to the last step? Here are some reminders from Section 3.1. First recall that exp[a] = e^a. Next, you need to remember that we can write $a^{1/2} = \sqrt{a}$. It is also true that $\prod \exp[f_i] = \exp\left[\sum f_i\right]$ because $e^a e^b = e^{a+b}$, that multiplying a constant $\prod_i^n a = a \cdot a \cdot \cdots \cdot a = a^n$ and finally that $\sum (a_i - b_i)^2 = (a-b)'(a-b)$.

In Section 5.2 we choose a formula, $\hat{\boldsymbol{\beta}}$, based on the idea of minimizing the sum of squared errors of prediction. But the least squares principle is just one way to choose a formula. The Maximum likelihood principle gives us an alternative logical path to follow in coming up with parameter estimates. The probability that our model is true is proportional to the likelihood of the sample, called ℓ or more specifically Pr(**y**). Therefore, it makes sense to pick $\hat{\boldsymbol{\beta}}$ such that ℓ is as large as possible.

It actually turns out to be simpler to maximize the log of the likelihood of the sample. The maximum point of ℓ is the same as maximum point of L = ln(ℓ), so this does not impact anything

except that it makes our life easier. After all, the likelihood of independent observations involves multiplication, and the ln function takes multiplication into addition which simplifies our task. Returning to the regression model, we have

$$L = \ln(\mathbf{y}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{(\mathbf{y}-\mathbf{X\beta})'(\mathbf{y}-\mathbf{X\beta})}{2\sigma^2} \tag{5.19}$$

with derivative

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \frac{1}{2\sigma^2}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}). \tag{5.20}$$

If we take $\partial L/\partial \boldsymbol{\beta} = 0$, multiply both sides by $2\sigma^2$, and solve for $\boldsymbol{\beta}$ we end up with the same formula that we came up with using the least squares principle, namely $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Thus $\hat{\boldsymbol{\beta}}$ is the least squares and the maximum likelihood estimator. Things don't always work out this way; sometimes least squares and ML estimators may be different and therefore in competition with each other. ML always has much to recommend it though. Whenever ML estimators exist, they can be shown to be efficient [see Equation (5.13)].

But now it is time to return to the theme of this chapter, confirmatory factor analysis. We need to be able to develop ML estimators for our three parameter matrices; $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$. Let us return to that task.

5.5 *Sums of Squares of the Regression Model*

Now that we have a formula $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, we can go back to our original objective function, $f = \mathbf{e}'\mathbf{e}$. We frequently call this scalar the *sum of squares error*, written alternatively as SS_{Error} or SSE. Now

$$SS_{Error} = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{5.21}$$

$$= [\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]'[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

so that therefore

$$SS_{Error} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$SS_{Error} = SS_{Total} - SS_{Predictable} \tag{5.22}$$

The error sum of squares can be seen as a remainder from the total raw sum of squares of the dependent variable, after the predictable part of has been subtracted. Or, to put this another way, the SS_{Total} can be seen as the sum of the $SS_{Error} + SS_{Predictable}$.

There are many ways of expressing the $SS_{Predictable}$, including

$$\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}.$$

In order to prove to yourself that these are all equivalent, substitute the formula for $\hat{\boldsymbol{\beta}}$ into each of the alternative versions of the formula above and then simplify by canceling any product of the form $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$.

Taking the last version of the $SS_{Predictable}$ on the right, note that

$$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = [\hat{\boldsymbol{\beta}}'\mathbf{X}'][\mathbf{X}\boldsymbol{\beta}] = [\mathbf{X}\hat{\boldsymbol{\beta}}]'[\mathbf{X}\boldsymbol{\beta}] = \hat{\mathbf{y}}'\hat{\mathbf{y}} \ .$$

Thus $SS_{Predictable}$ is the sum of the squares of the predictions of the model, the $\hat{y}_i$. Another way to write the SS_{Error} is as

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}'$$

$$= \mathbf{y}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}')$$

$$= \mathbf{y}'\mathbf{e}.$$

However, the quantity $\mathbf{y}'\mathbf{e}$ (SS_{Error}) is not the same as $\hat{\mathbf{y}}'\mathbf{e}$ since

$$\hat{\mathbf{y}}'\mathbf{e} = (\mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$= (\hat{\boldsymbol{\beta}}'\mathbf{X}') (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad (5.23)$$

$$= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0.$$

Note that the last line above involves two equivalent versions of $SS_{Predictable}$, which, being equivalent, have a difference of 0. The upshot is that the predicted scores, $\hat{\mathbf{y}}$, and the errors, $\mathbf{e}$, are orthogonal vectors [Equation (1.17)] with a correlation of 0.

5.6 *The Covariance Estimator for β*

We can conveniently produce the $\hat{\boldsymbol{\beta}}$ vector from the covariances of all the variables; x variables and y included. We are going to place y in the first row and column of the covariance matrix, $\mathbf{S}$ [see Equation (2.12)]. The $\mathbf{S}$ matrix is partitioned (Section 1.4) into sections corresponding to the y variable and the x's:

$${}_k\mathbf{S}_k = \left[\begin{array}{c|c} s_{yy} & \mathbf{s}'_{xy} \\ \hline \mathbf{s}_{xy} & \mathbf{S}_{xx} \end{array}\right]. \qquad (5.24)$$

The scalar s_{yy} represents the variance of the y variable, $\mathbf{S}_{xx}$ is the covariance matrix for the independent variables, and $\mathbf{s}_{xy} = \mathbf{s}'_{yx}$ is the vector of covariances between the dependent variable and each of the independent variables. There is no information about the levels of the y or x

variables and so we will not be able to calculate $\hat{\beta}_0$ from **S**, but we can calculate all of the other k* β values using

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{xx}^{-1}\mathbf{s}_{xy} . \tag{5.25}$$

If we need to know what the value of $\hat{\beta}_0$ is, we can calculate it as follows:

$$\beta_0 = \bar{x}_y - \hat{\boldsymbol{\beta}}' \bar{\mathbf{x}}_x$$

where $\bar{x}_y$ is the mean of the dependent variable and the column vector $\bar{\mathbf{x}}_x$ contains the means of each of the independent variables.

5.7 *Regression with Z-Scores*

Instead of just using deviation scores and eliminating β_0, as was done in the previous section, we can also create a version of the $\boldsymbol{\beta}$ vector, $\boldsymbol{\beta}^*$ say, based on standardized versions of the variables and which therefore does not carry any information about the metric of the independent and dependent variables. This can sometimes be useful for comparing particular values in the $\boldsymbol{\beta}$ vector and other purposes.

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{Z}_x'\mathbf{Z}_x)^{-1}\mathbf{Z}_x\mathbf{z}_y \tag{5.26}$$

$$= \mathbf{R}_{xx}^{-1}\mathbf{r}_{xy}, \tag{5.27}$$

where $\mathbf{Z}_x$ represents the matrix of observations on the independent variables, after having been converted to Z-scores, and $\mathbf{z}_y$ is defined analogously for the **y** vector. The second way that we have written this, in Equation (5.27), is by using the partitioned correlation matrix, just as we did with the variance matrix above in Equation (5.24). Here the correlations among the independent variables are in the matrix $\mathbf{R}_{xx}$, and those between the independent variables and the dependent variable are in the vector $\mathbf{r}_{xy}$. The partitioned matrix is shown below:

$${}_k\mathbf{R}_k = \left[\begin{array}{c|c} 1 & \mathbf{r}_{xy}' \\ \hline \mathbf{r}_{xy} & \mathbf{R}_{xx} \end{array}\right] \text{ where} \tag{5.28}$$

$$\mathbf{R}_{xx} = \begin{bmatrix} 1 & r_{x_1,x_2} & \cdots & r_{x_1,x_{k^*}} \\ r_{x_2,x_1} & 1 & \cdots & r_{x_2,x_{k^*}} \\ \cdots & \cdots & \cdots & \cdots \\ r_{x_{k^*},x_1} & r_{x_{k^*},x_2} & \cdots & 1 \end{bmatrix}$$

is the matrix of correlations among the k* independent variables, and is therefore k* by k*, the same as S_{xx}, and

$$\mathbf{r}_{xy}' = \mathbf{r}_{yx} = [r_{y,x_1} \quad r_{y,x_2} \quad \cdots \quad r_{y,x_{x_{k^*}}}]$$

is the vector of correlations between the dependent variable and each of the k* independent variables.

It is interesting to note that in the calculation of $\hat{\boldsymbol{\beta}}$ as well as the standardized $\hat{\boldsymbol{\beta}}^*$, the correlations among all the independent variables figure into the calculation into each $\hat{\beta}_i$. Of course, if $\mathbf{R}_{xx} = \mathbf{I}$, this would simplify things quite a bit. In this case, each independent variable would be orthogonal from all the others and the calculation of each $\hat{\beta}_i$ could be done sequentially in any order, instead of simultaneously as we have done above. We can also see here why our regression model is unprotected from misspecification in the form of missing independent variables. If there is some other independent variable of which we are not aware, or at least that we did not measure, our calculations are obviously not taking it into account, even though its presence could easily modify the values of all the other β's. The only time we can be protected from the threat of unmeasured independent variables is when we can be totally sure that all unmeasured variables would be orthogonal to the independent variables that we did measure. How can we ever be sure of this? We are protected from unmeasured independent variables when we have a designed experiment that lets us control the assignment of subjects (or in general "experimental units", whatever they might be) to the values of the independent variables.

5.8 *Partialing Variance*

Lets assume we have two different sets of independent variables in the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. Each of these has n observations, so they both have n rows, but there are differing numbers of columns in $\mathbf{X}_1$ and $\mathbf{X}_2$. Our model is still $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ but

$$\mathbf{X} = [\mathbf{X}_1 \mid \mathbf{X}_2] \text{ and}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ -- \\ \boldsymbol{\beta}_2 \end{bmatrix},$$

where $\boldsymbol{\beta}_1$ is the vector with as many elements as there are columns in $\mathbf{X}_1$ while $\boldsymbol{\beta}_2$ is the vector corresponding to each of the independent variables in $\mathbf{X}_2$. Note that in this case $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors, not individual beta values. The reason we are doing this is so that we can look at the regression model in more detail, tracking the relationship between two different sets of independent variables. Now we can rewrite $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ as

$$\hat{\mathbf{y}} = [\mathbf{X}_1 \quad \mathbf{X}_2] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

$$= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2.$$

The normal equations [c.f. Equation 5.7] would be

$$\begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{bmatrix} [\mathbf{X}_1 \quad \mathbf{X}_2] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{bmatrix} \mathbf{y}$$

but we could also look at the normal equations one set of X variables at a time, as

$$\mathbf{X}_1'\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbf{X}_1'\mathbf{y}\,, \tag{5.29}$$

$$\mathbf{X}_2'\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2'\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbf{X}_2'\mathbf{y}\,. \tag{5.30}$$

If we substract $\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$ from Equation (5.29) we end up with

$$\mathbf{X}_1'\mathbf{X}_1\boldsymbol{\beta}_1 = \mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$$

which, after we solve for β_1, gives us the estimator

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2\,. \tag{5.31}$$

The first component of the right hand side of Equation (5.31) is just the usual least squares formula that we would see if there was only set $\mathbf{X}_1$ of the independent variables and $\mathbf{X}_2$ was not part of the model. Instead, something is being subtracted away from the usual formula. To shed more light on this, we can factor the premultiplying matrix $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ to get

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'[\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2]\,.$$

What is the term in brackets? None other than the error for the regression equation if there was only $\mathbf{X}_2$ and $\mathbf{X}_1$ was not part of the model. In other words, $\hat{\boldsymbol{\beta}}_1$ is being calculated not using **y**, but using the error from the regression of **y** on $\mathbf{X}_2$. The variance that is at all attributable to $\mathbf{X}_2$ has been swept out of the dependent variable **y** before $\hat{\boldsymbol{\beta}}_1$ gets calculated, and vice versa.

5.9 T*he Intercept-Only Model*

Define

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{5.32}$$

and define

$$\mathbf{M} = \mathbf{I} - \mathbf{P}, \tag{5.33}$$

i. e. $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Keeping these definitions in mind, let us now consider the simplest of all possible regression models, namely, a model with only an intercept term,

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \cdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} \hat{\beta}_0 = {}_n 1_1 \hat{\beta}_0\,.$$

In this case, the $\hat{\boldsymbol{\beta}}$ vector is just the scalar $\hat{\beta}_0$ and so it's formula becomes

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y}$$

$$= n^{-1}\mathbf{1}'\mathbf{y}$$

$$= [n^{-1} \quad n^{-1} \quad \cdots \quad n^{-1}]\mathbf{y}$$

$$= \frac{1}{n}\sum_{i}^{n} y_i$$

so that our model $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is just

$$_n\hat{y}_1 = {}_n 1_1 \bar{y}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \cdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{y} \\ \cdots \\ \bar{y} \end{bmatrix}.$$

The matrix $\mathbf{P}$ is given by the expression

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$= {}_n\mathbf{1}_1(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$$

$$= \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$$

so in that case the predicted values of $\mathbf{y}$ are

$$\hat{\mathbf{y}} = \mathbf{P}\mathbf{y} = \begin{bmatrix} \bar{y} \\ \bar{y} \\ \cdots \\ \bar{y} \end{bmatrix}$$

and the Sum of Squares Predictable are

$$SS_{Predicted} = \mathbf{y}'\mathbf{P}\mathbf{y} = \bar{y}\sum y_i \ .$$

The **M** matrix also takes on a particular form in the intercept-only model.

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I} - \mathbf{P}$$

$$= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$$

The M matrix transforms the observations in y into error, but in this case the "error" is equivalent to deviations from the mean (in other words d_i values):

$$\mathbf{e} = \mathbf{M}\mathbf{y} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \cdots \\ y_n - \bar{y} \end{bmatrix}.$$

The SS_{Error} is the quadratic form with M in the middle,

$$SS_{Error} = \mathbf{y}'\mathbf{M}\mathbf{y} = \sum y_i(y_i - \bar{y})$$

$$= \mathbf{y}'(\mathbf{y} - \mathbf{1}\bar{y}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{1}\bar{y} = \mathbf{y}'\mathbf{y}$$

$$= \sum y_i^2 - \bar{y}\sum y_i = \sum y_i^2 - \left(\frac{\sum y_i}{n}\right)\sum y_i,$$

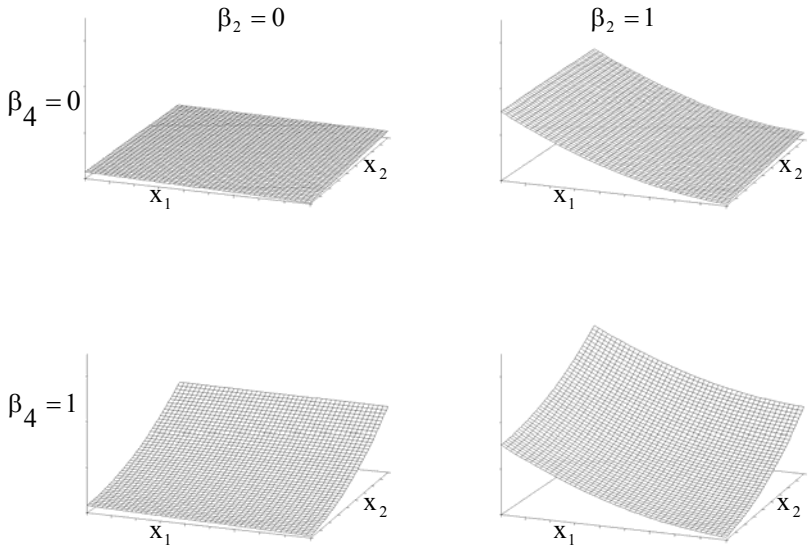
which the reader will recognize as the scalar, the corrected sum of squares from Equation (2.11).

5.10 *Response Surface Models*

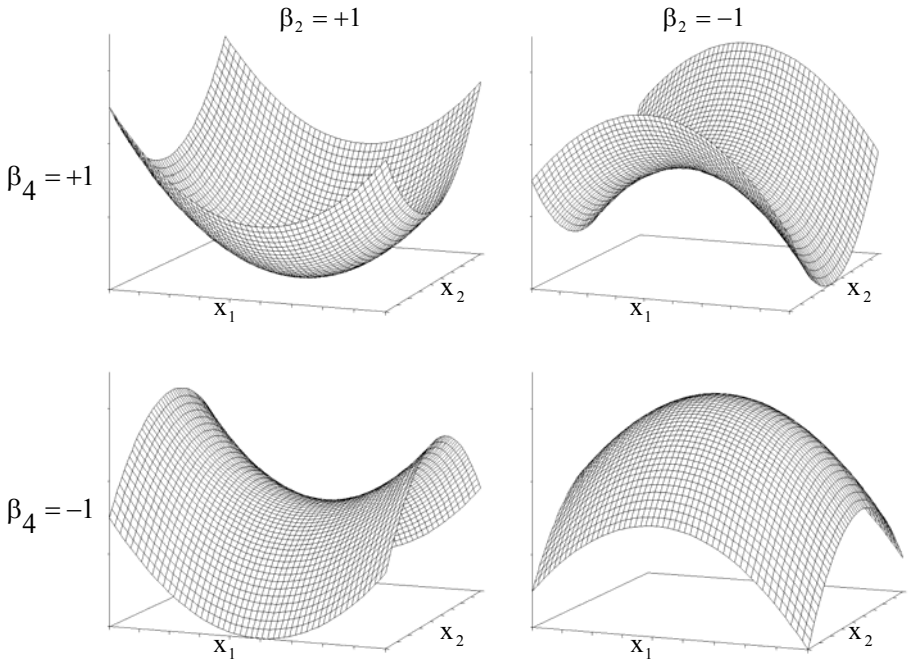
While it is known as the linear model, one can fit more complicated curves than lines or planes. It is relatively straightforward to include quadratic or higher order polynomials in a regression model, merely by squaring or cubing one of the independent variables (it is wise to mean center first). For example, consider the model

$$\hat{y}_i = \beta_0 + x_{i1}\beta_1 + x_{i1}^2\beta_2 + x_{i2}\beta_3 + x_{i2}^2\beta_4 .$$

The second and fourth independent variables are squared versions of the first and third. In order to demonstrate the wide variety of shapes we can model using polynomial equations, consider the figure below where β_2 and β_4 are either 0 or 1:



Or consider the following diagram in which the sign of β_2 and β_4 is either positive or minus:



References

Mosteller, Frederick and John W. Tukey (1977) *Data Analysis and Regression*. Reading, MA: Addison-Wesley.

Chapter 6: Testing Linear Hypotheses

Prerequisites: Chapter 5

6.1 *The Distribution of the Regression Model Estimator*

According to Theorem (4.9), if we have a random vector **a** such that the variance of **a** is known, V(**a**) = **C,** lets say, then we can deduce the variance of any linear combination of **a**. Using the matrix **D**′ to create a set of linear combinations, we would have, in that case, V(**D**′**a**) = **D**′**CD**. We can use this key theorem to deduce the variance of $\hat{\boldsymbol{\beta}}$, the vector of parameter estimates from the regression model, i. e.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Looking at the formula for $\hat{\boldsymbol{\beta}}$, we see that we can apply the theorem with **y** playing the role of the random vector "**a**", and the premultiplying matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ in its Oscar winning performance as "**D**", creates k linear combinations from **y**. We know the variance of **y**,

$$V(\mathbf{y}) = V(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = V(\mathbf{e}) = \sigma^2\mathbf{I}$$

since **y** must have the same variance as **e**. This is so because adding a constant to a random vector does not change the variance of that vector, as is pointed out in Theorm (4.8). Given that, we can apply the theorem of Equation (4.9) such that

$$V(\hat{\boldsymbol{\beta}}) = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma^2\mathbf{I}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad (6.1)$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

To get to the last line we have used a variety of theorems from Chapter 1, including the associative property of scalar multiplication [Theorem (1.29)], and the fact that if $\mathbf{A} = \mathbf{A}'$, then $\mathbf{A}^{-1} = (\mathbf{A}^{-1})'$ which is presented in Equation (1.40). Now that we have a formula for the variance of $\hat{\boldsymbol{\beta}}$, we are getting closer to being able to make inferences about $\boldsymbol{\beta}$, the population value. Of course we are interested in the population, not just the particular sample that we happened to have observed. To make the leap from the sample to the population we need to talk about the probability distribution of $\hat{\boldsymbol{\beta}}$. Another very important theorem about linear combinations comes next. Lets assume we have a n by 1 random vector **a** and a constant vector **b**′. Then

Central Limit

$$_1\mathbf{b}'_n\ {}_n\mathbf{a}_1 \rightarrow \text{normality as} \qquad (6.2)$$

$$n \rightarrow \infty .$$

What this *Central Limit* theorem states is that a linear combination of a random vector tends towards normality as n, the number of elements in that vector increases towards infinity. In practice, n need only get to about 30 for this theorem to apply. What's more, the theorem in no

way depends on the distribution of the random vector **a**. To take one extreme example, **a** might contain a series of binary values; 0's or 1's; and the theorem would still apply! Turning back to the least squares estimator, $\hat{\boldsymbol{\beta}}$, if we have a sample size more than 30, we can be fairly confident that $\hat{\boldsymbol{\beta}}$ will be normally distributed, even if the error vector **e**, and hence **y**, are not normally distributed. We can therefore conclude that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\,. \tag{6.3}$$

It is now time to use a distribution that is applicable when the sample size is less than 30, the *t-distribution* (more information can be seen in Section 4.6). Consider the normally distributed scalar q, that is $q \sim N[E(q), V(q)]$. In that case the ratio

$$\frac{q - E(q)}{\sqrt{\hat{V}(q)}} \sim t_{df}\,. \tag{6.4}$$

The subscript df on the *t* represents the degrees of freedom for the *t*-distribution, that is the effective number of observations used to estimate V(q) using $\hat{V}(q)$. More specifically, in the case of a particular element of $\hat{\boldsymbol{\beta}}$, say $\hat{\beta}_i$, we would have

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{V}(\hat{\beta}_i)}} \sim t_{n-k}\,. \tag{6.5}$$

We have already determined $V(\hat{\beta}_i)$ in Equation (6.1). In order to refer to this variance better, let us define

$$\mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1} = \{d^{ij}\}.$$

The superscript notation, used with the element d^{ij}, is often used to describe the elements of the inverse of a matrix. Note that d^{ii} is the ith diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Now we are in a position to say that

$$V(\hat{\beta}_i) = \sigma^2 \cdot d^{ii}\,. \tag{6.6}$$

All that remains to construct our *t* is to figure out how to estimate σ^2. This is done using

$$\hat{\sigma}^2 \equiv s^2 = \frac{SS_{Error}}{n-k} = \frac{\sum_i^n e_i^2}{n-k} \quad \text{so that} \tag{6.7}$$

$$\hat{V}(\hat{\beta}_i) = s^2 \cdot d^{ii}\,. \tag{6.8}$$

Instead of using Equation (6.7) to calculate s^2, we can also use the covariance approach (see Equation (5.25):

$$s^2 = s_{yy} - \mathbf{s}_{yx}\mathbf{S}_{xx}\mathbf{s}_{xy}.$$

In addition to being the empirical estimate of the variance of the e_i, s^2 is also the variance of $\mathbf{y} \mid \mathbf{X}$, that is, $\mathbf{y}$ conditional on the observed values of $\mathbf{X}$.

6.2 *A 1 - α Confidence Interval*

Finally, we are ready to make statements about the population values of $\hat{\boldsymbol{\beta}}$. There are two broad ways of doing this. The first, which will be given immediately below, is called a *confidence interval*. The second will be covered in the next section and involves all-or-nothing decisions about hypotheses. A 1 - α confidence interval for the element $\hat{\beta}_i$ is given by

$$\hat{\beta}_i \pm t_{\alpha/2,n-k}\sqrt{s^2 d^{ii}} \tag{6.9}$$

which means that

$$\Pr\left[\hat{\beta}_i - t_{\alpha/2,n-k}\sqrt{s^2 d^{ii}} \le \beta_i \le \hat{\beta}_i + t_{\alpha/2,n-k}\sqrt{s^2 d^{ii}}\right] = 1-\alpha, \tag{6.10}$$

where $t_{\alpha/2,\,n-k}$ is the tabled *t*-statistic with n - k degrees of freedom such that $\Pr(t \ge t_{\alpha/2}) = \alpha/2$. The upshot is that, with a probability of 1 - α, we can capture the population value of a parameter of interest between the minus and plus values of the confidence interval. The benefit of this procedure is that we can pick α a priori according to our tolerance for risk. Of course picking a smaller value of α (which reduces the risk of missing the target, β_i) implies a larger value of *t* in the formula which in turn expands the distance between the left and right end points of the interval.

Despite the elegance of confidence intervals, marketers do not usually use them. Marketing theory rarely provides us with enough information to motivate us to look at particular values of the β_i. At best, it seems our theories may be capable of letting us intuit the sign of β_i. We can then decide if we were right about our intuition using a yes or no decision, a procedure that we will now address.

6.3 *Statistical Hypothesis Testing*

Questions about marketing theory, as well as practitioner issues, that are explored using samples, are often solved through the use of *statistical hypothesis testing*. For example, we might be interested in testing the hypothesis

$$H_0: \beta_i = c$$

where c is a constant suggested by some *a priori* theory. It is important to note that the entire logical edifice that we are going to build in this section is based on the presumption that this hypothesis was indeed specified a priori, that is to say, specified before the researcher has looked at the data. In that case we need to create a mutually exclusive hypothesis that logically includes all possible alternative hypotheses. Thus, between the two hypotheses we have exhaustively described the outcome space; all outcome possibilities have been covered. Given the hypothesis above, the alternative must be

$$H_A: \beta_i \neq c.$$

We need to acknowledge that the two hypotheses are not symmetric. For one thing, Ho is specific while H_A is more general. You will note that H_0 is always associated with an equality. For another thing, the two sorts of mistakes that we can make, namely, believing in H_0 while H_A is actually true; vs. believing in H_A while H_0 is true; are not symmetric. Part of the definition of H_0 is that it is the hypothesis that we will believe in by default, unless the evidence is overwhelmingly against it. In some cases we can define H_0 for its "safety." That is, if we have two mutually exclusive hypotheses, and falsely believing in one of them, even though the other is true, is not so damaging or expensive, we would want to pick that one as H_0.

We now need to summarize the evidence for and against H_0 and H_A. Here is where the *t* statistic comes in. We will assume that H_0 is true. In that case,

$$\hat{t} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{V}(\hat{\beta}_i)}} = \frac{\hat{\beta}_i - c}{\sqrt{s^2 d^{ii}}} \sim t_{n-k}. \tag{6.11}$$

We can now evaluate the probability of this evidence assuming that H_0 is true by simply looking up the probability of $\hat{t}$ based on the *t*-distribution. Specifically, we reject H_0 if

$$|\hat{t}| > t_{\alpha/2, n-k}, \tag{6.12}$$

where $t_{\alpha/2, n-k}$ is the tabled *t*-statistic with n - k degrees of freedom such that $\Pr(t \geq t_{\alpha/2}) = \alpha/2$. The value α can once again be chosen *a priori* according to one's tolerance for the risk of falsely rejecting H_0, an error often referred to as being of *Type I*. The value α is divided in two simply because H_A has two tails, that is to say, it is the nature of H_0 that it can be wrong in either of two directions.

In some sorts of hypotheses we do not need to divide α by two. If we have H_0: $\beta_i \geq c$, which implies an alternative of H_A: $\beta_i < c$, there is only one direction or tail in which H_0 can be wrong. In that case we reject H_0 if

$$\hat{t} > t_{\alpha, n-k}. \tag{6.13}$$

The inequality obviously reverses direction if H_0 involves a "≤". Note that one way or the other, H_0 allows the possibility of an equality. The logic of hypothesis testing is based on H_0. It is the only hypothesis being tested. Rejecting H_0 we learn something, we can make a statement about the population. Otherwise we have simply failed to reject it and we must leave it at that.

Generally speaking, those writing articles for marketing journals tend to automatically pick α = .05. It's a social convention, but the arbitrariness of ".05" should not obscure the value we get out of picking some value a priori. In some practitioner applications the two possible types of errors can be assigned a monetary value and the choice of α can be optimized.

6.4 *More Complex Hypotheses and the t-statistic*

It is possible to look at more complex questions, for example is $\beta_1 = \beta_2$? We will write the question as a linear combination of the $\boldsymbol{\beta}$ vector:

$$H_0 : \mathbf{a}'\boldsymbol{\beta} = c$$

$$H_0 : \begin{bmatrix} 0 & 1 & -1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \cdots \\ \beta_{k*} \end{bmatrix} = 0.$$

We can create a *t*-test using the same technique as before as long as we can figure out the denominator of the *t*. The theorem we discussed at the beginning of the chapter, Theorem (4.9) which lets us derive the variance of a linear combination of a random variable can guide us once again:

$$\begin{aligned} V(\mathbf{a}'\hat{\boldsymbol{\beta}}) &= \mathbf{a}'V(\hat{\boldsymbol{\beta}})\,\mathbf{a} \\ &= \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{a}. \end{aligned} \tag{6.14}$$

By substituting the empirical estimate, s^2 for the population value σ^2, we get the formula for the *t* that lets us test the linear hypothesis H_0 against the alternative, H_A: $\mathbf{a}'\boldsymbol{\beta} \neq c$

$$\hat{t} = \frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - c}{\sqrt{s^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}}. \tag{6.15}$$

As before, we would reject H_0 if $|\hat{t}| > t_{\alpha/2,n-k}$.

We might note that the basic *t*-test discussed in the previous section to test H_0: $\beta_i = 0$ is a special case of this procedure with $\mathbf{a}' = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$. In general, if you can quantify a hypothesis as a single linear combination, so that the right hand side is a scalar and there is just one equal sign, you can test it with a *t*-test. But we can test even more complex hypotheses than these, and that is the subject of the next section, Section 6.5.

6.5 *Multiple Degree of Freedom Hypotheses*

We will now look at more complicated hypotheses that require more than a single linear combination. Where before our hypothesis was represented in $\mathbf{a}'$, now we will have a series of hypotheses in the q rows of the hypothesis matrix **A**. We can simultaneously test all q of these hypotheses,

$$H_0 : \mathbf{A}\boldsymbol{\beta} = {}_{q}\mathbf{c}_{1}$$

$$H_0 : \begin{bmatrix} \mathbf{a}'_{1.} \\ \mathbf{a}'_{2.} \\ \cdots \\ \mathbf{a}'_{q.} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} c_1 \\ c_2 \\ \cdots \\ c_q \end{bmatrix} \tag{6.16}$$

As an example, suppose we wanted to simultaneously test that $\beta_2 = 0$, and that $\beta_3 = 0$, or more concisely, that $\beta_2 = \beta_3 = 0$. We can use an **A** matrix as below,

$$H_0 : \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The flexibility of linear hypotheses cannot be exaggerated. Suppose we want to test that a set of β coefficients are equal; $\beta_1 = \beta_2 = \beta_3$. That can be coded into the **A** matrix as

$$H_0 : \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

To test these sorts of hypotheses, we will be using the *F distribution*, which is more general than the *t*. In fact, an F with one degree of freedom in the numerator is equivalent to a *t* squared. (This is briefly discussed in Section 4.7.) An F is a ratio of variances. Under the null hypothesis, both the numerator and the denominator variances measure the same thing so that the average F is one. In the case of the linear hypothesis H_0: $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, the numerator is the variance attributable to the hypothesis. In this context the variance is called a mean square - in other words it is an average sum of squares. To calculate the sum of squares that will be used for this mean square, we have:

$$SS_H = (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})' \left[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\right]^{-1} (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}). \tag{6.17}$$

Since $\boldsymbol{\beta}$ is a column vector, and this is a single quadratic form, SS_H is a scalar. For this to work the **A** matrix, which is q by k, has to have q independent rows, and certainly q must be less than or equal to k. Otherwise, the matrix within the brackets will not be capable of being inverted. Given that **A** has q independent rows, we can set up the ratio

$$\frac{SS_H / q}{SS_{Error} / n - k} \sim F_{q,n-k} \tag{6.18}$$

which can be used to test the hypotheses embodied in the **A** matrix.

Typically a variance is an average sum of squares divided by "n - 1" which represents the degrees of freedom of that variance. In this case, in the numerator, the average is being taken over the q rows of **A**. In other words, the number of observations - the degrees of freedom - is q. The denominator, which the reader should recognize as the variance of the e_i, called s^2, has n – k degrees of freedom. (We remind you that k represents the number of other parameters estimated in the regression model. We have already estimated values for the $\boldsymbol{\beta}$ vector.) leaving n – k observations for estimating s^2.

6.6 *An Alternative Method to Estimate Sums of Squares for an Hypothesis*

Let us return to one of the multiple degrees of freedom hypotheses we looked at above,

$$H_0 : \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

We are hypothesizing that two of the betas are zero, which implies that the independent variables associated with them vanish from the regression equation, being multiplied by zeroes. Lets call the model that is missing x_2 and x_3 the "Restricted Model." We could calculate the Sum of Squares Error for this model and compare it to the usual Sum of Squares Error. The difference, illustrated below, provides an alternative way of assessing the hypothesis:

$$SS_H = SS_{Error}\,(\text{Restricted Model}) - SS_{Error}\,(\text{Full Model})$$

Since the restricted model has fewer variables, it's SS_{Error} cannot be less than the SS_{Error} for the full model, thus SS_H must be positive; it is after all a sum of squares, so it had better be positive!

6.7 *The Impact of All the Independent Variables*

We often wonder if any of our independent variables are doing anything at all, if between them, we are achieving any prediction or explanation of the dependent variable. We can express this question using the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k^*} = 0\,. \tag{6.19}$$

The only β value missing from the hypothesis is β_0, which is usually not of any theoretical importance. The hypothesis asks if we can get any additional prediction, above and beyond the mean which is represented by β_0. The F given below,

$$\hat{F} = \frac{SS_{Error}\,(\text{Restricted to } \beta_0) - SS_{Error}\,(\text{Full})\,/\,k^*}{SS_{Error}\,(\text{Full})\,/\,n-k} \tag{6.20}$$

can be compared to the tabled value of $F_{\alpha,\,k^*,\,n-k}$. We can also summarize the predictive power of all the independent variables (except x_0) using *Big R Squared*, also known as the *squared multiple correlation* or SMC, shown below:

$$R^2 = \frac{SS_{Error}(\text{Restricted to } \beta_0) - SS_{Error}(\text{Full})}{SS_{Error}(\text{Full})}. \tag{6.21}$$

Now we will look at some alternative formulae for these Sums of Squares for Error. For example,

$$SS_{Error}(\text{Restricted to } \beta_0) = \sum_i (y_i - \bar{y})^2 \quad \text{and}$$

$$SS_{Error}(\text{Full}) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2 .$$

Using these terms, we can say that

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \text{ or in words} \tag{6.22}$$

Corrected SS = SS Due to Real Independent Variables + SS Error.

We can prove this by looking at the definition of SS_{Error}:

$$\mathbf{e'e} = \mathbf{y'y} - \hat{\boldsymbol{\beta}}\mathbf{X'X}\hat{\boldsymbol{\beta}}$$

$$= \mathbf{y'y} - \hat{\mathbf{y}}'\hat{\mathbf{y}}.$$

By rearranging we have

$$\mathbf{y'y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e'e}$$

$$\mathbf{y'y} - n\bar{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{y} + \mathbf{e'e}$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

This allows us to restate R^2 as

$$R^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

$$= 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

$$= \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \quad .$$

In summary, R^2 summarizes the proportion of the corrected Sum of Squares, and of the variance, of y which is explained by each of the independent variables, $x_1, x_2, \ldots, x_{k^*}$. The hypothesis H_0: $\rho^2 = 0$ (note that rho, ρ, is the Greek equivalent to r) is equivalent to the hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_{k^*} = 0$.

6.8 *Generalized Least Squares*

There are many circumstances where we cannot believe the Gauss-Markov assumption. Suppose for example that the variance of the errors is not $\sigma^2\mathbf{I}$ but rather follows some more general form, $\sigma^2\mathbf{V}$ where **V** is a symmetric matrix. If **V** is diagonal, the technique of this section is called *weighted least squares* or WLS. If **V** is symmetric, it is called *generalized least squares*, or GLS. Of course, if the elements of **V** are not known, we would run out of degrees of freedom trying to estimate the elements of both $\boldsymbol{\beta}$ and **V**. But in many cases, we have an a priori notion of what **V** should look like. For example, we can take advantage of the fact that the variance of the population proportion π is known and is in fact equal to $\pi(1 - \pi)/n$. If our dependent variable consists of a set of proportions, we can modify the Gauss-Markov assumption accordingly and perform weighted least squares. Instead of minimizing $\mathbf{e}'\mathbf{e}$, we minimize

$$\mathbf{f} = \mathbf{e}'\mathbf{V}^{-1}\mathbf{e}, \tag{6.23}$$

where the diagonal elements of V consist of the values $\pi(1 - \pi)/n$ for appropriate to each observed proportion. We can look at this technique as minimizing the sum of squares for a set of transformed errors. The transformed errors have constant variance and therefore are appropriate for the Gauss-Markov assumption. Our estimate of the unknowns becomes

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \, . \tag{6.24}$$

We can estimate σ^2 using

$$s^2 = \frac{SS_{Error}}{n - k}$$

where

$$SS_{Error} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

We can construct *t*-statistics that allow us to test hypotheses of the form

$$H_0: \beta_i = 0$$

using the ith diagonal element of $s^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ in the denominator to create a *t*. One can also test one degree of freedom hypotheses such as

$$\mathbf{a}'\boldsymbol{\beta} = c$$

using

$$\hat{t} = \frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - c}{s^2\mathbf{a}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{a}}$$

and for more complex hypotheses of the form

$$H_0: \mathbf{A}\boldsymbol{\beta} - \mathbf{c} = \mathbf{0}$$

we use

$$SS_H = (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{A}]^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})$$

to construct an F ratio numerator, with s^2 in the denominator.

This result is discussed in more detail in Section 17.4.

6.9 *Symmetric and Idempotent Matrices in Least Squares*

Define $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and define $\mathbf{M} = \mathbf{I} - \mathbf{P}$, i. e. $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Now recall Equation (5.21) for the SS_{Error}:

$$\begin{aligned}
\mathbf{e}'\mathbf{e} &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \mathbf{y}'\mathbf{I}\mathbf{y} - \mathbf{y}'\mathbf{P}\mathbf{y} \\
&= \mathbf{y}'[\mathbf{I} - \mathbf{P}]\,\mathbf{y} \\
&= \mathbf{y}'\mathbf{M}\mathbf{y}.
\end{aligned} \tag{6.25}$$

What this tells us is that the SS_{Error} is a quadratic form, with the matrix **M** in the middle. The $SS_{Predicted}$ is a quadratic form also, with **P** in the middle,

$$\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{y}$$

and as we might imagine, the raw total sum of squares of the dependent variable is a quadratic form, with the identity matrix in the middle:

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{I}\mathbf{y}.$$

So now we have some relationships among SS_{Total}, $SS_{Predictable}$ and SS_{Error}, namely

$$SS_{Total} = SS_{Predictable} + SS_{Error}$$

$$\mathbf{y}'\mathbf{I}\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{y} + \mathbf{y}'\mathbf{M}\mathbf{y} \text{ and} \tag{6.26}$$

$$\mathbf{I} = \mathbf{P} + \mathbf{M}. \tag{6.27}$$

At this point we might note that the Identity matrix **I** is of full rank (Section 3.7), that is to say, $|\mathbf{I}| \neq 0$, but both **P** and **M** are not with **P** having rank k and **M** rank n - k, the same as their degrees of freedom.

What's more, **P** transforms **y** into $\hat{\mathbf{y}}$, and **M** transforms **y** into **e** as can be seen below:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\,\mathbf{y} \\ &= \mathbf{P}\mathbf{y}\end{aligned} \tag{6.28}$$

and

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= \mathbf{M}\mathbf{y}\end{aligned} \tag{6.29}$$

So that we can think of **P** as the *prediction transform* or *prediction operator*, that is, a set of linear combinations that transform **y** into $\hat{\mathbf{y}}$, while **M** is the *error transform* or *error operator* that transforms **y** into **e**. These matrices have some even more unusual properties, namely:

Symmetry $\mathbf{M} = \mathbf{M}'$, $\mathbf{P} = \mathbf{P}'$ (6.30)

Idempotency $\mathbf{MM} = \mathbf{M}$, $\mathbf{PP} = \mathbf{P}$, (6.31)

and also,

$$_1\mathbf{1}'_n \; _n\mathbf{M}_n = _1\mathbf{0}'_n$$

$$_1\mathbf{1}'_n \; _n\mathbf{P}_n = _1\mathbf{0}'_n \quad \text{and} \tag{6.32}$$

$$\mathbf{PM} = _n\mathbf{0}_n.$$

More details of on the importance of M and P can be found in Section 4.5. In summary, since

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{I}\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{y} + \mathbf{y}'\mathbf{M}\mathbf{y},$$

we can show that these sums of squares components are distributed as Chi Square.

References

Graybill, Franklin (1976) *Theory and Application of the Linear Model*. Belmont, CA: Wadsworth.

Sawyer, Alan G. and J. Paul Peter (1983) "The Significance of Statistical Significance Tests in Marketing Research," *Journal of Marketing Research,* 20 (May), 122-33.

Chapter 7: The Analysis of Variance

Prerequisites: Chapter 6

7.1 *History and Overview of ANOVA*

The analysis of variance is often used to test for group differences – very frequently different groups of consumers who have been exposed to various *treatments*. The word treatment obviously makes reference to the early days of the technique from biology early in the 20th century. In the context of marketing, a classic and simple example might involve different ads viewed by the different groups. Of course ANOVA is applicable to analyses of pre-existing groups as well.

The historical roots of ANOVA go back long before the existence of computers and before text writers acknowledged that the regression technique of Chapters 5 and 6, and ANOVA, are basically one and the same. Of course, today, all the major statistical packages compute ANOVA as a special case of regression. And understanding ANOVA in this way will add to the student's intuition about what is going on. However, there are at least two different ways of notating ANOVA: an older method that relied on calculating machines and that uses multiple subscripts on the dependent variable, and the newer way that is optimized for computer calculation that uses one subscript as the observations are stacked in the vector **y**. In what follows we will offer a brief review of the older notation while demonstrating how it relates to the newer regression-centric view.

In what follows we will also assume that we have some sort of qualitative variable that divides the population into A groups indexed by a = 1, 2, ···, A. The observations from these groups might be represented as y_{ia}, that is, observation i from group a. A pictorial representation of the situation might look like the following

Group 1	Group 2	...	Group A
y_{11}	y_{12}		y_{1A}
y_{21}	y_{22}		y_{2A}
⋯	⋯		⋯
$y_{n_1 1}$	$y_{n_2 2}$		$y_{n_A A}$

You can see that the second subscript is indexing group membership while the first keeps track of the individual within that group. Further, in group a, the sample size is n_a with that observation being the last case in group a. This is known as a *one-way analysis of variance*, since there is but a single qualitative variable that identifies group membership. The traditional test of the null hypothesis involves the population means and whether they are all equal, viz.

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_A . \tag{7.1}$$

In general, we would estimate the population mean μ_a using the sample mean $\hat{\mu}_a = \bar{y}_{\cdot a}$. The subscript for the $\bar{y}_{\cdot a}$, the "·a" is taken from Equation (1.2) and is now holding the place of the eliminated first subscript in the data, the one that tracks the individual observation. Remaining with the older tradition, we say that our model is

$$y_{ia} = \mu + \alpha_a + e_{ia}, \tag{7.2}$$

with μ being the overall mean, and the α_a quantifying the impact of group membership. The e_{ia} represent error in the model, and in this case we can say that it is an error particular to group a. The problem is that we have exactly A unique groups – and A values of $\bar{y}_{\cdot a}$ in our data – but we have A + 1 parameters. That is, there are A α_a plus one μ. We need to restrict the α_a in some way. This problem is related to the idea that in the statement of the null hypothesis in Equation (7.1), there are A – 1 equal signs, not A of them. We are not interested in the levels of the group means per se, but in the differences between the levels of the group means. It turns out there are at least three popular ways to parameterize this model (of course there are an infinite number of ways to do it in general). The first one, covered next, is called *effect coding*.

7.2 *Effect Coding*

One thing we can do is impose the restriction

$$\sum_{a}^{A} \alpha_a = 0, \tag{7.3}$$

for example by setting $\alpha_A = -\alpha_1 - \alpha_2 - \cdots - \alpha_{A-1}$. The α_a represent the <u>effect</u> of being in group a:

$$\alpha_a = \bar{y}_{\cdot a} - \bar{y}_{\cdot\cdot} \tag{7.4}$$

where $\bar{y}_{\cdot\cdot}$ is clearly equivalent to μ.

At this time, let us think about how this model, as parameterized above, relates to regression. In the regression model $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$, the qualitative independent variable must be represented somehow using the columns of $\mathbf{X}$. The α_a must end up in the $\boldsymbol{\beta}$ matrix, or at least A – 1 of them must do so. We can, as we saw above, solve for the last one by subtraction. To illustrate how to implement *effect coding* lets say we have A = 4 groups. We do not have to show all of the subjects in all of the groups since the model for all of the subjects within each group must be identical. It will suffice to show the model for the i-th subject in each group. To the extent that any two members of the same group do not have the same score, this contributes to the error term. Now, our model will be

$$\begin{bmatrix} \hat{y}_{i1} \\ \hat{y}_{i2} \\ \hat{y}_{i3} \\ \hat{y}_{i4} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}. \tag{7.5}$$

It is worth contemplating the columns of $\mathbf{X}$ for a bit. The first one is clearly just the classic y-intercept, just as it has always been in Chapters 5 and 6. The last three columns code for group membership. The first vector coding for groups has a plus one for group 1 and a -1 for the last group. Zeroes appear in every other row of that column. The second group membership vector repeats the pattern but the plus one goes against group two. Finally, the last vector has a one in the next to last position, a minus one in the last position and zeroes elsewhere. To summarize, each column $x_{\cdot j}$ (j = 1, 2, ···, A-1) gets a 1 for group j, a negative 1 for group A, and everything else is null. Writing out the model in scalar terms reveals

$$\hat{y}_{i1} = \beta_0 + \beta_1,$$

$$\hat{y}_{i2} = \beta_0 + \beta_2,$$

$$\hat{y}_{i3} = \beta_0 + \beta_3 \quad \text{and}$$

$$\hat{y}_{i4} = \beta_0 - \beta_1 - \beta_2 - \beta_3 .$$

The null hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

is mathematically equivalent to

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 .$$

While the proof of this equivalence will be left to the interested reader, one can see that both statements have three equalities. Using the methods of Chapter 6, we can set up the hypothesis matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

which having three rows, provides an overall three degree of freedom test of no mean differences. Individual one degree of freedom tests for any of the β_j may or may not be of interest. H_0: $\beta_j = 0$ is equivalent to H_0: $\mu_j - \mu = 0$, that is, that there is no significant effect of being in group j.

7.3 *Dummy Coding*

In our model,

$$y_{ia} = \mu + \alpha_a + e_{ia}, \tag{7.6}$$

there are multiple ways to resolve the ambiguities and identify the model. We now cover the second one in which we impose the restriction

$$\alpha_A = 0 \tag{7.7}$$

which then implies that

$$\mu = \bar{y}_{\cdot A} \text{ and}$$

$$\alpha_a = \bar{y}_{\cdot a} - \bar{y}_{\cdot A} .$$

The coding for the design matrix looks like this:

$$\begin{bmatrix} \hat{y}_{i1} \\ \hat{y}_{i2} \\ \hat{y}_{i3} \\ \hat{y}_{i4} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

The columns of **X** are often called dummy variables since each value is either a '1' or a '0'. This means that

$$\hat{y}_{i1} = \beta_0 + \beta_1,$$

$$\hat{y}_{i2} = \beta_0 + \beta_2,$$

$$\hat{y}_{i3} = \beta_0 + \beta_3 \quad \text{and}$$

$$\hat{y}_{i4} = \beta_0.$$

You can see that column $\mathbf{x}_{\cdot j}$ gets a '1' for group j, j = 1, 2, ···, A - 1. Everything else gets a '0'. As before, H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ tests H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, and we can construct the **A** hypothesis matrix as above in equation (7.5). Test of individual β_j values are probably not interesting since H_0: $\beta_j = 0$ is equivalent to H_0: $\mu_j - \mu_A = 0$. However, this might be interesting if the last group, group A, is some sort of control group and the researcher wants to compare some of the other groups to the last one.

Note that both systems of coding lead to the same 3 degree of freedom F with the same value. What varies is how these three degrees of freedom are partitioned. And now we look at the final method of partitioning group effects, orthogonal coding.

7.4 *Orthogonal Coding*

In the previous two methods of coding, effect and dummy coding, the columns of **X** are correlated which is to say they are not orthogonal, a concept defined in Equation (1.17). In this section we describe a method of coding the design matrix in such a way that **X′X** is a diagonal matrix. Of course this means that the columns of **X** are all mutually orthogonal, meaning that the inner product is zero. There are very many ways of doing this, but here is one simple scheme that can be used to create orthogonal columns in **X**:

$$\mathrm{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 0 & 2 & -1 \\ 1 & 0 & 0 & 3 \end{bmatrix}.$$

The pattern should be clear - column j has j '-1's and one 'j'. Here we see that H_0: $\beta_1 = 0$ is equivalent to H_0: $\mu_1 = \mu_2$; H_0: $\beta_2 = 0$ is equivalent to H_0: $(\mu_1 + \mu_2)/2 = \mu_3$; and H_0: $\beta_3 = 0$ is equivalent to H_0: $(\mu_1 + \mu_2 + \mu_3)/3 = \mu_4$.

One can modify the scheme to test certain planned comparisons of interest. Suppose we had planned a priori to test H_0: $\mu_2 = \mu_3$. We can set the second column of **X** to embody this comparison:

$$X = \begin{bmatrix} 1 & 0 & 2 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & 0 & 0 & 3 \end{bmatrix}$$

which the reader can see as effectively identical to the example immediately above, but changing the order of the rows. At this point we need only test the hypothesis that $\beta_1 = 0$.

Now suppose we wish to compare groups 1 and 2 against 3 and 4, i.e. that $\frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}$. We can use **X** as below:

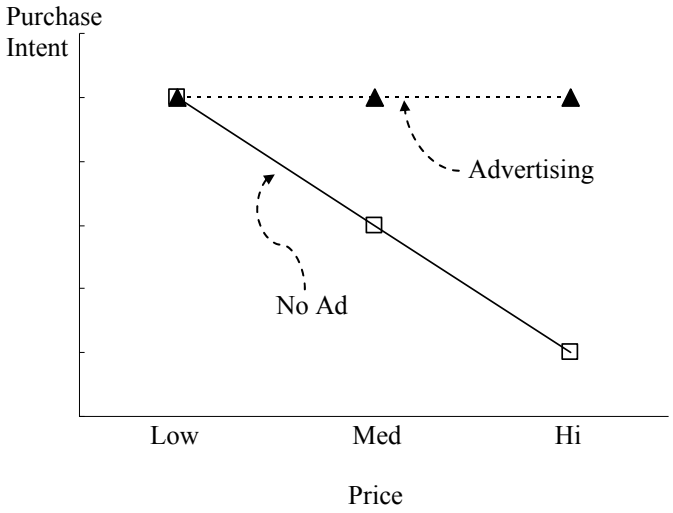
$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Here we can test our hypothesis using β_1. The pattern of signs in the second column of X (the column pertaining to β_1) allows you to interpret the sign of β_1. If β_1 is positive it means that the first two means are greater than the second two.

Note that in all the cases we have discussed in this section, we have orthogonal columns of **X**. This leads to an ease of interpretation of the β's.

7.5 *Interactive Effects*

In many cases in marketing the impact of one independent variable depends on the specific values of another independent variable. For example, we might find that as price increases, consumer purchase intention is reduced, except when there is the presence of advertising. This is illustrated in the hypothetical interaction plot below:



An interaction limits our ability to generalize. If you were to summarize the impact of Price on Purchase Intent, you would have to take into account the value of the other independent variable, Advertising. By the same token, if you were to try to describe what effect Advertising has on Intent, you would have to pull Price into the explanation. An interaction is characterized by non-parallel lines in an interaction plot, as is shown above. Interactions of many forms are possible, but the linear model can subsume any interactive effect by including columns in the design matrix **X** which consist of the products of other columns of **X**. To see this, look at the design matrix pictured below:

$$\begin{bmatrix} \hat{y}_{iHA} \\ \hat{y}_{iHN} \\ \hat{y}_{iMA} \\ \hat{y}_{iMN} \\ \hat{y}_{iLA} \\ \hat{y}_{iLN} \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 0 & 2 & 0 & -2 \\ 1 & 1 & 0 & 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}$$

Ad: A vs. N

Price: H, M, L

$x_{\cdot 4} = x_{\cdot 1} x_{\cdot 2}$

$x_{\cdot 5} = x_{\cdot 1} x_{\cdot 3}$

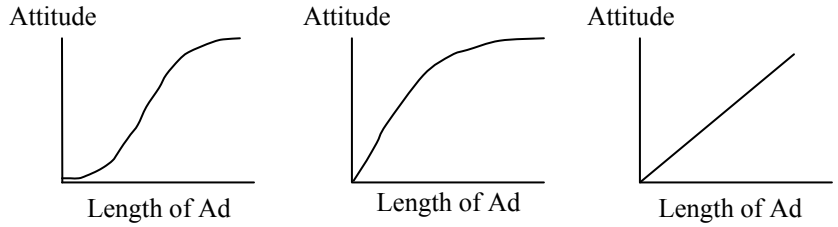
The subscripts on the dependent variable values run from L to M to H (low, medium and high) to index the level of the price variable and from A to N to indicate advertising vs. no-advertising. Column 0 of the **X** matrix codes for the usual intercept term. Column 1 uses orthogonal coding to register the difference in the level of advertising, while columns 2 and 3 use orthogonal coding to track the 3 levels of Price. With three levels, Price has 2 degrees of freedom, which is to say, 2 columns in **X**. The fourth column of **X** is the product of columns 1 and 2, while the fifth column is the product of columns 1 and 3. The interaction between Price and Advertising also has 2 degrees of freedom. The reader might notice that all six columns of **X** are mutually orthogonal.

7.6 *Quantitative Independent Variables*

We can actually use the linear regression model to fit a non-linear model. Almost any quantitative function can be approximated by a polynomial of sufficiently high order. Consider the model below:

$$y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \cdots + x_i^m\beta_m + e_i \tag{7.8}$$

To make this model work, one should first deviate the x_i from the mean to avoid problems of high correlation between the columns of the **X** matrix. With a relatively small number of levels of the quantitative independent variable, you can use the method of orthogonal polynomials instead. Any function can be represented as a polynomial with sufficiently high order. A curve with one elbow can be expressed as a quadratic function, one with two elbows can be imitated with a cubic function, and so on from quartic, quintic, etc. For example, we might be concerned with the shape of the relationship between the length of an ad viewed by subjects, and their attitude towards that ad. Imagine that one group saw a 1 minute ad, another a 2 minute ad, and there were also 3 and 4 minute groups. Presuming that the ad is affective, the relationship could take on a variety of forms, such as those pictured below:



On the far right is pictured a very simple linear assumption, in the middle a curve with one elbow, and on the left a more complex curve requiring a cubic component. We might construct the design matrix as below using

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \end{bmatrix}$$

but it would be smarter to use a columns that were not so highly correlated. As mentioned above, if you column-center the linear component and then use it as a basis for creating the other columns, this will help. You can also use orthogonal polynomials (see the tables in Bock 1975 for example):

$$\mathbf{X} = \begin{bmatrix} 1 & -3 & 1 & 1 \\ 1 & -1 & -1 & -3 \\ 1 & 1 & -1 & 3 \\ 1 & 3 & 1 & -1 \end{bmatrix}.$$

One could then test the necessity of the cubic term, assuming a linear and quadratic component using a *t*-test. If that proves non-significant, one could go on and test the necessity of the quadratic term.

7.7 *Repeated Measures Analysis of Variance*

A special case of the analysis of variance occurs when we have a set of *commensurate variables*, or *commensurate measures*. The expression implies that the same scale is repeatedly applied on several measurement occasions. For example, perhaps consumers are asked to rate four brands using a particular measure. Repeated measures are multivariate in nature, meaning that there is more than one dependent variable. In the example with four brands, there would be four dependent variables. We define y_{ij} as the measurement on person i, on measure j, with i = 1, 2, ···, n and j = 1, 2, ···, p. There are two ways to treat such data. We can place all of the measurements in a matrix, **Y**, with a row for each subject and a column for each measure. This is the multivariate approach, a topic covered in Chapter 8. For now, we will note that with four brands, and p = 4, the hypothesis that the means of the four brands are equal, i.e. that the columns of **Y** have equal means, is equivalent to the hypothesis that the three columns of $\widetilde{\mathbf{Y}}$ below have means of zero. The matrix $\widetilde{\mathbf{Y}}$ is given by

$$\widetilde{\mathbf{Y}} = \mathbf{YM} \tag{7.9}$$

$$\widetilde{\mathbf{Y}} = \mathbf{Y}\begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & 1 \\ 0 & -2 & 1 \\ 0 & 0 & -3 \end{bmatrix}. \tag{7.10}$$

The hypothesis matrix **M**, when used to postmultipy the original data matrix **Y**, transforms the columns of **Y** into new columns in $\widetilde{\mathbf{Y}}$. The first new column consists of the difference between the old columns 1 and 2. The second new column in $\widetilde{\mathbf{Y}}$ is the difference between the combination of columns 1 and 2 and column 3, and so forth.

The univariate approach stacks all of the data in a single vector, called **y**, in such a way that each subject's data appears contiguously, i.e.

$$\mathbf{y}' = [y_{11} \quad y_{12} \quad \cdots \quad y_{1p} \quad y_{21} \quad y_{22} \quad \cdots \quad y_{2p} \quad \cdots \quad y_{n1} \quad y_{n2} \quad \cdots \quad y_{np}]$$

We can then say that

$$\mathrm{V}(\mathbf{y}) = \left[\begin{array}{c|c|c|c} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \hline \cdots & \cdots & \cdots & \cdots \\ \hline \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{array}\right] \tag{7.11}$$

where each $\boldsymbol{\Sigma}$ and each **0** is a p by p matrix. There are n of them, so that the entire variance matrix of **y** is np by np. That the covariance matrix of each subject, $\boldsymbol{\Sigma}$, is homogeneous or identical from

one subject to the next, is only an assumption, analogous to the assumption of homogeneity of variance of the scalar σ^2 in regular ANOVA.

To use the univariate approach to repeated measures, the variance of the transformed measures must be homogeneous and independent, that is

$$\mathbf{M}'\mathbf{\Sigma}\mathbf{M} = \sigma^2\mathbf{I} \tag{7.12}$$

where the **M** matrix is the hypothesis matrix from above, **Σ** is the p by p (in our example with four brands, four by four) covariance matrix of the original measures, and $\sigma^2\mathbf{I}$ is a scalar matrix with identical values along the diagonal (three identical values in our example with four brands). Often this assumption is called *sphericity*. If this assumption is met, we can use univariate analysis of variance as will now be described using an example.

7.8 *A Classic Repeated Measures Example*

Imagine that we have three factors including one between subjects variable that divides subjects into two groups, a within-subjects factor with three levels and a within subjects variable that has four levels. All told our design is a 2 × 3 × 4 design, with the three factors named A, B and C. We can further imagine that we have 10 subjects, and since each subject is measured 12 times (since the repeated measures part of the design, B x C, involves 12 measures), we have a total of 120 data points. The results of such an ANOVA are typically described in an ANOVA table. A table for this design could look like this;

Source of Variance	**df**	**Error Term**
A	1	S(A)
S(A)	8	-
Between-Subjects Total	9	-
B	2	S(A) · B
C	3	S(A) · C
BC	6	S(A) · BC
AB	2	S(A) · B
AC	3	S(A) · C
ABC	6	S(A) · BC
S(A) · B	16	-
S(A) · C	24	-
SA(A) · BC	48	-
Within-Subjects Total	110	-
TOTAL	119	-

The notation in the table bears some explanation. S(A) is used to represent Subjects within levels of the A factor. In other words, subjects are *nested* within groups since the same subject does not appear in more than one group. In contrast, Subjects are *crossed* with the two repeated measures: B and C. In addition, the factor Subjects is a random effect. This means that the "levels" of Subjects were randomly sampled from some larger population to which we would like to generalize our results. In contrast, A, B and C are fixed effects whose levels are chosen for their inherent interest to the experimenter, and hopefully for that person, the reviewers.

You might note that the correct error term for the grouping factor is Subjects within groups. The correct error term for any repeated measures factor is that factor by Subjects interaction. In general terms, consider a purely within-subject effect, *w*, a purely between-subject effect, *b*, and

their interaction, *wb*. Either *w* or *b* may be main effects, interactions, or special contrasts. The error term for *b* is Subjects nested in groups. The error term for *w* is Subjects · *w* and the error term for *wb* is also Subjects · *w*. Homogeneity of Subject variance within groups is a needed assumption, as is the spherecity of transformed measures as described above in Equation (7.12).

References

R. Darrell Bock (1975) *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill.

Box, George E. P., William G. Hunter, J. Stuart Hunter (1978) *Statistics for Experimenters*. New York: Wiley.

Glantz, Stanton A. and Bryan K. Slinker (1990) *Primer of Applied Regression and Analysis of Variance*. New York: McGraw-Hill.

Kirk, Roger E. (1982) *Experimental Design. Second Edition*. Pacific Grove, CA: Brooks/Cole.

Mendenhall, William (1968) *The Design and Analysis of Experiments*. Belmont,CA: Duxbury.

Chapter 8: The Multivariate General Linear Model

Requirements: Sections 3.4, 3.5 - 3.8, 4.3 Chapter 7

8.1 *Introduction*

The main difference between this chapter and the chapters on the General Linear Model; 5, 6 and 7; lies in the fact that here we are going to explicitly consider multiple dependent variables. Multiple dependent variables are to some extent discussed in Chapter 7 in the context of the analysis of variance. In that chapter, however, we made an assumption about the error distribution which allowed us to treat the problem as essentially univariate [see Equation (7.12)]. In this chapter, we will be dealing with multiple dependent variables in the most general way possible, namely the multivariate general linear model. Before we begin, it will be necessary to review some of the fundamentals of hypothesis testing, and then after, to introduce some mathematical details of use in this area.

8.2 *Testing Multiple Hypotheses*

In Chapter 6, we covered two different approaches to testing hypotheses about the coefficients of the linear model. In Equation (6.15) we had

$$\hat{t} = \frac{\mathbf{a}'\boldsymbol{\beta} - c}{\sqrt{s^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}}$$

that allows us to test one degree of freedom questions of the form $\mathbf{a}'\boldsymbol{\beta} = c$, while in Equation (6.18) we have the test statistic

$$\hat{F} = \frac{SS_H / q}{SS_{Error} / n - k}$$

that allows us to test multiple degree of freedom questions $\mathbf{A}\boldsymbol{\beta} = \mathbf{C}$. In the former case we have n - k degrees of freedom, and in the latter, q and n - k degrees of freedom. In that chapter we made the implicit assumption that these tests had been planned *a priori*, and that they were relatively few in number. In the case of multiple dependent variables, this second assumption becomes far less tenable. We begin by discussing a way to test hypotheses even when there are a large number of them. We then discuss the case where this large number of hypotheses might even be post hoc.

8.3 *The Dunn-Bonferroni Correction*

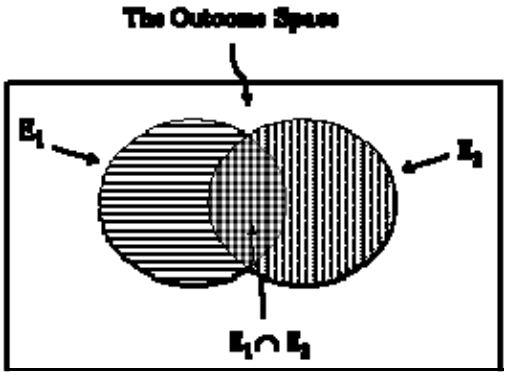
What can we do if we wish to test a large number of hypotheses, say, H_1, H_2, ···, H_r? For any particular hypothesis, we can limit the probability that we reject H_0 when it was indeed true of the population, that is we can limit

$$\Pr(\text{Type I Error on } H_i) = \alpha_i.$$

But what is the probability of at least one Type I error in a sequence of r hypotheses? To delve into this question it will be useful to utilize the notation of Set Theory, where $\cup$ symbolizes union and $\cap$ symbolizes intersection. The probability of at least one Type I error is

$$\alpha^* = \Pr(\text{Type I error on } H_1 \cup \text{Type I Error on } H_2 \cup \cdots \cup \text{Type I Error on } H_r). \qquad (8.1)$$

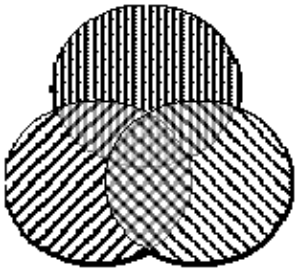
We can think of α^* as the overall α rate, the probability of at least on Type I Error. Define E_i as a Type I error event for H_i. From probability theory, with r = 2 hypotheses, lets say, the situation is illustrated below:



Two parts of the outcome space are shaded, the two parts that correspond to E_1 (a Type I Error on H_1) and E_2 (a similar result on H_2). There is some overlap, namely the part of the space comprising the intersection of E_1 and E_2. It can be shown that

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2).$$

Needless to say, one has to subtract out the $\Pr(E_1 \cap E_2)$ so that it is not counted twice when adding up $\Pr(E_1) + \Pr(E_2)$. For r = 3 hypotheses we have a diagram as below



and we can say

$$\Pr(E_1 \cup E_2 \cup E3) = \Pr(E_1) + \Pr(E_2) + \Pr(E_3) -$$

$$\Pr(E_1 \cap E_2) - \Pr(E_1 \cap E_3) - \Pr(E_2 \cap E_3) + \Pr(E_1 \cap E_2 \cap E_3).$$

Here, we needed to subtract all of the two-way intersections but then we had to add back in the third way intersection which was subtracted once too often. In any case, it is clear that the simple sum of the probabilities, $\sum_i^r \Pr(E_i)$ is an upper bound on the probability of at least one Type I Error since we have not subtracted out any of the intersecting probabilities. We can then safely say that

$$\Pr(E_1 \cup E_2 \cup \cdots \cup E_r) \le \Pr(E_1) + \Pr(E_2) + \cdots + \Pr(E_r).$$

Now of course, $Pr(E_i) = \alpha_i = \alpha$ for all i, so in that case we can state

$$\alpha^* = Pr(E_1 \cup E_2 \cup \cdots \cup E_r) \leq \sum_i^r Pr(E_i) = \sum_i^r \alpha_i$$

in which case

$$\alpha^* \leq r \cdot \alpha.$$

If we select α so that

$$\frac{\alpha^*}{r} \leq \alpha$$

we set an upper limit on our overall α. For example, with r = 10 a priori hypotheses, if I want my overall Type I rate to be $\alpha^* = .05$, I would pick $\alpha = .05/10 = .005$ for each hypothesis.

This logic is of course flexible enough to be applicable to any sort of hypotheses whether they be about factor analysis loadings, differences between groups, or tests of betas. A problem with this approach becomes apparent when r gets big. It then becomes very conservative. At that point it is reasonable to use a different logic, a logic that is also applicable to post hoc hypotheses. We now turn to that.

8.4 *Union-Intersection Protection from Post Hoc Hypotheses*

This technique, also known as the Roy-Scheffé approach, is one that protects the marketing researcher from the worst data sniffing case possible, in other words, any post hoc hypothesis. As with the Dunn-Bonferroni test, it is applicable to any sort of hypothesis testing. And as with the Dunn-Bonferroni the overall probability of at least one Type I event is

$$\alpha^* = Pr(\text{Type I error on } H_1 \cup \text{Type I Error on } H_2 \cup \cdots \cup \text{Type I Error on } H_r)$$

$$= Pr(E_1 \cup E_2 \cup \cdots \cup E_r).$$

This probability is equivalent to 1 - Pr(No Type I Events). Define the complement of E_i as $\overline{E}_i$, a non-Type I event. We can then re-express the above equation, expressed as a union, as

$$\alpha^* = 1 - Pr(\overline{E}_1 \cap \overline{E}_2 \cap \cdots \cap \overline{E}_r). \tag{8.2}$$

which is instead expressed as an intersection. A commonality to all hypothesis testing situations is that $\overline{E}_i$ occurs when the calculated value of the test statistic, $\hat{\theta}_i$, exceeds a critical value, θ_i. Perhaps θ_i is a *t*, and F, or an eigenvector of $E^{-1}H$. In any of these cases,

$$\begin{aligned} \alpha^* &= Pr(\hat{\theta}_1 < \theta_0 \cap \hat{\theta}_2 < \theta_0 \cap \cdots \cap \hat{\theta}_r < \theta_0) \\ &= 1 - Pr(\hat{\theta}_{max} < \theta_0) \end{aligned} \tag{8.3}$$

where $\hat{\theta}_{max}$ is the largest value of $\hat{\theta}$ that you could ever mine out of your data. Here is an example inspired from ANOVA. Suppose we wanted to test

$$H_0: \mathbf{c}'\boldsymbol{\mu} = 0$$

where $\boldsymbol{\mu} = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_k]'$ is the vector of population means from a one way univariate ANOVA, in other words the topic of Chapter 7 where the interest is on testing hypotheses about differences among the groups. Here we wish to be protected from

$$\hat{t}^2_{max} = (\mathbf{c}'\overline{\mathbf{y}})^2 \Big/ \frac{s^2}{n}\mathbf{c}'\mathbf{c}.$$

Picking elements of the vector **c** so as to make this *t* as large as possible leads to the Scheffé (1959) post-hoc correction. More information on post hoc (and a priori) tests among means can be found in Keppel (1973).

8.5 *Details About the Trace Operator and It's Derivative*

The trace operator was introduced in Chapter 1. To briefly review, the trace of a square matrix, say **A**, is defined as $\text{Tr}(\mathbf{A}) = \sum a_{ii}$, i.e. the sum of the diagonal elements. Some properties of $\text{Tr}(\cdot)$ follow. Assuming that **A** and **B** are square matrices we can say

Transpose $$\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{A}') \tag{8.4}$$

Additivity $$\text{Tr}(\mathbf{A} + \mathbf{B}) = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{B}) \tag{8.5}$$

Then, for **A** m · n and **B** n · m we have

Commutative $$\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA}) \tag{8.6}$$

which further implies, for **C** m · m

Triple Product $$\text{Tr}(\mathbf{ABC}) = \text{Tr}(\mathbf{CAB}) \tag{8.7}$$

In Chapter 3, we discuss the derivative of a scalar function of a vector, and a vector function of a vector. Here we want to look at the derivative of a scalar function of a matrix, that function being, of course, the trace of that matrix. To start off, note that by definition

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \frac{\partial f(\mathbf{X})}{\partial x_{12}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{1n}} \\ \frac{\partial f(\mathbf{X})}{\partial x_{21}} & \frac{\partial f(\mathbf{X})}{\partial x_{22}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{2n}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{m1}} & \frac{\partial f(\mathbf{X})}{\partial x_{m2}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{bmatrix}, \tag{8.8}$$

where $f(\mathbf{X})$ is a scalar function of the matrix **X**. Now we can begin to talk about the $\text{Tr}(\cdot)$ function which is a scalar function of a square matrix. For **A** m · m we have

$$\frac{\partial \mathrm{Tr}(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{I} \tag{8.9}$$

For **A** m · n and **B** n · m we can say

$$\frac{\partial \mathrm{Tr}(\mathbf{AB})}{\partial \mathbf{A}} = \mathbf{B}' \tag{8.10}$$

which also implies, from Equation (3.19)

$$\frac{\partial \mathrm{tr}(\mathbf{AB})}{\partial \mathbf{A}'} = \mathbf{B} \,. \tag{8.11}$$

Finally, assuming we have **A** m · m and **B** m · m,

$$\frac{\partial \mathrm{tr}(\mathbf{A}'\mathbf{B}\mathbf{A})}{\partial \mathbf{A}} = (\mathbf{B} + \mathbf{B}')\mathbf{A} \,. \tag{8.12}$$

8.6 *The Kronecker Product*

We now review the definition of the *Kronecker product,* sometimes called the *Direct product*, with operator ⊗. By definition,

$$_{\mathrm{mp}}\mathbf{C}_{\mathrm{nq}} = {}_{\mathrm{m}}\mathbf{A}_{\mathrm{n}} \otimes {}_{\mathrm{p}}\mathbf{B}_{\mathrm{q}} = \{a_{ij}\mathbf{B}\} \,. \tag{8.13}$$

For example,

$$\begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}\mathbf{B} \\ \hline a_{21}\mathbf{B} \end{bmatrix}.$$

Here are some properties of the Kronecker product. We can say that

Transpose $$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'. \tag{8.14}$$

For **A** m · n, **B** n · p and **C** p · q, it is the case that

Associative $$\mathbf{AB} \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{BC}. \tag{8.15}$$

For **A** and **B** m · n and **C** p · q,

Distributive $$(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}. \tag{8.16}$$

For $\mathbf{A}$ m · n, $\mathbf{B}$ n · p and $\mathbf{C}$ q · r and $\mathbf{D}$ r · s,

$$(\mathbf{A} \otimes \mathbf{C})(\mathbf{B} \otimes \mathbf{D}) = \mathbf{AB} \otimes \mathbf{CD}. \tag{8.17}$$

8.7 *The Vec Operator*

For a matrix $\mathbf{A}$, lets say m by n, we define

$$\text{vec}(\mathbf{A}) = \text{vec}\begin{bmatrix} \mathbf{a}'_{1\cdot} \\ \mathbf{a}'_{2\cdot} \\ \cdots \\ \mathbf{a}'_{m\cdot} \end{bmatrix} = \begin{bmatrix} \mathbf{a}'_{1\cdot} & \mathbf{a}'_{2\cdot} & \cdots & \mathbf{a}'_{m\cdot} \end{bmatrix}. \tag{8.18}$$

While other definitions of Vec(·) are possible, this one, that does so one row at a time, will prove useful to us when we start to look at the multivariate GLM. In particular, the following theorem will be quite useful. For $\mathbf{A}$ m · n, $\mathbf{B}$ n · p and $\mathbf{C}$ p · q,

$$\text{Vec}(\mathbf{ABC}) = (\mathbf{A} \otimes \mathbf{C}') \text{ Vec}(\mathbf{B}). \tag{8.19}$$

8.8 *Eigenstructure for Asymmetric Matrices*

Suppose we needed to maximize $\mathbf{x}'\mathbf{Hx}$ subject to $\mathbf{x}'\mathbf{Ex} = 1$. Then

$$f(\mathbf{x}) = \mathbf{x}'\mathbf{Hx} - \lambda(\mathbf{x}'\mathbf{Ex} - 1) \tag{8.20}$$

and to minimize we set

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{Hx} - 2\lambda\mathbf{Ex} = 0. \tag{8.21}$$

Rearranging a bit we have

$$(\mathbf{H} - \lambda\mathbf{E})\mathbf{x} = (\mathbf{E}^{-1}\mathbf{H} - \lambda\mathbf{I})\mathbf{x} = 0. \tag{8.22}$$

You will note the eigenstructure discussed in Chapter 3 is a special case of the current discussion with $\mathbf{E} = \mathbf{I}$. In our case, as $\mathbf{E}^{-1}\mathbf{H}$ is asymmetric, the eigenvectors are not orthonormal [defined in Equation (3.33)]. Instead we have the relation

$$\mathbf{E}^{-1}\mathbf{H} = \mathbf{XLX}^{-1}. \tag{8.23}$$

For symmetric matrices we have had $\mathbf{X}^{-1} = \mathbf{X}'$, but not in this case.

8.9 *Eigenstructure for Rectangular Matrices*

For completeness, we note that any m · n matrix $\mathbf{A}$ or rank r can be decomposed into the triple product

$$\mathbf{A} = \mathbf{X}\mathbf{L}^{1/2}\mathbf{V}' \tag{8.24}$$

where $\mathbf{X}$ is $m \cdot r$, $\mathbf{L}^{1/2}$ is $r \cdot r$ and $\mathbf{V}$ is $n \cdot r$. This is called *singular value decomposition*. The matrix $\mathbf{X}$ contains the *left eigenvectors* of $\mathbf{A}$ while $\mathbf{V}$ contains the *right eigenvectors* of $\mathbf{A}$. Further, $\mathbf{V}'\mathbf{V} = \mathbf{I}$ and $\mathbf{X}'\mathbf{X} = \mathbf{I}$. There are important relationships between the eigenvalues of a rectangular matrix and a cross product matrix. We have

$$\mathbf{A}'\mathbf{A} = (\mathbf{X}\mathbf{L}^{½}\mathbf{V}')\,(\mathbf{V}\mathbf{L}^{½}\mathbf{X}') = \mathbf{X}\mathbf{L}\mathbf{X}' \tag{8.25}$$

and

$$\mathbf{A}\mathbf{A}' = (\mathbf{V}\mathbf{L}^{½}\mathbf{U}')\,(\mathbf{U}\mathbf{L}^{½}\mathbf{V}') = \mathbf{V}\mathbf{L}\mathbf{V}' \tag{8.26}$$

If $\mathbf{A}$ is already symmetric then $\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}'$ so $\mathbf{X} = \mathbf{V}$.

8.10 *The Multivariate General Linear Model*

The multivariate general linear model is a straightforward generalization of the univariate case in Equation (5.3). Instead of having one dependent variable in one column of the vector $\mathbf{y}$, we have a set of p dependent variables in the several columns of the matrix $\mathbf{Y}$. The model is therefore

$$\begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} & \cdots & \hat{y}_{1p} \\ \hat{y}_{21} & \hat{y}_{22} & \cdots & \hat{y}_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{y}_{np} & \hat{y}_{np} & \cdots & \hat{y}_{np} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k^*} \\ 1 & x_{21} & \cdots & x_{2k^*} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n2} & \cdots & x_{nk^*} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{k^*1} & \beta_{k^*2} & \cdots & \beta_{k^*p} \end{bmatrix} \tag{8.27}$$

which, as you can see, implies that the number of columns of the $\mathbf{B}$ matrix match the number of columns of the $\mathbf{Y}$ matrix. Perhaps this concept is better represented using the dot subscript reduction operator (Section 1.1), which allows us to present the model as

$$[\hat{\mathbf{y}}_{\cdot 1} \quad \hat{\mathbf{y}}_{\cdot 2} \quad \cdots \quad \hat{\mathbf{y}}_{\cdot p}] = \mathbf{X}\,[\boldsymbol{\beta}_{\cdot 1} \quad \boldsymbol{\beta}_{\cdot 2} \quad \cdots \quad \boldsymbol{\beta}_{\cdot p}] \tag{8.28}$$

with each column of $\mathbf{Y}$ entering into a regression equation with the corresponding column of $\mathbf{B}$ serving as the coefficient vector. We can express the model most succinctly by using

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}. \tag{8.29}$$

Next we define the $n \cdot p$ error of prediction matrix as $\boldsymbol{\varepsilon}$, i. e.

$$\boldsymbol{\varepsilon} = \hat{\mathbf{Y}} - \mathbf{Y}$$

so that

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}. \tag{8.30}$$

8.11 *A Least Squares Estimator for the MGLM*

How do we come up with estimators for the unknowns in the **B** matrix? When Y the error **e** was only a vector, as in Chapter 5, we could pick our objective function as $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. The matrix $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ on the other hand, is not a scalar but a p · p sum of squares and cross products matrix. In this case what we do is to minimize the trace of $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ as we will now see. Our objective function is

$$f = \text{Tr}[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}] \tag{8.31}$$

which, according to Equation (8.30), can be expanded to

$$f = \text{Tr}[(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})]. \tag{8.32}$$

Factoring the product leads to four components as below;

$$f = \text{Tr}[\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{XB} - \mathbf{B}'\mathbf{X}'\mathbf{Y} + \mathbf{B}'\mathbf{X}'\mathbf{XB}].$$

But since Equation (8.5) notes that the trace of a sum is equivalent to the sum of the traces, we can now say

$$f = \text{Tr}(\mathbf{Y}'\mathbf{Y}) - \text{Tr}(\mathbf{Y}'\mathbf{XB}) - \text{Tr}(\mathbf{B}'\mathbf{X}'\mathbf{Y}) + \text{Tr}(\mathbf{B}'\mathbf{X}'\mathbf{XB}).$$

More simplification is possible. From Equation (8.4) we note that $\text{Tr}(\mathbf{B}'\mathbf{X}'\mathbf{Y}) = \text{Tr}(\mathbf{Y}'\mathbf{XB})$ and from Equation (8.7) we note that $\text{Tr}(\mathbf{Y}'\mathbf{XB})$ is equivalent to $\text{Tr}(\mathbf{BY}'\mathbf{X})$. We can now rewrite f as

$$f = \text{Tr}(\mathbf{Y}'\mathbf{Y}) - 2\text{Tr}(\mathbf{BY}'\mathbf{X}) + \text{Tr}(\mathbf{B}'\mathbf{X}'\mathbf{XB}).$$

In order to make f as small as possible, it is necessary to find the $\partial f/\partial \mathbf{B}$. Using Equations (8.10) as well as (8.12), we have

$$\frac{\partial f}{\partial \mathbf{B}} = -2\mathbf{X}'\mathbf{Y} + [\mathbf{X}'\mathbf{X} + (\mathbf{X}'\mathbf{X})']\mathbf{B}.$$

But since $\mathbf{X}'\mathbf{X}$ is symmetric, we can simplify a bit more and have

$$\frac{\partial f}{\partial \mathbf{B}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{XB}. \tag{8.33}$$

After setting Equation (8.33) equal to zero, this now leads us to the multivariate analog of the normal equations [Equation (5.7)] as below:

$$\mathbf{X}'\mathbf{XB} = \mathbf{X}'\mathbf{Y} \tag{8.34}$$

so that

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{Y} \tag{8.35}$$

Each column of $\hat{\mathbf{B}}$ has the same formula as the univariate model, i. e.

$$\hat{\boldsymbol{\beta}}_{\cdot j} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{\cdot j}.$$

8.12 *Properties of the Error Matrix* $\boldsymbol{\varepsilon}$

In order to talk about the distribution of the error matrix $\boldsymbol{\varepsilon}$, we will have to rearrange it somewhat using the Vec(·) function of Section 8.7. We will assume, in a multivariate analog to the Gauss Markov Assumption of Chapter 5, that the distribution of the n by p matrix $\boldsymbol{\varepsilon}$ is

$$\text{Vec}(\boldsymbol{\varepsilon}) \sim N({}_{np}\mathbf{0}_1, {}_n\mathbf{I}_n \otimes {}_p\boldsymbol{\Sigma}_p). \tag{8.36}$$

The Vec operator has unpacked the $\boldsymbol{\varepsilon}$ matrix, one row at a time, in other words, one consumer's data at a time. Since there are n consumers with p measurements each, the mean vector of Vec($\boldsymbol{\varepsilon}$) is np by 1. The covariance matrix for Vec($\boldsymbol{\varepsilon}$), since the latter has np elements, must be np by np. This covariance matrix has a particular structure that logically, and visually, is reminiscent of the structure we assume in the univariate case presented in Equation (5.16), that of $\sigma^2\mathbf{I} = \mathbf{I} \cdot \sigma^2$. Here, instead we have the partitioned matrix

$$\mathbf{I} \otimes \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix} \tag{8.37}$$

with each $\boldsymbol{\Sigma}$ and each null matrix $\mathbf{0}$ being p · p. The $\boldsymbol{\Sigma}$ in the ith diagonal partition represents the (homogeneous) variance matrix for observation i. The $\mathbf{0}$ in the i, jth position implies that rows i and j of $\boldsymbol{\varepsilon}$, corresponding to subjects i and j, are independent.

8.13 *Properties of the* $\mathbf{B}$ *Matrix*

It is now timely to contemplate the expectation and variance of our estimator of Equation (8.35). Before proceeding, if you wish you can review some of the rules of expectations and variance presented in Section 4.1. The expectation will be straightforward, as

$$\text{E}(\hat{\mathbf{B}}) = \text{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}]$$

which for fixed $\mathbf{X}$ and Equation (4.5) leads to

$$\text{E}(\hat{\mathbf{B}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{E}(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{B} = \mathbf{B}.$$

In order to derive the $\text{V}(\hat{\mathbf{B}})$, we will need Theorem (4.9) as well as the more recent Theorem (8.19). OK, let us proceed by noting that

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{I}.$$

Now with $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ playing the role of "**A**", **Y** playing the role of "**B**", and the p by p identity matrix **I** playing the role of "**C**", we apply Theorem (8.19) to show that

$$\text{Vec}(\hat{\mathbf{B}}) = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \otimes \mathbf{I}]\text{Vec}(\mathbf{Y})$$

Now we just need to recall that $\text{Var}[\text{Vec}(\mathbf{Y})] = \mathbf{I} \otimes \mathbf{\Sigma}$ and to apply Theorem (4.9) and take it to the bank:

$$\text{Var}[\text{Vec}(\hat{\mathbf{B}})] = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \otimes \mathbf{I}](\mathbf{I} \otimes \mathbf{\Sigma})[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{I}]. \tag{8.38}$$

Note that in the above we have taken advantage of Equation (8.14) to express

$$[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \otimes \mathbf{I}]' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{I}\ .$$

Now applying Equation (8.17) two times to Equation (8.38) we can express it as

$$\text{Var}[\text{Vec}(\hat{\mathbf{B}})] = (\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{\Sigma}. \tag{8.39}$$

8.14 *The Multivariate General Linear Hypothesis*

In Chapter 6 we looked at q degree of freedom hypotheses of the form

$$H_0\text{: } \mathbf{A\beta} - \mathbf{c} = \mathbf{0},$$

where the matrix **A** had q rows and where **0** is a q by 1 column of zeroes. In this chapter, since the **B** matrix has multiple columns of possible interest, as compared to **β** which is a column vector, we allow ourselves the possibility to test linear hypotheses about these several columns of **B**. The general form of the hypothesis is then

$$H_0\text{: } \mathbf{ABM} - \mathbf{C} = \mathbf{0}. \tag{8.40}$$

The q rows of **A** test hypotheses concerning the k independent variables. **A** is therefore q · k with q ≤ k. The ℓ columns of **M** test hypotheses about the p dependent variables. **M** is necessarily p · ℓ with $\ell \leq$ p. Next, in Section 8.15 we will look at some examples of **A** and **M**.

8.15 *Some Examples of MGLM Hypotheses*

In our first example, we have k = 3 with $\mathbf{x}_{\cdot 0}$ being the usual column of 1's, $\mathbf{x}_{\cdot 1}$ being income, and then $\mathbf{x}_{\cdot 2}$ being education. On the dependent variable side, we have p = 2 with $\mathbf{y}_{\cdot 1}$ a measure of attitude towards a particular brand and $\mathbf{y}_{\cdot 2}$ being a likelihood of purchase measure. Imagine for a moment that we want to find out if education and income, taken jointly, impact the two dependent variables. Our hypothesis matrices would then take the form as shown below,

$$\mathbf{ABM} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{12} & \beta_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In the second example, k = 1 and $\mathbf{x}_{\cdot 0}$ is the usual vector of n 1's. However, p = 4 where $\mathbf{y}_{\cdot 1}$ through $\mathbf{y}_{\cdot 4}$ are evaluations of four product concepts on a 10 point scale. In this second example, the question of interest is, "Do the product evaluations differ?" In this case, we will use the multivariate approach to repeated measures. The current approach is in contrast to the univariate approach covered in Section 7.7. Here we have

$$\mathbf{aBM} = 1 \cdot [\beta_{01} \quad \beta_{02} \quad \beta_{03} \quad \beta_{04}] \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Since there are no real independent variables, the matrix **B** is actually a row vector with only the intercepts present. In an intercept only model (see Section 5.9), the β_0 values are simply the means of the dependent variables. The **M** hypothesis matrix transforms the four variable means into three mean-differences. Thus, the hypothesis is of three degress of freedom which test for equality among the levels of the four original dependent variables.

Our example number 3 includes k = 4 with an intercept term plus three attitude variables. For dependent variables, we have p = 3 behavioral measures. Our hypothesis will be an omnibus question designed to ask whether attitude influences behavior:

$$\mathbf{ABM} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} \\ \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \end{bmatrix} \mathbf{I}.$$

Finally, in our fourth example, we have experimental data in which we had a 2 × 2 ANOVA with four groups of consumers. Half the groups saw a high price, and half a low price. Half the groups saw the presence of advertising with half seeing no advertising. There is also the potential interaction of these two factors. Two measures were $\mathbf{y}_{\cdot 1}$; an affective response and $\mathbf{y}_{\cdot 2}$; a cognitive response. The hypothesis concerns the one degree of freedom interaction between price and advertising. Does such an interaction occur for affect and cognition?

$$\mathbf{ABM} = [0 \quad 0 \quad 0 \quad 1] \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

8.16 *Hypothesis and Error Sums of Squares and Cross-Products*

In the univariate linear model, we calculate the hypothesis sum of squares, which is a scalar that corresponds to the single dependent variable. The following equation produces the sum of squares and cross products matrix for the hypothesis embodied in Equation (8.40). As such, it is the multivariate analog to the univariate version presented in Equation (6.17):

$$\mathbf{H} = (\mathbf{A}\hat{\mathbf{B}}\mathbf{M} - \mathbf{C})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\mathbf{B}}\mathbf{M} - \mathbf{C}) . \tag{8.41}$$

The result is ℓ by ℓ with ℓ being the number of columns of **M** and **C**, or in other words, the number of transformed dependent variables in the hypothesis in Equation (8.40). The error sums of squares and cross-products for the hypothesis, in contrast to the single sum of squares for the univariate version in Equation (5.22), is also an $\ell \cdot \ell$ matrix:

$$\mathbf{E} = \mathbf{M}' [\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \mathbf{Y}] \mathbf{M} . \tag{8.42}$$

Again in the univariate case, in Equation (6.18) we formed an F-ratio using the sum of squares for the hypothesis, and the sum of squares for the error. Modifying the form of Equation (6.18) somewhat, we can express the calculated F as

$$\hat{F} = \frac{SS_h / q}{SS_{Error} / n - k} = \frac{h / q}{e / n - k} = e^{-1}h \cdot \frac{n - k}{q}.$$

In the multivariate case we will do something similar, but the degrees of freedom are absorbed into the multivariate tables. But more importantly, since $\mathbf{E}^{-1}\mathbf{H}$ is an $\ell \cdot \ell$ matrix, we must decide how to summarize all of those numbers in a way that allows us to make an all-or-nothing decision about the hypothesis in Equation (8.40).

Eigenstructure affords an optimal method for summarizing a matrix, and in Section 8.8 we studied the eigenstructure of asymmetric matrices like $\mathbf{E}^{-1}\mathbf{H}$. We are now ready to test our multivariate linear hypothesis.

8.17 *Statistics for Testing the Multivariate General Linear Hypothesis*

If we define s as the rank of $\mathbf{E}^{-1}\mathbf{H}$, we then have the eigenvalues λ_1 , λ_2 , ···, λ_s of the system

$$(\mathbf{E}^{-1}\mathbf{H} - \lambda\mathbf{I})\mathbf{x} = 0. \tag{8.43}$$

In general, s = Min(q, ℓ), that is, whichever is smaller, the number of rows of **A** or the number of columns of **M**. The eigenstructure of $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$ will be of interest also:

$$[\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1} - \theta\mathbf{I}]\mathbf{x} = 0 \tag{8.44}$$

with

$$\theta_i = \frac{\lambda_i}{1 + \lambda_i} \tag{8.45}$$

so that

$$\lambda_i = \frac{\theta_i}{1-\theta_i}. \tag{8.46}$$

In a logical sense, the λ_i are analogous to F ratios, being the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$, while the θ_i are more analogous to squared multiple correlations, being the eigenvalues of $\mathbf{H}(\mathbf{H} + \mathbf{E}^{-1})$. Now there are four different ways to test the multivariate hypothesis, proposed by four different statisticians. In addition, there is an F approximation that is somewhat commonly used as well. The four are:

Hotelling-Lawley Trace
$$\mathrm{Tr}(\mathbf{E}^{-1}\mathbf{H}) = \sum_i^s \lambda_i \tag{8.47}$$

Roy's Largest Root
$$\theta_1 = \frac{\lambda_1}{1+\lambda_1} \tag{8.48}$$

Pillai's Trace
$$\mathrm{Tr}[\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}] = \sum_i^s \theta_i = \sum_i^s \frac{\lambda_i}{1+\lambda_i} \tag{8.49}$$

Wilk's Lambda
$$\Lambda = \frac{|\mathbf{H}|}{|\mathbf{H}+\mathbf{E}|} = \prod_i^s \frac{1}{1+\lambda_i} \tag{8.50}$$

An especially good set of tables for these statistics can be found in Timm (1975).

The F approximation is based on Wilk's determinantal criterion in Equation (8.50). That formula is

$$F' = \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \cdot \frac{rt-2u}{\ell q} \tag{8.51}$$

where, as before, q is the number of rows or the rank of $\mathbf{A}$, ℓ is the number of columns or the rank of $\mathbf{M}$, but there are some other parameters. The values

$$u = \frac{\ell q - 2}{4},$$

$$r = n - k - \frac{\ell - q + 1}{2},$$

$$t = \begin{cases} \dfrac{\ell^2 q^2 - 4}{\ell^2 + q^2 - 5} & \text{if } \ell^2 + q^2 - 5 > 0 \\ 1 & \text{if } \ell^2 + q^2 - 5 \le 0 \end{cases}$$

and n is the sample size while k is the number of columns of $\mathbf{X}$. The degress of freedom for F′ are $\ell \cdot q$ in the numerator and rt - 2u in the denominator. The approximation is exact if $s = \mathrm{Min}(\ell, q) \le$

2, which is to say that the rank of $\mathbf{E}^{-1}\mathbf{H}$ is 2 or less. You will note the eigenstructure discussed in Chapter 3 is a special case of the following discussion with $\mathbf{E} = \mathbf{I}$.

Earlier, in Section 8.4, we spoke of correcting a statistical test for having a large number of tests and also for post hoc data snooping. If we consider the hypothesis

$$H_0: \mathbf{a}'\mathbf{B}\mathbf{m} = 0$$

where we try to pick the elements in the vectors **a** and **m** to make the significance test as large as possible, then $\hat{\theta}_{max}$, from Equation (8.3) is Roy's largest root. Unlike the Dunn-Bonferroni approach, the Union-Intersection approach controls for a high number of tests and also takes into account the correlations between the dependent variables. Another example would be where we try to maximize the correlation between a linear combination of x variables and a linear combination of the y variables. This is called canonical correlation.

8.18 *Canonical Correlation*

In the multivariate general linear model, since there are p elements to the **y** vector and the k variables in the **x** vector, we face an embarrassment of riches in trying to summarize the relationship between the two sets of variables. Shown below, we see the partitioned matrix of all the variables, partitioned into y and x sets:

$$R = \left[\begin{array}{c|c} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \hline \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{array}\right]$$

The p · k matrix $\mathbf{R}_{yx}$ certainly has information in it about the relationship between the two sets of variables, containing as it does, the correlations between the sets. But in order to summarize the relationship between the two sets, we want a scalar. One obvious approach is to create new two new scores, one from the x set and one from the y set such that the correlation between the two scores is as high as possible. In essence, the problem is to pick the p elements of $\mathbf{c}'$ in

$$u = \mathbf{c}'\mathbf{z}_y \tag{8.52}$$

and the k elements of $\mathbf{d}'$ in

$$v = \mathbf{d}'\mathbf{z}_x \tag{8.53}$$

such that

$$\rho^2 = \frac{(\mathbf{c}\mathbf{R}_{yx}\mathbf{d})^2}{\mathbf{c}'\mathbf{R}_{yy}\mathbf{c} \cdot \mathbf{d}'\mathbf{R}_{xx}\mathbf{d}} \tag{8.54}$$

is maximized. This leads to two different eigenvector problems,

$$[\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} - \rho^2 I]\mathbf{c} = 0 \tag{8.55}$$

and

$$[\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx} - \rho^2 \mathrm{I}]\mathbf{d} = 0. \tag{8.56}$$

We can pick the smaller problem to solve and then deduce the other eigenvector using either

$$\mathbf{d} = \frac{1}{\rho^2}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{c} \tag{8.57}$$

or

$$\mathbf{c} = \frac{1}{\rho^2}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{d}. \tag{8.58}$$

The canonical correlation can be thought of as a linear hypothesis of the form of Equation (8.40) with

$${}_k\mathbf{A}_{k^*} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = [\,{}_k\mathbf{0}_1 \;\vdots\; {}_{k^*}\mathbf{I}_{k^*}]$$

and $\mathbf{M} = {}_p\mathbf{I}_p$. The number of canonical correlations and eigenvector combinations depends on s, which in this case is simply whichever is smaller, k or p. The first canonical correlation squared corresponds to Roy's Largest Root in Equation (8.48), which can be used to test the hypothesis that the canonical correlation is zero. One can also use Pillai's Trace [Equation (8.49)] to test whether all of the canonical correlations are zero, i. e.

$$\mathrm{H}_0\colon \rho_1^2 = \rho_2^2 = \cdots = \rho_s^2 = 0.$$

Placing each of the eigenvectors, $\mathbf{a}_{\cdot 1}$, $\mathbf{a}_{\cdot 2}$, ⋯, $\mathbf{a}_{\cdot s}$ into columns of the matrix **A** (not the hypothesis matrix), we have rows of **A** that correspond to y variables and columns of **A** that correspond to different canonical variables from Equation (8.52). We can standardize the elements of **A** using

$$\mathbf{C}_s = \mathbf{C}(\mathbf{C}'\mathbf{R}_{yy}\mathbf{C})^{-1/2}$$

and for the x set we have

$$\mathbf{D}_s = \mathbf{D}(\mathbf{D}'\mathbf{R}_{xx}\mathbf{D})^{-1/2}.$$

It is also instructive to look at the correlations between each of the canonical variables in Equation (8.53) and the variables of the x set, and the canonical variables in Equation (8.52) and the variables of the y set. We have for each combination

$$\mathrm{Cov}(\mathbf{u}, \mathbf{z}_y) = \mathbf{C}_s'\mathbf{R}_{yy},$$

$$\mathrm{Cov}(\mathbf{v}, \mathbf{z}_y) = \mathbf{D}_s'\mathbf{R}_{xx},$$

$$\mathrm{Cov}(\mathbf{u}, \mathbf{z}_x) = \mathbf{C}_s'\mathbf{R}_{yx},$$

$$\mathrm{Cov}(\mathbf{v}, \mathbf{z}_y) = \mathbf{D}_s' \mathbf{R}_{xy.}$$

8.19 *MANOVA*

We will begin with an example with a purely between subjects design, and two different dependent variables. Imagine that we have four groups of subjects, each group having seen a different advertisement. Thus, k = 4 with $\mathbf{x}_{\cdot 0}$ being the usual vector of constants and $\mathbf{x}_{\cdot 1}$, $\mathbf{x}_{\cdot 2}$ and $\mathbf{x}_{\cdot 3}$ coding for group membership. To keep things simple, lets say that $\mathbf{y}_{\cdot 1}$ contains the respondent's answer to the question, "How much do you like the product?" while $\mathbf{y}_{\cdot 2}$ has data on "Intention to buy." In summary, **Y** is n · 2, **X** is n · 4 and **B** is 4 · 2 with

$$\hat{\mathbf{Y}} = \mathbf{XB}.$$

It would be natural to test the hypothesis of no group differences for the two dependent variables. This hypothesis is much the same as canonical correlation, its just that the emphasis is slightly different. We calculate the hypothesis sum of squares and cross product matrix

$$\mathbf{H} = (\mathbf{A}\hat{\mathbf{B}}\mathbf{M} - \mathbf{C})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\mathbf{B}}\mathbf{M} - \mathbf{C}),$$

with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and $\mathbf{M} = {}_2\mathbf{I}_2$ and the error sum of squares and cross products matrix,

$$\mathbf{E} = \mathbf{M}' [\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \mathbf{Y}] \mathbf{M},$$

invert this latter matrix in order to find the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$, calculate the four criteria and the F approximation, and see to the fate of H_0. In addition, the eigenvectors for the y set,

$$\mathbf{v} = \mathbf{d}'\mathbf{y},$$

can tell us the optimal combination of y's for detecting group differences. Similarly, the eigenvectors for the x set reveal the best possible contrast among the group means.

8.20 *MANOVA and Repeated Measures*

To start off this section, we will pick an example with no grouping variables, just one group of consumers who rate a product using the same scale under p = 3 different scenarios. The multivariate model is then

$$\hat{\mathbf{Y}} = \mathbf{XB}$$

$$\begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} & \cdots & \hat{y}_{1p} \\ \hat{y}_{21} & \hat{y}_{22} & \cdots & \hat{y}_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{y}_{n1} & \hat{y}_{n2} & \cdots & \hat{y}_{np} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} [\beta_{01} \quad \beta_{02} \quad \beta_{03}].$$

To test the hypothesis that all scenarios lead to equal ratings, we use

$$H_0: \ \mathbf{ABM} = \mathbf{C}$$

$$H_0: \ 1 \cdot [\beta_{01} \quad \beta_{02} \quad \beta_{03}] \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = [0 \quad 0].$$

We can conceptualize the process here a little bit differently. For each subject, you could transform the scores prior to the analysis by applying the **M** hypothesis matrix directly to the **Y** matrix. In that case, you could simply test whether the β_0 values of the transformed measures were zero. So if we define

$$\tilde{\mathbf{Y}} = \mathbf{XB}$$

where **M** is exactly as before, and now we test to see if

$$H_0: \ 1 \cdot [\tilde{\beta}_{01} \quad \tilde{\beta}_{02}] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = [0 \quad 0]$$

where the parameters $\tilde{\beta}_{01}$ and $\tilde{\beta}_{02}$ would be estimated from $\tilde{\mathbf{Y}}$ instead of **Y**. Both approaches are equivalent because the hypotheses

$$H_0: \ \mu_{\tilde{y}_1} = \mu_{\tilde{y}_2} = 0 \tag{8.59}$$

and

$$H_0: \ \mu_{y_1} = \mu_{y_2} = \mu_{y_3} \tag{8.60}$$

are equivalent. Using the transformed dependent variable matrix $\tilde{\mathbf{Y}}$ and testing the Hypothesis of Equation (8.59) is an example of Hotelling's T^2 (pronounced Tao Squared), which is the multivariate analog of the household variety *t*-statistic. The T^2 is used to test hypotheses of the form

$$H_0: \boldsymbol{\mu}_y = \mathbf{c}$$

with $\boldsymbol{\mu}_y$ being the vector of population means for the dependent variables. Hotelling's T^2 can also be used to test multivariable mean differences across two groups, just as the *t* does where there is but one dependent variable.

Now we put together an example where there are different groups of subjects as well as repeated measurements. As before, we assume that all subjects rate a product under p = 3 different scenarios. But now there are actually four different treatment groups, each group having seen a different advertisement for the product. In that case, k = 4 so that the **B** matrix is 4 by 3. Each column of **B** corresponds to one of the three rating scenarios. The first row of **B** contains the intercept terms, while the next three rows pertain to group differences.

Is there an impact of advertisement? In the univariate approach, we add up the three measures to create for each subject i, $\tilde{y} = y_1 + y_2 + y_3$. We test the hypothesis using

$$H_0: \mathbf{A}\tilde{\boldsymbol{\beta}} = \mathbf{c}$$

$$H_0: \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \tilde{\beta}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

which is covered in Chapter 7. In the multivariate approach covered in this chapter, we do not transform the dependent variables, we leave them as they are. We have

$$H_0: \mathbf{ABM} = \mathbf{C}$$

$$H_0: \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} \\ \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \end{bmatrix} \mathbf{I} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This approach confounds the main effect of group with the simple main effect of advertisement on $\mathbf{y}_{\cdot 1}$, on $\mathbf{y}_{\cdot 2}$ and on $\mathbf{y}_{\cdot 3}$. In other words, from column 1 of **Y** we look to see what effect there is of group membership, we do the same thing with columns 2 and 3. But this claims some of the variance that would ordinarily be considered part of the advertisement × scenario interaction. The main effect of advertisement would generally look only at a summary of the group differences holding the scenario constant.

Is there an effect of scenario? Here we start with the univariate approach. If we define

$$\mathbf{M} = \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}$$

and assume that

$$\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M} = \sigma^2\mathbf{I}$$

as we did in Equation (7.12), we can utilize the univariate approach to repeated measures and use the F-test discussed in Section 7.7 with an error term of subjects × scenario interaction. In the univariate approach all scores are placed in a single column vector. In contrast, in the multivariate case each scenario constitutes a different column of **Y** and we test

$$H_0: \ \mathbf{ABM} = \mathbf{C}$$

$$H_0: \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} \\ \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \end{bmatrix}$$

There also exists an approach in between the univariate and multivariate methods. One could Test

$$H_0: \ \mathbf{M}'\mathbf{\Sigma}\mathbf{M} = \sigma^2\mathbf{I}$$

and pick the univariate approach if you fail to reject and the multivariate approach if you reject. Another approach was proposed by Greenhouse and Geisser (1959) who suggested that we could correct the univariate F to the degree that

$$\mathbf{M}'\mathbf{S}\mathbf{M} \neq \hat{\sigma}^2\mathbf{I}. \tag{8.61}$$

Here in Equation (8.61) we have replaced $\mathbf{\Sigma}$ with it's estimator, **S**.

If we wish to test the advertisement × scenario interaction according to the univariate approach, we would need to assume that $\mathbf{M}'\mathbf{\Sigma}\mathbf{M} = \sigma^2\mathbf{I}$, place all scores in the vector **y**, and use the interaction of subjects × scenario as the error term.

In order to test the advertisement × scenario interaction according to the multivariate model, we can combine the **A** matrix from the advertisement main effect and the **M** matrix from the scenario main effect. In that case we have

$$H_0: \ \mathbf{ABM} = \mathbf{C}$$

$$H_0: \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} \\ \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

8.21 *Classification*

To motivate this section, which will discuss the technique known as the discriminant function, we begin the discussion with a little two group example. Imagine we are trying to decide who to include in direct mailing. Our goal is to classify our customers into two groups based on whether they will, or will not, respond to the mailout. From a sample of our customer base, we have collected some data which we will get to in just a minute. For now, we note that the cost, or

disutility, of misclassifying someone in group i, mistakenly placing them in group j is c_{ij}. Given our two groups, we might then tabulate the cost matrix as

		Classification Decision	
		Group 1	Group 2
Reality	Group 1	0	c_{12}
	Group 2	c_{21}	0

For each individual we have a p element row vector from the matrix $\mathbf{Y}$, $\mathbf{y}'_{i\cdot}$, containing numeric variables. The probability density for the individuals in group j is $f_j(\mathbf{y}_{i\cdot})$, while π_j is the relative size of group j, also called the *prior probability*. The *conditional probability* an individual with vector $\mathbf{y}_{i\cdot}$ comes from group j is

$$\Pr(j \mid \mathbf{y}_{i\cdot}) = \frac{\pi_j f_j(\mathbf{y}_{i\cdot})}{\sum_m \pi_m f_m(\mathbf{y}_{i\cdot})}$$

We want to minimize our expected cost which in the two group case is given by

$$\Pr(1 \mid \mathbf{y}_{i\cdot})\, c_{12} + \Pr(2 \mid \mathbf{y}_{i\cdot})\, c_{21}$$

and we can decide that individual is in group 1 if

$$f_1(\mathbf{y}_{i\cdot}) \cdot \pi_1 \cdot c_{12} > f_2(\mathbf{y}_{i\cdot}) \cdot \pi_2 \cdot c_{21}$$

or rearranging we can say that we should decide that the individual is in group 1 if

$$\frac{f_1(\mathbf{y}_{i\cdot})}{f_2(\mathbf{y}_{i\cdot})} > \frac{\pi_2}{\pi_1} \cdot \frac{c_{21}}{c_{12}}. \tag{8.62}$$

If the π_j are unknown or assumed to be equal, and $c_{21} = c_{12}$, then it is only the right hand side of the above Equation (8.62) and what matters is the relative height of the two densities. The crossover point of Equation (8.62) would be the place where the densities themselves cross over.

The usual assumption is that an observation vector from group j

$$\mathbf{y}_{i\cdot} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

which implies from Equation (4.17) that

$$f_i(\mathbf{y}_{i\cdot}) = \frac{1}{|\boldsymbol{\Sigma}_j|^{1/2} (2\pi)^{p/2}} \exp\left[-(\mathbf{y}_{i\cdot} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{y}_{i\cdot} - \boldsymbol{\mu}_j)/2\right].$$

Taking Equation (8.62) and taking logs to both sides, we would then place a case in group 1 if

$$\ln \frac{f_1(\mathbf{y}_{i\cdot})}{f_2(\mathbf{y}_{i\cdot})} > \ln \frac{\pi_2 c_{21}}{\pi_1 c_{12}}$$

$$\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}-\frac{1}{2}\left[(\mathbf{y}_{i\cdot}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{y}_{i\cdot}-\boldsymbol{\mu}_1)-(\mathbf{y}_{i\cdot}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{y}_{i\cdot}-\boldsymbol{\mu}_2)\right]>\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}.$$

If we assume that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ the above expression simplifies to

$$\left[\mathbf{y}_{i\cdot}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]-\frac{1}{2}\left[(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mu_1-\boldsymbol{\mu}_2)\right]>\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}. \tag{8.63}$$

To get to this point it helps to realize that $(\mathbf{a}-\mathbf{b})'\mathbf{C}(\mathbf{a}-\mathbf{b}) = \mathbf{a}'\mathbf{C}\mathbf{a} - 2\mathbf{a}'\mathbf{C}\mathbf{b} + \mathbf{b}'\mathbf{C}\mathbf{b}$ and that $(\mathbf{a}+\mathbf{b})'\mathbf{C}(\mathbf{a}-\mathbf{b}) = \mathbf{a}'\mathbf{C}\mathbf{a} - \mathbf{b}'\mathbf{C}\mathbf{b}$. Noting also that $\ln\frac{a}{b} = -\ln\frac{b}{a}$, if we substract $\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}$ from both sides of the above equation we get

$$\left[\mathbf{y}_{i\cdot}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]-\frac{1}{2}\left[(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]-\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}>0. \tag{8.64}$$

Define the left hand side of this last equation as ν_{12}. Our decision to place a case in group 1 is made if

$$\nu_{12} > 0.$$

For population 1 we have

$$\nu_{12} \sim N\left(\ln\frac{\pi_1 c_{12}}{\pi_2 c_{21}}+\Delta_{12}^2,\ \Delta_{12}^2\right)$$

where

$$\Delta_{12}^2 = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

which is known as the Mahalanobis distance between the mean vectors of the two populations. Knowing the distribution of ν_{12} allows us to estimate the probability and the total costs of misclassification. We also define the raw discriminant function as

$$\nu = \mathbf{d}'\mathbf{y}_{i\cdot}$$

where

$$\mathbf{d} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

We can also standardize the function using

$$\nu_s = \Delta_{12}^{-1}\mathbf{d}'\mathbf{y}_{i\cdot} = \mathbf{d}_s'\mathbf{y}_{i\cdot}$$

Back to the decision,

$$\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}} + \left[\mathbf{y}_{i\cdot}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right] - \frac{1}{2}\left[(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right] > 0$$

$$\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}} + \mathbf{y}_{i\cdot}'\mathbf{d} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\mathbf{d} > 0.$$

Rearranging, our decision "1" is taken if

$$\nu = \mathbf{y}_{i\cdot}'\mathbf{d} > \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\mathbf{d} - \ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}$$

or in the standardized version

$$\nu_s = \mathbf{y}_{i\cdot}'\mathbf{d}_s > \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\mathbf{d}_s - \Delta_{12}^{-1}\ln\frac{\pi_2 c_{21}}{\pi_1 c_{12}}.$$

The discriminant function maximizes the separation between the values $\bar{\nu}_1$ and $\bar{\nu}_2$, the means for the two groups on the discriminant scores. When we don't know the $\boldsymbol{\mu}_j$ or $\boldsymbol{\Sigma}$, we split our samples into validation and holdout samples.

8.22 *Multiple Group Discriminant Function*

The problem can be approached as a special case of MANOVA. For example, assuming that we have k = 4 groups with p discriminating dependent variables, and the general linear hypothesis

$$H_0: \mathbf{ABM} = 0,$$

we would use the hypothesis matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

with $\mathbf{M} = \mathbf{I}$. Just as we did before in Equations (8.41) and (8.42), we would calculate the hypothesis and error sum of squares matrices **H** and **E**. In order to find a score, $\nu = \mathbf{d}'\mathbf{y}_{i\cdot}$, with ν have as large a between groups sum of squares as possible, we will utilize the eigenstructure of $\mathbf{E}^{-1}\mathbf{H}$ as before. We pick values in the vector **d** such that our F test for group differences on ν is as large as possible. In other words, we maximize the between groups sum of squares for ν divided by it's within groups sum of squares, that is to say $\frac{\mathbf{d}'\mathbf{H}\mathbf{d}}{\mathbf{d}'\mathbf{E}\mathbf{d}}$, over all possible values of **a**. It is customary to scale **a** such that the within-group variance (mean square) is

$$\frac{\mathbf{d}'\mathbf{E}\mathbf{d}}{n-k} = \mathbf{d}'\mathbf{S}\mathbf{d} = 1.$$

References

Greenhouse, Samuel W. and S. Geisser (1959) On Methods in the Analysis of Profile Data. *Psychometrika*, 24, 95-112.

Hair, Joseph F., Rolph E. Anderson, Ronald L. Tatham and William C. Black (1995) *Multivariate Data Analysis. Fourth Edition*.. Englewood Cliffs, NJ: Prentice-Hall.

Keppel, Geoffrey (1973) *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs, New Jersey: Prentice-Hall.

Lattin, James, J. Douglas Carroll and Paul E. Green (2003) *Analyzing Multivariate Data.* Pacific Grove, CA: Brooks/Cole.

Marascuilo, Leonard A. and Joel R. Levin (1983) *Multivariate Statistics in the Social Sciences*. Monterey, CA: Brooks/Cole.

Scheffé, Henri (1959) *The Analysis of Variance*. New York: Wiley.

Sharma, Subhash (1996) *Applied Multivariate Techniques*. New York: Wiley

Tatsuoka, Maurice M. (1971) *Multivariate Analysis*. New York: Wiley.

Timm, N. H. (1975) *Multivariate Analysis with Applications in Education and Psychology*. Monterey, CA: Brooks/Coles.

Section III: Covariance Structure

Chapter 9: Confirmatory Factor Analysis

Prerequisites: Chapter 5, Sections 3.9, 3.10, 4.3

9.1 *The Confirmatory Factor Analysis Model*

The difference between the models discussed in this section, and the regression model introduced in Chapter 5 is in the nature of the independent variables, and the fact that we have multiple dependent variables. The independent variables are unobserved constructs, also known as *factors*, dimensions or *latent variables*. At this point the student might ask, how scientific is it to speak of unobserved variables in a model? We will soon see that if the model of unobserved independent variables is correct, it makes a strong prediction about the structure of the covariances among the observed dependent variables. For this reason, these models are a special case of models known as *covariance structure models*.

Given that we are dealing with unobserved variables, it will be useful to shift our notation somewhat. In regression, we look at a particular variable as a column vector that displays the individual observations which comprise the rows. In factor analysis, the individual observations cannot be fully observed since the right hand side variables, the factors, are not observed. Instead, we will propagate our model using a typical observation, call it observation i, but leaving off the subscript i. What's more, instead of arranging our matrices such that the each column is a different variable and each row is a different observation, we will be looking at the transpose.

Of course, this is in contrast to the notation employed in Chapters 5 through 8. In that later chapter, we study the model

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}$$

where the columns of **Y** (and the parameter vector **B** as well as the error matrix $\boldsymbol{\varepsilon}$) represent the p different dependent variables. If we were to take the transpose of both sides of that model we would have

$$\mathbf{Y}' = \mathbf{B}'\mathbf{X}' + \boldsymbol{\varepsilon}'.$$

You will note that, since the product of a transpose is the transpose of the product in reverse order [Equation (1.34)], **B** and **X** are now reversed. Also, the data matrices $\mathbf{Y}'$ and $\mathbf{X}'$ now have a row for each variable, instead of a column as before. Next, as described above, rather than look at every subject, we look at a typical observation, for example, number i:

$$\mathbf{y}_{i\cdot} = \mathbf{B}'\mathbf{x}_{i\cdot} + \boldsymbol{\varepsilon}_{i\cdot}.$$

The dot, which is a subscript reduction operator, is mentioned in Section 1.1. One final change is convenient. If we totally drop the subscripts from $\mathbf{y}_{i\cdot}$, $\mathbf{x}_{i\cdot}$ and $\boldsymbol{\varepsilon}_{i\cdot}$, we would just have

$$\mathbf{y} = \mathbf{B}'\mathbf{x} + \boldsymbol{\varepsilon}.$$

This is how we will describe the model in this chapter. We will call the regression weights λ's instead of β's and the independent variables will be η's instead of x's.

We start out with a scalar representation of the situation:

$$
\begin{aligned}
y_1 &= \lambda_{11}\eta_1 + \lambda_{12}\eta_2 + \cdots + \lambda_{1m}\eta_m + \varepsilon_1 \\
y_2 &= \lambda_{21}\eta_1 + \lambda_{22}\eta_2 + \cdots + \lambda_{2m}\eta_m + \varepsilon_2 \\
\cdots &= \quad \cdots \\
y_p &= \lambda_{p1}\eta_1 + \lambda_{p2}\eta_2 + \cdots + \lambda_{pm}\eta_m + \varepsilon_p \ .
\end{aligned}
\tag{9.1}
$$

The left hand side shows p different variables. Perhaps y_1 through y_3 represent three measures of consumer "greenness", that is, a tendency to buy environmental friendly products. Perhaps y_4 through y_6 represent three different measures of innovativeness. In any case, the point is that the y's are p *manifest* or observed variables. As has been mentioned, we are representing the data from a typical subject, the i-th, but the subscript i is left off according to the traditions in this area. On the right hand side, you have regression coefficients, the λ_{ij}, which are basically β weights. In the context of factor analysis, regression weights are called factor *loadings*. The reason that they have two subscripts is that you need one subscript to keep track of the dependent variable, or the equation, and another subscript to keep track of the independent variable. And speaking of which, these are the η values of which there are m. The η's are the *common factors* which explain much of the behavior of the y's, at least the part of their behavior that they have in common – the covariances. Finally, we have the ε_i which are called *unique factors*. This is not exactly the same thing as the error in a regression model. In regression, the error is an *error-in-equations*, also called *specification error.* That is to say, unless a regression model has an R^2 of 1, the model is missing some explanatory independent variables or is otherwise *mispecified*. In factor analysis, the ε's are *errors-in-variables*, or *measurement error*. The three variables we devised to measure "greenness", for example, might not do so perfectly. We generally assume that the part that the three variables have in common, as quantified by their covariances, must be due to the fact that all three are at least partially measuring what they are supposed to be measuring. But each one of the three has some variance that is unique to it. That is what the ε_i account for.

We can write the model in matrix terms,

$$
\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_p \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \cdots \\ \eta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_p \end{bmatrix}.
$$

$$
\mathbf{y} = \mathbf{\Lambda\eta} + \boldsymbol{\varepsilon} \ . \tag{9.2}
$$

By all rights, in addition to the $\mathbf{y}$ vector, the $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ vectors should have a subscript i since they are random variables, sampled from the population for which this model holds. On the other hand, $\mathbf{\Lambda}$ is a constant matrix, holding parameters that describe this population.

So how does this model with unobserved variables make contact with reality? In order to show how it does so, we need to start with some assumptions and some definitions. We will assume that $E(\mathbf{y}) = \mathbf{0}$, a p by 1 null vector. This does not reduce the generality of the model at all, since covariances are not affected by the addition or subtraction of a constant [see Theorem (4.8)]. In order to estimate the model, we will make the assumptions that

$$
\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{\Psi}),
$$

$$
\boldsymbol{\varepsilon} \sim \mathbf{N(0, \Theta)}
$$

and that

$$\text{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) = \mathbf{0}.$$

Like the $\mathbf{y}$ vector, $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ are mean-centered. We will also see quite a bit of the coviarance matrices for $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$, with $V(\boldsymbol{\eta}) = \boldsymbol{\Psi}$ and $V(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$. At this point, we are ready to see what the covariance matrix of the y's should look like. We have by the definition of variance in Equation (4.7)

$$\begin{aligned} V(\mathbf{y}) \equiv \boldsymbol{\Sigma} &= E(\mathbf{y}\mathbf{y}') \\ &= E\left[(\boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon})(\boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon})'\right] \\ &= \boldsymbol{\Lambda}\, E(\boldsymbol{\eta}\boldsymbol{\eta}')\, \boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\, E(\boldsymbol{\eta}\boldsymbol{\varepsilon}') + E(\boldsymbol{\varepsilon}\boldsymbol{\eta}')\, \boldsymbol{\Lambda}' + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'), \end{aligned}$$

but of the four components from left to right, pieces two and three vanish since $\text{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\eta}) = \mathbf{0}$. We have made use of Equation (4.5) and (4.6). We can rewrite $E(\boldsymbol{\eta}\boldsymbol{\eta}') = \boldsymbol{\Psi}$, which was defined above as the covariance matrix of the η's when we were talking about assumptions. In piece four we have $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Theta}$ which was also defined above as the variance of the unique factors. Putting all of these conclusions together, we end up with the fact that the variance of $\mathbf{y}$ is

$$V(\mathbf{y}) = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}\,. \tag{9.3}$$

9.2 *A Confirmatory Factor Analysis Example*

Now is the section of the chapter where we look at an example confirmatory factor analysis that is just complicated enough to be a valid example, but is simple enough to be, well; a silly example. Lets say we have devised three questionnaire items which measure the consumers' attitude towards Beer B, and three other items that measure attitudes towards Beer C. Our six item survey then contains the variables listed in the table:

Variables	**Description**
y_1	Measurement 1 of B
y_2	Measurement 2 of B
y_3	Measurement 3 of B
y_4	Measurement 1 of C
y_5	Measurement 2 of C
y_6	Measurement 3 of C

To finish describing the model, we will hypothesize that there are two factors, B (η_1) and C (η_2). Our model would then look like

$$\mathbf{y} = \mathbf{\Lambda\eta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}$$

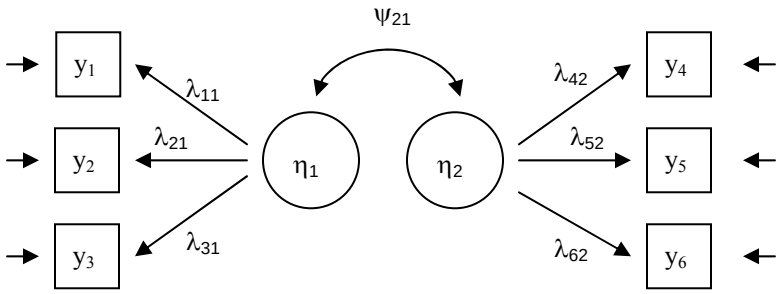
Again, remember that the $\mathbf{y}$, $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ vectors are random variables, but $\mathbf{\Lambda}$ is a parameter matrix and the unknowns in it must be estimated from the sample. To fully estimate the model, we also have two other parameter matrices,

$$\mathbf{\Psi} = \begin{bmatrix} \psi_{11} & - \\ \psi_{21} & \psi_{22} \end{bmatrix} \text{ and}$$

$$\mathbf{\Theta} = \begin{bmatrix} \theta_{11} & 0 & \cdots & 0 \\ 0 & \theta_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \theta_{pp} \end{bmatrix}.$$

Note that the $\mathbf{\Psi}$ matrix is symmetric, being a covariance matrix and so we do not need to enumerate the upper triangular part of it. And by the definition of what we mean by a unique factor, the ε_i are independent which means that the variance matrix of the ε_i, $\mathbf{\Theta}$, is diagonal. As a general rule in covariance structure models, we need to specify variances and covariances of right hand side random variables, and we need to specify regression weights between right hand and left hand side variables.

Below you can see what we call the *Path Diagram* for this model:



A path diagram is a very common way of representing a covariance structure model, and there are a set of conventions that go along with this type of figure. Single-headed arrows represent directional causal pathways, and two-headed arrows are used to represent covariation. Unique

factors, and other sorts of error terms, are usually indicated by single-headed arrows without labels. Circles are used to convey the presence of latent variables, and boxes convey observed variables.

9.3 *Setting a Metric for Latent Variables*

The model as it has been presented so far cannot be uniquely identified. To illustrate this, lets pretend we have a single variable and a single factor. In that case everything boils down to scalars, and the model is $y = \lambda\eta + \varepsilon$ and from Equation (9.3), $V(y) = \lambda^2\psi + \theta$. Now define $\eta^* = a\cdot\eta$ so that $V(\eta^*) = a^2\psi = \psi^*$. Also, define $\lambda^* = \lambda/a$. In that case,

$$y = \lambda^*\eta^* + \varepsilon = \frac{\lambda}{a}\cdot a\eta + \varepsilon \quad \text{and also} \tag{9.4}$$

$$V(y) = \lambda^{*2}\,\psi^* + \theta = \frac{\lambda^2}{a^2}\cdot a^2\psi + \theta\,. \tag{9.5}$$

What this means is that if I have a model with parameters λ^* and ψ^*, and you have a model with parameters λ and ψ, both models would fit equally well and there would be no logical way to decide which was better. In fact, they would be completely equivalent. The source of this ambiguity lies in the fact that η is unobserved, and it is at most an interval scale. To further identify the model we must set intervals for it, a process called setting its metric. We can do this in one of two ways. We can fix one loading per factor to a constant, such as 1.0, or we can fix the variance of each factor to 1.0. Returning to our two factor example, the first method would yield

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & 1 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix}, \quad \text{and} \quad \mathbf{\Psi} = \begin{bmatrix} \psi_{11} & - \\ \psi_{21} & \psi_{22} \end{bmatrix}$$

while the second approach would give

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix}, \quad \text{and} \quad \mathbf{\Psi} = \begin{bmatrix} 1 & - \\ \psi_{21} & 1 \end{bmatrix}.$$

These two methods are equivalent, yielding the same Chi Square values, but the first method is slightly more general, being applicable in certain situations where the second method cannot be used. The first method ties the metric of each factor to the first variable that measures it. The second method turns the factors into z-scores, and the factor covariance matrix $\mathbf{\Psi}$ can then be interpreted as a correlation matrix. For both methods, the $\mathbf{\Theta}$ matrix has p free parameters.

9.4 *Degrees of Freedom for a Confirmatory Factor Analysis Model*

Factor analysis does not look directly at raw data. The input data for this technique are the elements of the sample covariance matrix **S**, which is a p by p symmetric matrix. Therefore **S** contains

$$\frac{p(p+1)}{2} \tag{9.6}$$

"data points", those being the p variances and the p(p-1)/2 unique covariances. For our 6 variable example, this would total 21. In our model, assuming we use the first method to fix the metric of the two factors, we have

4 λ's
3 ψ's
6 θ's
---
13 parameters

The degrees of freedom for the model are equal to the number of data points minus the number of unique free parameters that are estimated from those data. In our case, we have 21 – 13 = 8 degrees of freedom. We will be able to reject the model (or not as the case may be) using a χ^2 test with 8 degrees of freedom. In terms of hypotheses, we will be testing

$$H_0: \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \tag{9.7}$$

against the general alternative

$$H_A: \boldsymbol{\Sigma} = \mathbf{S} . \tag{9.8}$$

In some ways this pair of hypotheses is very similar to hypotheses that we saw in Chapter 6 with regression. However, here we have a different sort of emotional attachment to the hypotheses. In regression, which encompasses everything from the basic *t*-test through more complex possibilities, we are generally motivated to "hope for" H_A and hope against H_0. Here, our model is H_0, so in an emotional sense, the roles of the Type I and II errors are reversed. The truth is that the current situation is actually more natural, if we can use that word. In regression, the hypothesis we are testing is a sort of "straw man" that no one believes in anyway, and that we set up just to knock down. We will talk more about the "emotional reversal" of H_0 and H_A later when we discuss goodness of fit measures (that is, measures other than the traditional χ^2). But first, it is time to understand how we estimate the parameters of the model and come up with a χ^2 value to test it. That is the topic of the next two sections. We will be using an estimation philosophy known as Maximum Likelihood. In order to explore this topic, we will be returning to the much simpler regression model. Then we will venture forth and look at estimation for confirmatory factor analysis models.

9.5 *Maximum Likelihood Estimators for Factor Analysis*

Maximum likelihood is discussed in general in Section 3.10 and within the context of the regression model in Section 5.4. ML for factor analysis begins with the probability of observation i under the confirmatory factor analysis model. Here we have the multivariate normal distribution

[see Equation (4.17)] to deal with since we have p variables, not just one as we did with regression. We have

$$\Pr(\mathbf{y}_{i\cdot}) = \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}\mathbf{y}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\right] \tag{9.9}$$

for the p variables on observation i. For the whole sample we have

$$\ell_0 = \prod_i^n \Pr(\mathbf{y}_i) = \frac{1}{(2\pi)^{np/2}\,|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_i^n \mathbf{y}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\right]. \tag{9.10}$$

The summation in the exponent of the above equation makes sense if you keep in mind that $e^a \cdot e^b = e^{a+b}$. Now, to get ready for the next equation note that from Equation (1.27) and Section 1.7

$$\sum_i^n \mathbf{y}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i' = \mathrm{tr}[n\mathbf{S}\boldsymbol{\Sigma}^{-1}] \tag{9.11}$$

because

$$\sum_i^n \mathbf{y}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i = \mathrm{Tr}\left[\sum \mathbf{y}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\right] = \mathrm{Tr}\sum_i^n \mathbf{y}_i\mathbf{y}_i'\boldsymbol{\Sigma}^{-1} = \mathrm{Tr}[n\mathbf{S}\boldsymbol{\Sigma}^{-1}].$$

This is so since a scalar is equal to its trace, and the trace of a product is invariant to the sequence of that product assuming conformability. We now take the log of the likelihood in Equation (9.10) but substitute the identity from Equation (9.11) to end up with

$$\begin{aligned} \ln \boldsymbol{\ell}_0 = L_0 &= -\frac{1}{2}np\,\ln(2\pi) - \frac{1}{2}n\ln|\boldsymbol{\Sigma}| - \frac{1}{2}n\,\mathrm{tr}\,(\mathbf{S}\boldsymbol{\Sigma}^{-1}) \\ &= \text{constant} - \frac{1}{2}n\left[\ln|\boldsymbol{\Sigma}| + \mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1})\right]. \end{aligned} \tag{9.12}$$

The term "constant" above represents $-\frac{1}{2}np\,\ln(2\pi)$ which doesn't impact the optimal solution one way or the other since it does not depend on the parameters and so will not figure into the derivative. Now suppose I look at the likelihood under H_A: $\boldsymbol{\Sigma} = \mathbf{S}$. We will call that log likelihood L_A and we find that

$$L_A = \text{constant} - \frac{1}{2}n\left[\ln|\mathbf{S}| + p\right]. \tag{9.13}$$

Now we have two log likelihoods, one; L_0 which reflects the confirmatory factor analysis model, and another that gives us the log likelihood under the general alternative that $\boldsymbol{\Sigma}$ exhibits no particular structure, which is to say it is arbitrary. In other words, it is what it is.

It turns out that under very general conditions,

$$-2\ \ln\left[\frac{\ell_0}{\ell_A}\right] = -2\left[L_0 - L_A\right] \sim \chi^2(m), \tag{9.14}$$

where m represents the difference in the number of parameters estimated under the two models; the null (0) and the alternative (A). As we have already described, the alternative model estimates $\frac{p(p+1)}{2}$ parameters while the number of parameters in the null model depends on the specific theory as expressed in the matrices $\mathbf{\Lambda}$, $\mathbf{\Psi}$ and $\mathbf{\Theta}$. Plugging Equations (9.12) and (9.13) into Equation (9.14), the χ^2 value is then

$$\hat{\chi}^2 = n\left[\ln|\mathbf{\Sigma}| - \ln|\mathbf{S}| + tr(\mathbf{S}\mathbf{\Sigma}^{-1}) - p\right]. \tag{9.15}$$

As can be seen, as $\mathbf{\Sigma} \to \mathbf{S}, \hat{\chi}^2 \to 0$. Thus the closer the match between $\mathbf{\Sigma}$ and $\mathbf{S}$, the smaller the value of χ^2. But it is also true that as $n \to \infty, \hat{\chi}^2 \to \infty$, and conversely, as $n \to 0, \hat{\chi}^2 \to 0$. This means that all things being equal, it becomes easier to reject H_0 the larger the sample size, and it becomes harder to reject H_0 the smaller the sample size. This is how all efficient statistics function, but since we have an emotional attachment to H_0 instead of H_A, this would seem to have certain consequences both for individual researchers, and for the development of marketing as a whole.

It is necessary that we pick values for the unknowns in the matrices $\mathbf{\Lambda}$, $\mathbf{\Psi}$ and $\mathbf{\Theta}$ at the minimum value of Equation (9.15). Equation (9.15) is obviously nonlinear in the unknowns so this will entail nonlinear optimization as discussed in Section 3.9. For now we note that any computer algorithm that finds the minimum of Equation (9.15) will utilize the derivatives of that function to determine "which way is down". Any such algorithm, however, requires rational starting values to avoid ending up in a local, rather than the global, minimum of the function. As such, you should do the best job that you can by manually inserting starting values into whatever program you use to estimate the confirmatory factor model. Certainly, under any circumstances, you should be able to get the sign right for any loadings in the matrix $\mathbf{\Lambda}$. Diagonal elements of $\mathbf{\Theta}$ could be seeded with small positive values. Diagonal elements of $\mathbf{\Psi}$ are likely to resemble the variances of the measures, while off-diagonal elements could be smaller than the diagonal, and of appropriate sign. Of course, it is also important that any fixed elements in the matrices $\mathbf{\Lambda}$, $\mathbf{\Psi}$ and $\mathbf{\Theta}$ have appropriate starting values, as these will also end up as the final values!

9.6 *Special Case: The One Factor Model*

Consider a confirmatory factor model with one factor:

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_p \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_p \end{bmatrix} \eta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_p \end{bmatrix} .$$

If we fix $V(\eta) = \psi_{11} = 1$, the expression for the covariance matrix is simply

$$\mathbf{\Sigma} = \mathbf{\lambda}\mathbf{\lambda}' + \mathbf{\Theta} \tag{9.16}$$

and our measures y_1 , y_2 , …, y_p are called *congeneric tests*. In this context the single η is called a *true score*. As you might guess, this terminology comes from the field of educational and psychological measurement. If we further specialize the model so that all lambdas are equal, i. e.

$$\lambda_1 = \lambda_2 = \cdots = \lambda_p = \lambda ,$$

we have the model of *τ-equivalent tests*. Congeneric tests have p λ's and p θ's, but τ-equivalent tests have only one λ and p θ's. Finally, the model of *parallel tests* includes the additional restriction that

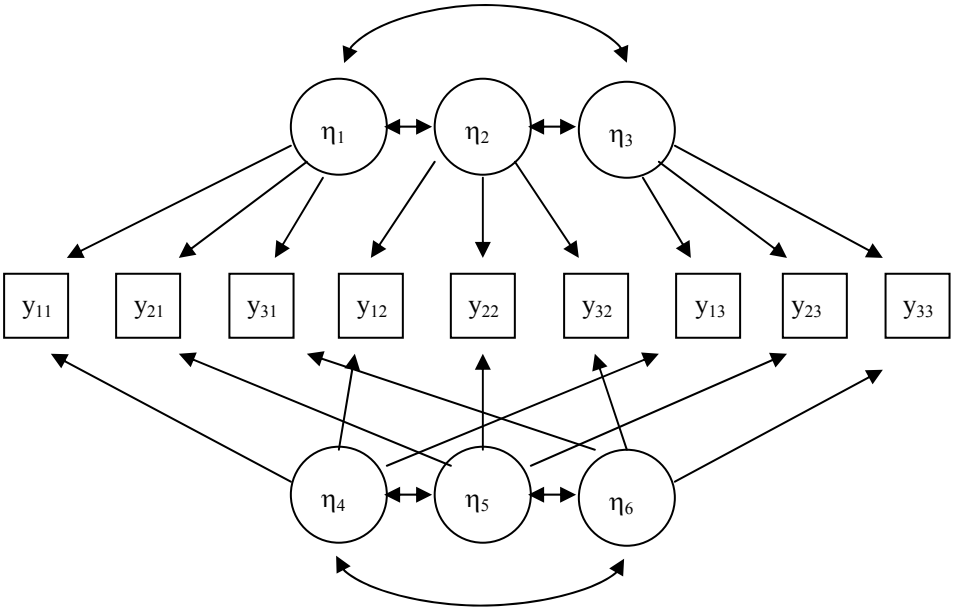
$$\theta_{11} = \theta_{22} = \cdots = \theta_{pp} = \theta .$$

Congeneric tests involve 2p free parameters to be estimated from the sample covariances, τ-equivalent tests have p + 1 parameters, and parallel tests have only 2 unknown parameters. Thus the model of parallel tests makes a very strong prediction about the structure of the covariance matrix using only 2 parameters. Having only 2 parameters means that the model has a larger number of degrees of freedom than τ-equivalence and especially congeneric tests. The degrees of freedom of the model represent restrictions that must be met in the covariance matrix. As such, parallel tests places many more restrictions on the covariance matrix which is shown below:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \lambda \\ \lambda \\ \cdots \\ \lambda \end{bmatrix} \begin{bmatrix} \lambda & \lambda & \cdots & \lambda \end{bmatrix} + \begin{bmatrix} \theta & 0 & \cdots & 0 \\ 0 & \theta & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \theta \end{bmatrix} .$$

9.7 *The Multi-Trait Multi-Method Model*

We sometimes have an opportunity to measure a set of traits using a common set of methods. For example we might measure the consumer's attitude towards a set of products repeating the same items to measure each product. With three traits (products) and three methods (items) we would have a path diagram as below. Note that to simplify an already complicated diagram, the unique factors were left off, as were the labels on the arrows.



and then the model would appear as

$$\begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \\ y_{12} \\ y_{22} \\ y_{32} \\ y_{13} \\ y_{23} \\ y_{33} \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 & 0 & \lambda_{14} & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 & \lambda_{25} & 0 \\ \lambda_{31} & 0 & 0 & 0 & 0 & \lambda_{36} \\ 0 & \lambda_{42} & 0 & \lambda_{44} & 0 & 0 \\ 0 & \lambda_{52} & 0 & 0 & \lambda_{55} & 0 \\ 0 & \lambda_{62} & 0 & 0 & 0 & \lambda_{66} \\ 0 & 0 & \lambda_{73} & \lambda_{74} & 0 & 0 \\ 0 & 0 & \lambda_{83} & 0 & \lambda_{85} & 0 \\ 0 & 0 & \lambda_{93} & 0 & 0 & \lambda_{96} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \\ \eta_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{31} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{32} \\ \varepsilon_{13} \\ \varepsilon_{23} \\ \varepsilon_{33} \end{bmatrix},$$

where η_1, η_2 and η_3 are trait factors and η_4, η_5 and η_6 are method factors. To finish specifying the model, we note that $V(\boldsymbol{\varepsilon}) = \text{Diag}(\theta_{11}\ \theta_{22} \cdots \theta_{99})$, meaning that the nine unique elements of $\boldsymbol{\Theta}$ are arrayed on it's diagonal, and that

$$V(\boldsymbol{\eta}) = \boldsymbol{\Psi} = \begin{bmatrix} 1 & & & & & \\ \alpha_{21} & 1 & & & — & \\ \alpha_{31} & \alpha_{32} & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & \beta_{21} & 1 & \\ 0 & 0 & 0 & \beta_{31} & \beta_{32} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{\alpha} & - \\ \mathbf{0} & \boldsymbol{\beta} \end{bmatrix}.$$

The three by three section of zeroes in $\boldsymbol{\Psi}$ is null because trait and method factors are assumed independent, an assumption that we would be testing when we look at the χ^2 for the model. Note that we have called the correlations among the trait factor α's and the correlations among the method factors β's. This does not change anything of course. This is just a confirmatory factor analysis model in which certain values in the $\boldsymbol{\Psi}$ matrix are playing slightly different roles from other values.

9.8 Goodness of Fit, Root Mean Square Error, and Other Output from the Model

With a large enough sample size, one can statistically reject even fairly good models. Conversely, with a small sample size it is possible to fail-to-reject models that are patently incorrect. Given that state of affairs, Bentler and Bonet (1980) proposed that in addition to comparing H_0 vs H_A, that we introduce a truly null hypothesis. I will call this latest hypothesis H_S for "straw man" hypothesis. Specifically we have

$$H_A: \boldsymbol{\Sigma} = \mathbf{S}$$

$$H_0: \boldsymbol{\Sigma} = \boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}$$

$$H_S: \boldsymbol{\Sigma} = \boldsymbol{\Psi} \text{ (with } \boldsymbol{\Psi} \text{ diagonal)}$$

For the straw man hypothesis, H_S, we have further restricted H_0 such that $\boldsymbol{\Lambda} = \mathbf{I}$, $\boldsymbol{\Theta} = \mathbf{0}$, and $\boldsymbol{\Psi}$ is diagonal. We have three hypotheses. For hypothesis j, with degrees of freedom df_j, we define

$$Q_j = \frac{\hat{\chi}^2}{df_j}$$

and then we define

$$\rho_{s0} = \frac{Q_S - Q_0}{Q_S - 1} \tag{9.17}$$

as one possible measure and

$$\Delta_{s0} = \frac{\hat{\chi}_s^2 - \hat{\chi}_0^2}{\hat{\chi}_s^2} \tag{9.18}$$

as another measure of *goodness of fit*. This latter index, Δ_{s0}, where the subscripts s and 0 highlight the fact that we are comparing hypotheses s and 0, represents the percent improvement in $\hat{\chi}^2$ from hypothesis s to hypothesis 0. The quantity 1 - Δ_{s0} gives us the remaining improvement that would be possible for H_A.

Joreskög has proposed an index simply termed *GFI* that consists of

$$\text{GFI} = 1 - \frac{\text{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S} - \mathbf{I}\right]^2}{\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})}$$

and an adjusted version,

$$\text{AGFI} = 1 - \frac{p(p+1)}{2 \cdot df_0}(1 - \text{GFI})$$

We should also mention that there exists a traditional measure of fit for any sort of model, the root mean square error, or

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}(s_{ij} - \sigma_{ij})^2}{p(p+1)/2}} \quad .$$

Note that the double summation operators in the numerator run through each of the unique elements in the covariance matrix. The RMSE gives you the average error across the elements of $\boldsymbol{\Sigma}$ as compared with $\mathbf{S}$.

We can also look at lack of fit for any individual fixed parameter. Of course, any free parameter estimated from the sample covariance matrix $\mathbf{S}$ does not contribute to lack of fit. It is the fixed parameters, generally the 0's in $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$ that are being tested in H_0 and it is these elements that cause a model to not fit. Given that we are picking free parameters in such a way that the derivative of Chi Square with respect to those parameters is 0, or assuming all of our free parameters are in the vector $\boldsymbol{\alpha}'$, we have solved for the free parameters when

$$\frac{\partial \hat{\chi}^2}{\partial \boldsymbol{\alpha}'} = \mathbf{0}'$$

because when the derivatives are zero, Chi Square is minimized. But this suggests a way to judge the fixed parameters. For any fixed parameter, say π, in general

$$\frac{\partial \hat{\chi}^2}{\partial \pi} \neq 0 \, .$$

These first derivatives provide a clue as to which parameter can be changed from fixed to free for the maximal benefit to $\hat{\chi}^2$. All that remains is that we scale the first derivative with the second derivative and we have what is called a *modification index*, or MI:

$$\text{MI} = \frac{\frac{n}{2}\left(\frac{\partial \hat{\chi}^2}{\partial \pi}\right)^2}{\frac{\partial (\hat{\chi}^2)^2}{\partial \pi \, \partial \pi}} \, .$$

General information on the second order derivative is given in Section 3.3 and its role in ML is discussed in Section 3.10.

References

Bentler, Peter and Douglas Bonett (1980) Significance Tests and Goodness-of-Fit in the Analysis of Covariance Structures. *Psychological Bulletin*, 88, 588-606.

Cole, David A and Scott E. Maxwell (1985) Multitrait-Multimethod Comparisons Across Populations: A Confirmatory Factor Analytic Approach. *Multivariate Behavioral Research*. 2, 389-417.

Jöreskog, Karl Gustav (1969) A General Approach to Confirmatory Maximum Likelihood Factor Analysis. *Psychometrika*. 34 (2), 183-200.

Chapter 10: Structural Equation Models

Prerequisites: Chapter 9

10.1 *The Basic Structural Equation Model*

In this chapter we are going to look at models where the theme is cause and effect. Unlike regression, these models are explicitly formulated as causal models, not just predictive models. We will also be using a notation that is quite similar to that used in Chapter 9 for Confirmatory Factor Analysis, which is to say that we will have a column vector, **y**, containing p dependent variables. The vector **y** is understood to represent an arbitrarily chosen observation from the population, maybe the ith. We will have a similar situation with the vector **x** that is a q by 1 column vector. In SEM (Structural Equation Model) terms, we say that **y** contains the *endogenous variables* and **x** contains the *exogenous variables*. An endogenous variable is one that appears at least once as the dependent variable in an equation. On the other hand, variables that do not appear on the left hand side are exogenous, or "given." In other words, all variances of, and covariances between, exogenous variables are determined outside of the system. They are not at issue. The variances and covariances of the endogenous variables are being modeled as a function of the exogenous variables. The basic model looks like

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_p \end{bmatrix} = \begin{bmatrix} 0 & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & 0 & \cdots & \beta_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{p1} & \beta_{p2} & \cdots & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_p \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1q} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pq} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_q \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \cdots \\ \zeta_p \end{bmatrix}$$

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\zeta}. \tag{10.1}$$

So we have p simultaneous equations. Note that for each of the causal parameters, the γ's and the β's, the subscripts follow the same pattern. The first subscript refers to the equation, in other words the y variable which is the effect. The second subscript refers to the cause.

The p by p **B** matrix contains the coefficients of the regressions of y variables on other y variables with 0's on the diagonal which implies that a variable cannot cause itself. The p by q matrix **Γ** contains the coefficients of the y's on the x's. The error vector, **ζ**, is p by 1. These errors are different than factor analysis errors, they represent *errors-in-equations*, in the way that these equations are specified. Thus they are also called *specification errors*.

In order to get to a point where we can estimate the model, we need to add some assumptions. To start off innocuously enough, we assume that E(**y**) = **0** and E(**x**) = **0**, which has absolutely no impact on the variances or covariances of these variables [see Equation (4.8)]. We then assume that the **x** and **ζ** vectors are independent,

$$\text{Cov}(\mathbf{x}, \boldsymbol{\zeta}) = \mathbf{0} \tag{10.2}$$

which is to say that the covariances between the x's and the ζ's consist of a q by p rectangular array of zeroes. We will also need to assume that the determinant

$$|\mathbf{I} - \mathbf{B}| \neq 0. \tag{10.3}$$

Now let us define

$$V(\mathbf{x}) = E(\mathbf{x}\mathbf{x}') = \mathbf{\Phi} \quad \text{and} \tag{10.4}$$

$$V(\boldsymbol{\zeta}) = E(\boldsymbol{\zeta}\boldsymbol{\zeta}') = \mathbf{\Psi} . \tag{10.5}$$

Note that we have "reused" the $\mathbf{\Psi}$ matrix from Chapter 9. In confirmatory factor analysis, $\mathbf{\Psi}$ was used for the factor covariance matrix. In fact, the use of $\mathbf{\Psi}$ as the covariance matrix of the ζ's is actually consistent with its Chapter 9 meaning. At this point we are ready to deduce what is known as *reduced form*. Reduced form requires that we solve for the **y** vector, as below:

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$$

$$\mathbf{y} - \mathbf{B}\mathbf{y} = \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$$

$$(\mathbf{I} - \mathbf{B})\mathbf{y} = \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$$

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Gamma}\mathbf{x} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta}$$

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{e} . \qquad . \tag{10.6}$$

The matrices $\mathbf{G} = (\mathbf{I} - \mathbf{B})^{-1}\Gamma$ and $\mathbf{e} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta}$ are defined merely for convenience, but **G** does highlight the fact that we can go from the structural parameters in **B** and $\mathbf{\Gamma}$ to the classic regression parameters with some algebra. Of course, that does not prove we can go in the opposite direction!

What is the variance of the y variables? We can use the reduced form derived above to simplify our explanation,

$$\mathbf{\Sigma} = E(\mathbf{y}\mathbf{y}') = E\left[(\mathbf{G}\mathbf{x} + \mathbf{e})(\mathbf{G}\mathbf{x} + \mathbf{e})'\right]$$

$$= E(\mathbf{G}\mathbf{x}\mathbf{x}'\mathbf{G}') + E(\mathbf{G}\mathbf{x}\mathbf{e}') + E(\mathbf{e}\mathbf{x}'\mathbf{G}') + E(\mathbf{e}\mathbf{e}') . \tag{10.7}$$

The 2nd and 3rd terms vanish. To see this, we look at the 2nd component which is given by

$$E(\mathbf{G}\mathbf{x}\mathbf{e}') = E\left\{\mathbf{G}\mathbf{x}\left[(\mathbf{I} - \mathbf{B}')^{-1}\boldsymbol{\zeta}\right]'\right\}$$

which, using the fact that the transpose of an inverse is the inverse of the transpose [Equation (1.40)], and passing the constants in **G** and $(\mathbf{I} - \mathbf{B})^{-1}$ through the Expectation operator [remember Equations (4.5) and (4.6)], is equivalent to

$$E(\mathbf{G}\mathbf{x}\mathbf{e}') = \mathbf{G}\,E(\mathbf{x}\boldsymbol{\zeta}')\,(\mathbf{I} - \mathbf{B})^{-1} .$$

Here we note that $E(\mathbf{x}\boldsymbol{\zeta}')$ that appears immediately above is another way to express the $Cov(\mathbf{x}, \boldsymbol{\zeta})$, and that covariance must be zero by previous assumption. The 3rd term is just the transpose of the 2nd. What the cancellation of the 2nd and 3rd components in equation (10.7) means is that we end up with the following expression for $\mathbf{\Sigma}$,

$$E(\mathbf{y}\mathbf{y}') = \mathbf{G}E(\mathbf{x}\mathbf{x}')\mathbf{G}' + E(\mathbf{e}\mathbf{e}') \tag{10.8}$$

At this point we might pause and note the similarity between this expression and it's equivalent for factor analysis, Equation (9.3)! Now, to further flesh out this last equation we need to remember that we had previously defined $V(\mathbf{x}) = E(\mathbf{x}\mathbf{x}') = \mathbf{\Phi}$, and $V(\boldsymbol{\zeta}) = E(\boldsymbol{\zeta}\boldsymbol{\zeta}') = \mathbf{\Psi}$. Proceeding along those lines we see that

$$E(\mathbf{y}\mathbf{y}') = (\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}'(\mathbf{I}-\mathbf{B}')^{-1} + (\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Psi}(\mathbf{I}-\mathbf{B}')^{-1}$$

$$= (\mathbf{I}-\mathbf{B})^{-1}[\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}' + \mathbf{\Psi}](\mathbf{I}-\mathbf{B}')^{-1} .$$

How about the covariance between **x** and **y**? That would be

$$E(\mathbf{x}\mathbf{y}') = E\left[\mathbf{x}(\mathbf{G}\mathbf{x}+\mathbf{e})'\right]$$

$$= E\left[\mathbf{x}\mathbf{x}'\mathbf{G}' + \mathbf{x}\mathbf{e}'\right]$$

$$= E(\mathbf{x}\mathbf{x}')\mathbf{G}' + E(\mathbf{x}\boldsymbol{\zeta}')(\mathbf{I}-\mathbf{B}')^{-1}$$

$$= \mathbf{\Phi}\mathbf{G}' + \mathbf{0} = \mathbf{\Phi}\mathbf{\Gamma}'(\mathbf{I}-\mathbf{B}')^{-1} .$$

The null matrix appears above because we have previously assumed that $E(\mathbf{x}\boldsymbol{\zeta}') = \mathbf{0}$ [in equation (10.2)], that is to say the x variables are not correlated with the errors in the equations. Putting all the pieces together,

$$\mathbf{\Sigma} = \left[\begin{array}{c|c} \mathbf{\Sigma}_{yy} & - \\ \hline \mathbf{\Sigma}_{xy} & \mathbf{\Sigma}_{xx} \end{array}\right]$$

$$= \left[\begin{array}{c|c} (I-B)^{-1}[\Gamma\Phi\Gamma' + \Psi](I-B')^{-1} & - \\ \hline \Phi\Gamma'(I-B')^{-1} & \Phi \end{array}\right] . \qquad (10.9)$$

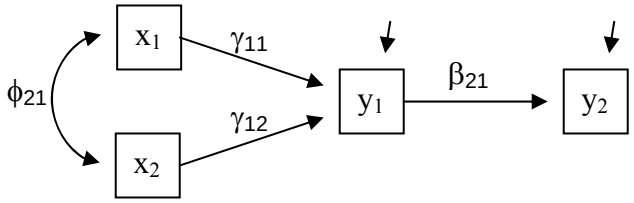
The structure above constitutes H_0 and H_A: $\mathbf{\Sigma} = \mathbf{S}$ is as before in Chapter 9.

10.2 *A Simple Example with Four Variables*

At this point I would like you to imagine that we have measured the following four variables:

Variable	Description
x_1	Perceived Attractiveness of Product
x_2	Perceived Cost of Product
y_1	Intention to Purchase
y_2	Purchasing Behavior

Now let us look at the path diagram for a causal model.



There are a few things we might note about this diagram. As is the tradition with confirmatory factor analysis, we usually leave off a label for errors; they are just represented as single headed unlabeled arrows. Covariances, such as the one between x_1 and x_2, are represented by two-headed arrows. Causal paths are represented by one-headed arrows. By tradition, the variances of the exogenous variables do not appear on path diagrams.

The structural equations for this model are

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \zeta_1$$
$$y_2 = \beta_{21}y_1 + \zeta_2$$

and in matrix terms

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}, \text{ i. e.}$$

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\zeta}.$$

In addition we need to specify the variances of any variable appearing on the right hand side:

$$V(\boldsymbol{\zeta}) = \begin{bmatrix} \psi_{11} & - \\ 0 & \psi_{22} \end{bmatrix}, \quad \text{and}$$

$$V(\mathbf{x}) = \begin{bmatrix} \phi_{11} & - \\ \phi_{21} & \phi_{22} \end{bmatrix} = \mathbf{S}_{xx} = \mathbf{\Phi}.$$

Since the x's are exogenous, their variances and covariances are given, and are estimated by the sample values. Thus they cannot contribute to the falsification of the model. Counting up all the free parameters, we have 1 β, 2 γ's, 2 ψ's and 3 ϕ's. There are $(4\cdot5)/2 = 10$ data values, leaving 2 degrees of freedom for the model. This can be seen in the path diagram by the fact that there are two missing arrows; the arrow that does not appear between x_1 and y_2, and the arrow not present between x_2 and y_2. It is actually these two missing arrows that are being tested by the Chi Square statistic for this model. Their absence is what we can falsify using the SEM technique.

10.3 *All y Models*

Any model that can be expressed with x and y variables can be expressed with y variables alone. Consider the following two sets of equations,

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$$

$$\mathbf{x} = \mathbf{0}\mathbf{y} + \mathbf{I}\mathbf{x} + \mathbf{0}\ ,$$

where the second set of equations, involving the x variables, is present just to create a similarity between the x's and y's. In fact, the second set really just sets $\mathbf{x} = \mathbf{x}$! Now define

$$\mathbf{z} = \begin{bmatrix} \mathbf{y} \\ - \\ \mathbf{x} \end{bmatrix},$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{B} & \mathbf{\Gamma} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \quad \text{and}$$

$$\mathbf{e} = \begin{bmatrix} \boldsymbol{\zeta} \\ \mathbf{0} \end{bmatrix}$$

so that we can rewrite the two sets of structural equations

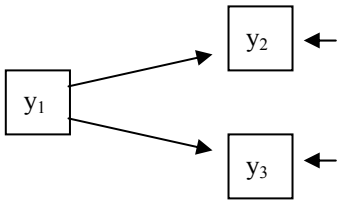
$$\mathbf{z} = \mathbf{G}\mathbf{z} + \mathbf{e} \quad \text{with}$$

$$\mathrm{V}(\mathbf{e}) = \begin{bmatrix} \Psi & - \\ \mathbf{0} & \mathbf{\Phi} \end{bmatrix} = \mathbf{A}\ .$$

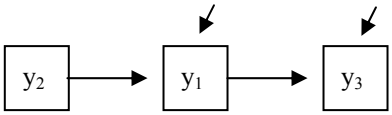
We define **z**, **G** and **A** temporarily just to illustrate the point. The point being that we need only one set of variables with one regression coefficient matrix and one variance matrix. It is most convenient to use **y**, **B** and $\mathbf{\Psi}$ to play these roles.

10.4 *In What Sense Are These Causal Models?*

Using Structural Equation models we have the potential to reject the hypothesis H_0 that embodies the causal model. Rejecting H_0 is a definitive event. If H_0 is not rejected, the results are a bit more ambiguous. All we can say in that case is that we have failed to reject the hypothesis. In other words, it is still in contention but by no means can it be considered proven. In point of fact, there are an infinite number of other possible models that could also be true. H_0 is merely among the survivors. To illustrate this point, consider the two causal structures below:



and



Note that the path diagrams above have been simplified somewhat from the traditional conventions. Both models have one degree of freedom that corresponds to the missing path between y_2 and y_3. In point of fact, the one degree of freedom, or the restriction implied by that degree of freedom, is identical in both cases. To explore the nature of this restriction, we revisit Section 5.8. Consider the regressions

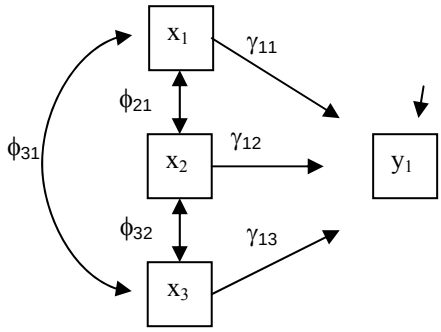
$$y_2 = y_1 + e_2 \text{ and}$$

$$y_3 = y_1 + e_3.$$

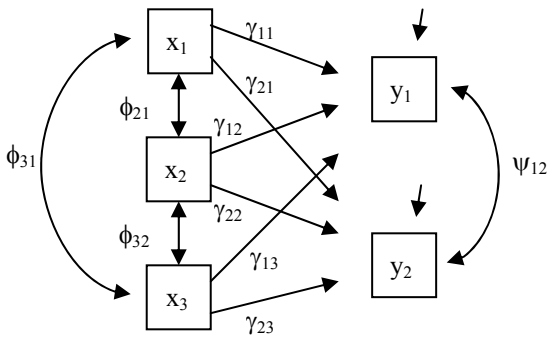
Both causal diagrams require only that the partial covariance $\sigma_{23\cdot1} = 0$ where $\sigma_{23\cdot1}$ is the $\text{Cov}(e_2, e_3)$ from the above two regression equations. Failure to reject does not prove your model.

10.5 *Regression As a Structural Equation Model*

Consider a regression model with three independent variables and one dependent variable. The path diagram for this appears below;



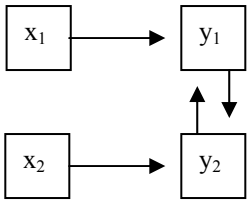
Now let us count up degrees of freedom for the model. We have six elements in the $\mathbf{\Phi}$ matrix (remember that the variances of the exogenous variables do not appear on a path diagram), there are three γ values, and one ψ. Among the four observed variables there are $4(5)/2 = 10$ covariances and variances. Thus there are exactly as many free parameters as there are data points. In effect, the parameters are just transformations of the data. We say in this case that the model is *just identified*. The model does not impose any restrictions on the $\mathbf{\Sigma}$ matrix, which is to say that it has 0 degrees of freedom. Now lets look at the multivariate case with multiple dependent variables. For example, below we can see a model with two y variables and three x variables:



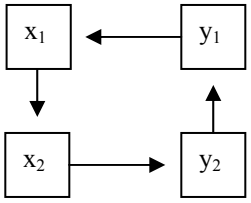
We leave it to the reader to calculate the degrees of freedom in this case and to verify that here too, we will end up with a just identified structural equation model. Regression, whether it is with one or more dependent variables, is not falsifiable in the sense of structural equation modeling. Regression is not a causal model.

10.6 *Recursive and Nonrecursive Models*

At this point we need to learn an important term that unfortunately sounds as if it means the opposite of what it actually means. A *recursive system* is characterized by $V(\boldsymbol{\zeta}) = \boldsymbol{\Psi}$ diagonal, and by the fact that it is possible to arrange the y variables so that **B** is lower (or upper) triangular. It is probably easier to illustrate the concept of a recursivity by referring to its opposite. Some example systems that are *non-recursive* are shown below.



and



Both of these would be called non-recursive. Generally, non-recursive models can be very difficult to estimate using structural equation models. There are certain specialized econometric techniques, discussed in Chapter 17, specially constructed to facilitate these sorts of models.

10.7 *Structural Equation Models with Latent Variables*

It is possible to combine the latent variables models of Chapter 9 with the structural equation models of this chapter. In other words, we can have path models between factors. While we have already shown we can always get by with just y-variables, here, if only for notational clarity, we will assume we have two sets of variables, an x set and a y set, and therefore we need two measurement models,

$$\mathbf{y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{10.10}$$

$$\mathbf{x} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \,. \tag{10.11}$$

The y-variables are a function of certain latent variables, the η's, while the x-variables are a function of other latent variables, the ξ's. The next step would be that we can have structural equation models amongst these latent variables as below:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \ . \tag{10.12}$$

Needless to say, there are a set of assumptions that we must make before we can use these models. These are listed now

$$\text{Cov}\ (\boldsymbol{\eta}, \boldsymbol{\varepsilon}) = \mathbf{0} \tag{10.13}$$

$$\text{Cov}\ (\boldsymbol{\xi}, \boldsymbol{\delta}) = \mathbf{0} \tag{10.14}$$

$$\text{Cov}\ (\boldsymbol{\xi}, \boldsymbol{\zeta}) = \mathbf{0} \tag{10.15}$$

$$\text{Cov}\ (\boldsymbol{\varepsilon}, \boldsymbol{\delta}, \boldsymbol{\zeta}) = \mathbf{0} \tag{10.16}$$

$$\text{Diag}\ (\mathbf{B}) = \mathbf{0} \tag{10.17}$$

$$|\ \mathbf{I} - \mathbf{B}| \neq 0 \ . \tag{10.18}$$

The first two assumptions in equations (10.13) and (10.14) are that the common factors and the unique factors are independent. In the structural equation model, the independent variable and the error must be uncorrelated [assumption (10.15)]. Each of the three types of errors are mutually uncorrelated [assumption (10.16)] . The diagonal of the **B** matrix is a set of p zeroes, and the expression (**I** – **B**) must be nonsingular, meaning that its determinant cannot be zero so that it can be inverted (as is discussed in Section 1.8).

To review, we have now introduced four parameter matrices: $\boldsymbol{\Lambda}_y$ which contains factor loadings for y variables, $\boldsymbol{\Lambda}_x$ which contains loadings for the regression of x variables on their factors, the ξ's, $\boldsymbol{\Gamma}$ containing regression coefficients for $\boldsymbol{\eta}$ on $\boldsymbol{\xi}$, and **B** with the regression coefficients for η's on other η's. To round out the picture, we have four variance matrices. The variance of all inputs must be specified, and that includes

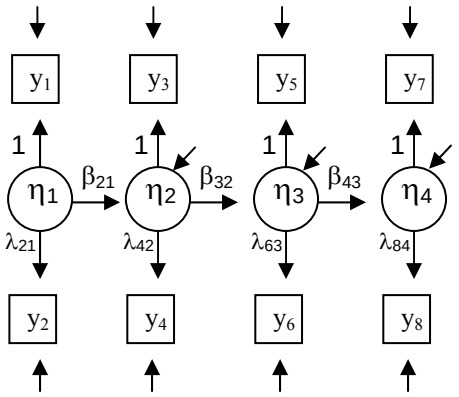
$$V(\boldsymbol{\xi}) = \boldsymbol{\Phi}, \tag{10.19}$$

$$V(\boldsymbol{\zeta}) = \boldsymbol{\Psi}, \tag{10.20}$$

$$V(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}_{\varepsilon} \ \text{and} \tag{10.21}$$

$$V(\boldsymbol{\delta}) = \boldsymbol{\Theta}_{\delta} \ . \tag{10.22}$$

Our first example involves a longitudinal study in which a group of customers is asked the same two items on four different purchase occasions. These two items are hypothesized to be unidimensional. Here, we have to admit that this is just an illustrative example since any two items are unidimensional! You need more than two items to create a scale, otherwise you are just modeling a plain household variety correlation. But, proceeding anyway, here is the path diagram:



Perhaps we are interested in the persistence of the attitude towards the brand over time. All of the variables have been labeled in such a way as to illustrate an all-y model. The measurement model is

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \lambda_{42} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_{63} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \lambda_{84} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{bmatrix}$$

with the structural model

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \beta_{21} & 0 & 0 & 0 \\ 0 & \beta_{32} & 0 & 0 \\ 0 & 0 & \beta_{43} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \end{bmatrix}.$$

To this we add the variance matrices of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\zeta}$, respectively,

$$\boldsymbol{\Theta}_\varepsilon = \begin{bmatrix} \theta_{11} & 0 & \cdots & 0 \\ 0 & \theta_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \theta_{88} \end{bmatrix} \text{ and}$$

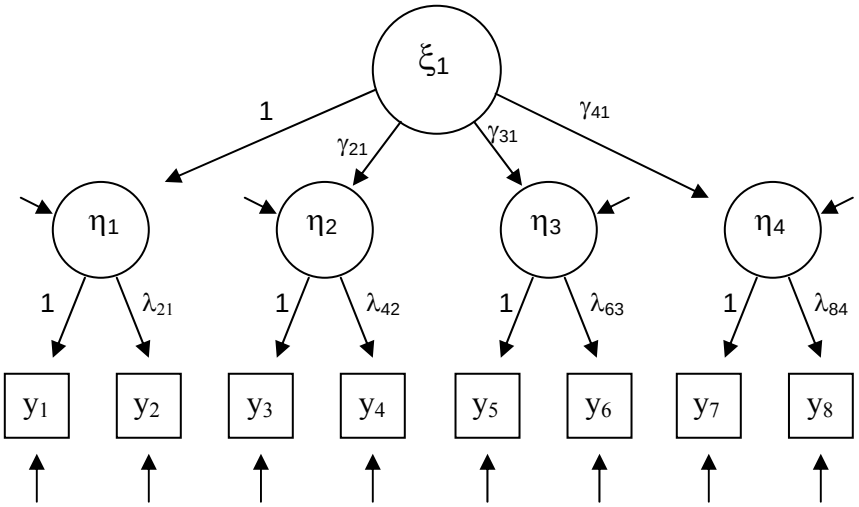
$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & 0 & 0 & 0 \\ 0 & \psi_{22} & 0 & 0 \\ 0 & 0 & \psi_{33} & 0 \\ 0 & 0 & 0 & \psi_{44} \end{bmatrix}.$$

In the $\boldsymbol{\Psi}$ matrix, the parameter ψ_{11} is exogenous.

It is important to be able to calculate the degrees of freedom for this or any other model you are working on. The raw data for the model, given that there are eight observed variables, is given by the expression 9(8)/2 = 36. From this we must subtract the four free elements in the loading matrix, three β's, eight elements in $\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}$ and then four elements on the diagonal of $\boldsymbol{\Psi}$. This leads to 15 degrees of freedom.

10.8 *Second Order Factor Analysis*

One very beautiful, if rarely applied, model is the second order factor model. In effect, the factors themselves may form a higher order factor. In other words, if the correlations amongst the factors have the right structure, these may be the result of a latent variable. A path diagram of this model appears below:



Note that the η's have their own loadings and their own unique factors. Here, the variable ξ_1 serves as the higher order factor. In general terms, the second order factor analysis model can be written as

$$\mathbf{y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \text{ and} \tag{10.23}$$

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \quad . \tag{10.24}$$

which the reader will recognize as a special case of a SEM with latent variables. We can write the model more compactly as

$$\mathbf{y} = \boldsymbol{\Lambda}_y\left[\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}\right] + \boldsymbol{\varepsilon} \ . \tag{10.25}$$

We need to assume that Cov($\boldsymbol{\varepsilon}$, $\boldsymbol{\zeta}$) = $\mathbf{0}$ and Cov($\boldsymbol{\xi}$, $\boldsymbol{\zeta}$) = $\mathbf{0}$. Here we also have V($\boldsymbol{\varepsilon}$) = $\boldsymbol{\Theta}_{\varepsilon}$, V($\boldsymbol{\zeta}$) = $\boldsymbol{\Psi}$ and V($\boldsymbol{\xi}$) = $\boldsymbol{\Phi}$. The variance matrix of y, $\boldsymbol{\Sigma}$, takes on a particularly aesthetic form with this model,

$$V(\mathbf{y}) = \boldsymbol{\Lambda}_y\left[\boldsymbol{\Gamma\Phi\Gamma}' + \boldsymbol{\Psi}\right]\boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_{\varepsilon} \quad , \tag{10.26}$$

with the internal part in the brackets being the V($\boldsymbol{\eta}$). Again, students should make certain they can calculate the degrees of freedom for this model.

10.9 *Models with Structured Means*

In order to look at means, something that is useful especially when there are multiple groups, we need to include a unit vector as an "independent variable" and analyze the raw SSCP matrix [see Equation (2.9)] instead of a covariance matrix. Our model is

$$\mathbf{y} = \boldsymbol{\nu}_y + \boldsymbol{\Lambda}_y\,\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{10.27}$$

$$\mathbf{x} = \boldsymbol{\nu}_x + \boldsymbol{\Lambda}_x\,\boldsymbol{\xi} + \boldsymbol{\delta} \tag{10.28}$$

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}\,\boldsymbol{\eta} + \boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\zeta}\,. \tag{10.29}$$

Now define E($\boldsymbol{\xi}$) = $\boldsymbol{\kappa}$. Then

$$E(\boldsymbol{\eta}) = (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma\kappa}), \tag{10.30}$$

$$E(\mathbf{x}) = \boldsymbol{\nu}_x + \boldsymbol{\Lambda}_x\,\boldsymbol{\kappa} \text{ and} \tag{10.31}$$

$$E(\mathbf{y}) = \boldsymbol{\nu}_y + \boldsymbol{\Lambda}_y\,(\mathbf{I} - \mathbf{B})^{-1}\,(\boldsymbol{\alpha} + \boldsymbol{\Gamma\kappa}) \quad . \tag{10.32}$$

In order to fit this model in the context of a SEM, we need to include a vector of 1's that we will call x_0. It will be the only variable labeled as an x. For the rest of the real x's and the y's, we will utilize an all-y model. For x_0 we have

$$\mathbf{1} = \mathbf{1}\xi_0 + \mathbf{0} \quad .$$

For all of the rest of the variables, we have

$$\begin{bmatrix}\mathbf{y}\\ \hline \mathbf{x}\end{bmatrix} = \left[\begin{array}{c|c|c} \boldsymbol{\Lambda}_y & \mathbf{0} & \boldsymbol{\nu}_y \\ \hline \mathbf{0} & \boldsymbol{\Lambda}_x & \boldsymbol{\nu}_x \end{array}\right]\begin{bmatrix}\boldsymbol{\eta}\\ \hline \boldsymbol{\xi}\\ \hline \mathbf{1}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}\\ \hline \boldsymbol{\delta}\end{bmatrix} \tag{10.33}$$

as the measurement model. The structural equation model looks like

$$\begin{bmatrix}\boldsymbol{\eta}\\ \hline \boldsymbol{\xi}\\ \hline 1\end{bmatrix} = \left[\begin{array}{c|c|c} \mathbf{B} & \boldsymbol{\Gamma} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0}' & \mathbf{0}' & 0 \end{array}\right]\begin{bmatrix}\boldsymbol{\eta}\\ \hline \boldsymbol{\xi}\\ \hline 1\end{bmatrix} + \begin{bmatrix}\boldsymbol{\alpha}\\ \hline \boldsymbol{\kappa}\\ \hline 1\end{bmatrix} 1 + \begin{bmatrix}\boldsymbol{\zeta}\\ \hline \boldsymbol{\xi} - \boldsymbol{\kappa}\\ \hline 0\end{bmatrix} \quad . \tag{10.34}$$

The means of the latent variables, the α_i, show up in the position usually occupied by the "$\boldsymbol{\Gamma}$" matrix, which in this case is a vector. There is a sequence of hypotheses and models that can be tested. If we assume there are two groups, we would start by testing

$$H_0: \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2. \tag{10.35}$$

Failure to reject this hypothesis implies that we should pool the groups. At this point any between group analysis stops.

$$H_0: \boldsymbol{\Lambda}_y^{(1)} = \boldsymbol{\Lambda}_y^{(2)} \tag{10.36}$$

Failure to reject the hypothesis in Equation (10.36) implies each population has the same factor structure. Otherwise, if you reject this hypothesis, it doesn't make sense to compare factor means across groups because these means correspond to different factors in the two groups. Therefore if we reject the hypothesis of Equation (10.36), between group analysis stops.

$$H_0: \boldsymbol{\nu}_y^{(1)} = \boldsymbol{\nu}_y^{(2)} \tag{10.37}$$

Failure to reject the above hypothesis implies that the items work the same way in each population. If you reject it, between group comparison stops.

$$H_0 : \boldsymbol{\Theta}_\varepsilon^{(1)} = \boldsymbol{\Theta}_\varepsilon^{(2)}$$

There are no consequences of either rejecting or failing to reject the above hypothesis. However, as always, we should seek to end up with the simplest model possible so failing to reject this one would be considered positive.

$$H_0 : \boldsymbol{\alpha}^{(1)} = \boldsymbol{\alpha}^{(2)} \tag{10.38}$$

This would ordinarily be considered the key hypothesis. Do the groups vary on the factor means? Finally, we could look at

$$H_0 : \boldsymbol{\Psi}^{(1)} = \boldsymbol{\Psi}^{(2)} \tag{10.39}$$

which asks whether the groups differ on the factor space.

References

Joreskog, Karl Gustav (1970). A General Method for Analysis of Covariance Structures. *Biometrika*. 57 (2), 239-51.

Kenny, David. A (1979) Correlation and Causality. New York: Wiley.

Chapter 11: Exploratory Factor Analysis

Prerequisites: Chapter 9, Sections 3.5 - 3.8

11.1 *Some Comments on the History of Factor Analysis*

In this chapter we are going to cover a set of techniques known as *Exploratory Factor Analysis*. Originally, these techniques were simply known as factor analysis, but when Confirmatory Factor Analysis was invented, the word "Exploratory" was added so as to differentiate the two types of factor analysis. At this point we will be briefly reviewing the basic factor analysis model. The derivation of that model is done with more detail in Chapter 9. The difference between exploratory and confirmatory analyses is partly stylistic. For one thing, in exploratory analysis it is traditional to use a correlation matrix instead of a covariance matrix. In that case, the model specifies that

$$\begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ z_p \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \cdots \\ \eta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_m \end{bmatrix}$$

$$\mathbf{z} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}\,. \tag{11.1}$$

Using Theorem (4.9) we can easily show that

$$\hat{\mathbf{R}} = \mathrm{E}\left[(\boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon})(\boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon})'\right]$$

$$= \boldsymbol{\Lambda}\,\mathrm{E}(\boldsymbol{\eta}\boldsymbol{\eta}')\,\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\,\mathrm{E}(\boldsymbol{\eta}\boldsymbol{\varepsilon}') + \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\eta}')\,\boldsymbol{\Lambda}' + \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}').$$

Defining $\mathrm{E}(\boldsymbol{\eta}\boldsymbol{\eta}') = \mathrm{V}(\boldsymbol{\eta}) = \boldsymbol{\Psi}$, $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \mathrm{V}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$, and knowing that the unique factor vector $\boldsymbol{\varepsilon}$ is independent of the common factors in the vector $\boldsymbol{\eta}$, we can conclude that

$$\hat{\mathbf{R}} = \boldsymbol{\Lambda}\,\boldsymbol{\Psi}\,\boldsymbol{\Lambda}' + \boldsymbol{\Theta}. \tag{11.2}$$

Thus the presence of unmeasured variables can be revealed by a particular structure in the observed correlation matrix. There are a variety of ways of uncovering the structure revealed in Equation (11.2), many of which were invented long before computers. In general, there are two steps involved in doing this. In the first step, the factors are extracted, but in an arbitrary way where the regression weights in $\boldsymbol{\Lambda}$ are generally not interpretable. In a second step, the factors are rotated into an orientation that is more interpretable and hopefully in alignment with theoretical expectations. This is all in contrast to the confirmatory approach, where we hypothesize a certain alignment of the loadings from the beginning, and test the proposed model.

One of the earliest ways, and still the most popular method of factor extraction, is called *Principal Factors*. We begin our discussion with that technique.

11.2 Principal Factors Factor Extraction

We will begin with the simplifying assumption that the unobserved factors are z-scores and are also uncorrelated. In that case $\boldsymbol{\Psi} = \mathbf{I}$ and the model of Equation (11.2) simplifies to

$$\hat{\mathbf{R}} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}.$$

The part of the correlation matrix due to the common factors, call it **R***, is given by

$$\hat{\mathbf{R}}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'. \tag{11.3}$$

The off-diagonal elements of $\hat{\mathbf{R}}$ and $\hat{\mathbf{R}}^*$ are identical since $\boldsymbol{\Theta}$ is diagonal. The $\boldsymbol{\Theta}$ matrix must be diagonal, being the covariance matrix of the unique factors, and "unique" after all, describes a set of independent factors. However, $\hat{\mathbf{R}}$ and $\hat{\mathbf{R}}^*$ do differ on the diagonal. Whereas **R** has unities on the diagonal, **R*** has the proportion of the variance of each variable that it has in common with the other variables. This proportion is known as the *communality* of the variable. A quick look at **R*** reveals it to appear as below

$$\mathbf{R}^* = \begin{bmatrix} h_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & h_2^2 & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & h_p^2 \end{bmatrix}$$

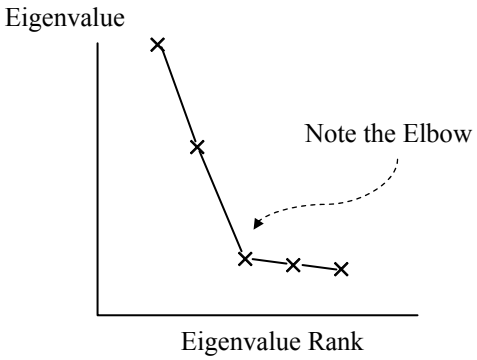
with h_i^2 being the communality of variable i. The goal of principle factors is to extract factors from **R*** in such a way as to explain the maximum amount of variance. Extracting the maximum amount of variance is also the goal of eigenstructure, as discussed in Section 3.5. Principle Factors is a technique that uses the eigenstructure of the **R*** matrix. But before we can proceed, we have to answer two related questions.

1. What are the values of the h_i^2?
2. How many factors are there?

If we knew how many factors there were, we could extract that many eigenvalues and eigenvectors from **R**, reproduce $\hat{\mathbf{R}}$ using the eigenvalues and eigenvectors, then look at the diagonal of this reproduced correlation matrix. Conversely, if we knew what the communalities were, we could deduce the number of factors because while the rank (see Section 3.7) of $\hat{\mathbf{R}}$ is p, the rank of $\hat{\mathbf{R}}^*$ depends on m, the number of factors as can be seen in Equation (11.3). $\hat{\mathbf{R}}^*$ is an outer product [Equation (1.21)] with a rank no greater than the number of columns of $\boldsymbol{\Lambda}$. Therefore the number of non-zero eigenvalues of $\hat{\mathbf{R}}^*$ would tells us exactly how many factors there are. So which comes first: the chicken in the form of the values of the h_i^2, or the egg in the form of the number of factors?

Even though this is called exploratory factor analysis, we would normally begin with some notion of m, the number of factors. This notion might come from substantive theory or from an educated guess. Another traditional method is to pick the number of factors based on the number of eigenvalues > 1. The logic here is that since an eigenvalue represents the variance of the factor, if a factor does not explain even as much as a single observed variable, it is not really pulling its weight.

Another approach is to use a so-called *Scree Chart*.



Given the Scree Chart above, we would pick m = 3 and extract 3 eigenvalues. The third one represents an inflection point, after which there is not much change.

Even if we start with some determined number of factors, it is good to start off with good estimates of the communalities. Here we take advantage of the fact that the lower bound for the communality for a particular variable is the squared multiple correlation, R^2, introduced in Equation (6.21), when that variable is regressed on all the other variables. So we have the relationship

$$R_j^2 \le h_j^2 \le 1 \tag{11.4}$$

where R_j^2 is the R^2 value for variable j, chosen as the dependent variable with all other variables used as independent variables. A very simple computational formula for this is

$$R_j^2 = 1 - \frac{1}{r^{jj}} \tag{11.5}$$

where r^{jj} is the jth diagonal element of $\mathbf{R}^{-1}$, the inverse of the correlation matrix of all the variables.

We are now ready to discuss the steps of the algorithm known as Principal Factors. We begin with the observed correlation matrix, **R**. According to Equation (11.4), we then can either use the lower bound to the communality, the Squared Multiple Correlation, or use the upper bound, unity. In either case, we find the eigenstructure of **R***, and then reproduce that matrix using only the m largest eigenvalues and their corresponding eigenvectors, i. e.

$$\hat{\mathbf{R}}^* = \mathbf{XLX}' .$$

Here the columns of the matrix **X** contain the eigenvectors while the diagonal elements of **L** contain the eigenvalues. Now we need only define

$$\mathbf{\Lambda} = \mathbf{XL}^{1/2}$$

where the square root of a matrix, $\mathbf{L}^{1/2}$ is uniquely identified since **L** is a diagonal matrix containing the eigenvalues on the diagonal and zeroes elsewhere. Remembering the definition of the Diag function [Equation (2.13)] of a square matrix, by subtraction we can deduce that

$$\boldsymbol{\Theta} = \mathbf{I} - \text{Diag}(\hat{\mathbf{R}}^*).$$

Sometimes Principal Factors is iterated using the following steps.

Step 0. Find the m largest roots of $\mathbf{R}$. Calculate $\hat{\mathbf{R}} = (\mathbf{X}\mathbf{L}^{1/2})(\mathbf{X}\mathbf{L}^{1/2})' = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$.

Step 1. Set $\mathbf{R}^* = \mathbf{R} - [\mathbf{I} - \text{diag}(\hat{\mathbf{R}})]$.

Step 2. Find the m largest roots of $\mathbf{R}^*$, recalculate $\hat{\mathbf{R}} = (\mathbf{X}\mathbf{L}^{1/2})(\mathbf{X}\mathbf{L}^{1/2})' = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$. If $\hat{\mathbf{R}}$ is not changing from iteration to iteration, stop. Otherwise go back to Step 1.

In Step 0 we can start with unities on the diagonal of $\mathbf{R}$ and the process will converge down to the h_j^2, or you start with squared multiple correlations and converge up.

11.3 *Exploratory Factor Analysis Is a Special Case of Confirmatory*

Before the maximum likelihood approach to factor analysis was invented by Lawley (summarized in Lawley and Maxwell 1963), factor analysis existed as a purely descriptive technique. Now we know that exploratory factor analysis is a special case of the confirmatory model discussed in Chapter 9. To implement the special case, we fix the absolute minimum number of parameters necessary to identify the model. The absolute minimum number of parameters that must be fixed to identify an m-factor model is m^2. These need to be arranged in the $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ matrices in a certain way, however. If we set $\boldsymbol{\Psi} = \mathbf{I}$ this fixes $\frac{m(m+1)}{2}$ parameters leaving

$$\begin{aligned} m^2 - \frac{m(m+1)}{2} &= \frac{2m^2}{2} - \frac{m^2+m}{2} \\ &= \frac{m(m-1)}{2} \end{aligned} \tag{11.6}$$

restrictions. If you have no hypotheses, other than a hypothesis as to the number of factors, m, these restrictions may be arbitrarily placed in $\boldsymbol{\Lambda}$ with column i getting i - 1 zeroes at the top. For example, with m = 3 factors we have $V(\boldsymbol{\eta}) = \boldsymbol{\Psi} = \mathbf{I}$ which imposes $\frac{3(3+1)}{2} = 6$ restrictions. We need $\frac{3(3-1)}{2} = 3$ more restrictions to make $m^2 = 9$ all together. In that case we can arbitrarily build $\boldsymbol{\Lambda}$ as

$$\Lambda = \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \\ \cdots & \cdots & \cdots \\ \lambda_{p1} & \lambda_{p2} & \lambda_{p3} \end{bmatrix}.$$

The $\hat{\chi}^2$ tests the null hypothesis that $\mathbf{\Sigma}$ stems from 3 factors vs. the alternative that $\mathbf{\Sigma}$ is arbitrary, or it stems from as many factors as there are variables, p. Once dimensionality has been statistically determined, rotation may generate hypotheses (for later confirmation) regarding the nature of the dimensions. If there are more than m^2 fixed constants rotation is not possible and meaning that the factor space has been restricted.

11.4 Other Methods of Factor Extraction

In addition to ML factor analysis, we have an approach called MINRES which seeks to minimize the residual or the difference between the predicted correlations in $\hat{\mathbf{R}}$ and the actual correlations in $\mathbf{R}$. The objective function is then

$$f = \sum_{j=2}^{p} \sum_{k=1}^{j} \left(r_{jk} - \sum_{l}^{m} \lambda_{jl} \lambda_{kl} \right)^2 . \qquad (11.7)$$

The reader will note that the component $\sum_{l}^{m} \lambda_{jl} \lambda_{kl}$ is a scalar version of the inner product of the jth and kth rows of $\mathbf{\Lambda}$, since it would be those two rows used to reproduce element r_{jk} in Equation (11.3).

Canonical factoring maximizes the canonical correlation (see Section 8.18) between the factors and the variables while *Image factoring* and *Alpha factoring* are based on the notion that items measuring any particular factor are sampled from some population of items that might be chosen. These techniques are discussed in Harman (1976).

11.5 *Factor Rotation*

After the factors have been extracted, whether this be by ML Factor Analysis, Principal Factors, or one of the other techniques, it is possible to rotate the factor axes into a position of possible higher theoretical value. That this is possible can be easily proven by noting that the extraction of factors based on Equation (11.3) is essentially arbitrary. If I define an orthonormal matrix $\mathbf{C}$ such that $\mathbf{CC}' = \mathbf{I}$, I can always create a new loading matrix, call it $\tilde{\mathbf{\Lambda}}$, as in

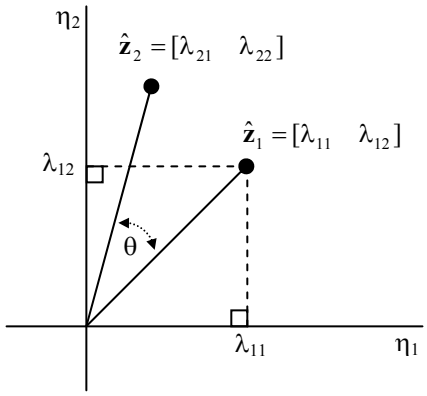
$$\tilde{\mathbf{\Lambda}} = \mathbf{\Lambda C}$$

so that

$$\mathbf{R}^* = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}' + \mathbf{\Theta}$$

$$= \mathbf{\Lambda CC}'\mathbf{\Lambda} + \mathbf{\Theta}$$

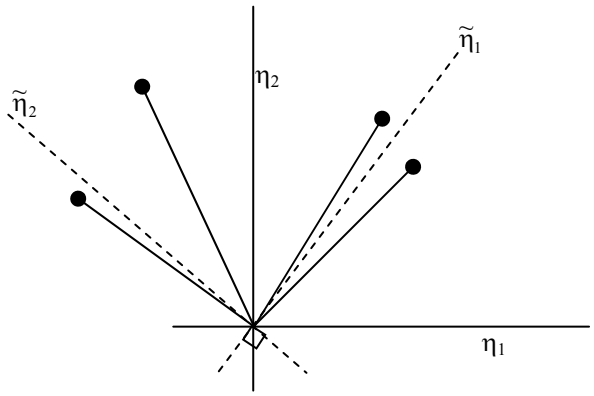
which of course yields the original Equation (11.3). An orthonormal matrix like **C** imposes a rigid rotation on axes which leaves angles and distances between points unchanged. The geometry of the situation might look like the figure below which shows two variables defined in a space with two factors.



Looking at the Figure, the length of the vector $\hat{\mathbf{z}}_j$ that corresponds to the (predicted) variable j is $\sqrt{\hat{\mathbf{z}}_j'\hat{\mathbf{z}}_j} = \sqrt{\sum_k \lambda_{jk}} = \sqrt{h_j^2}$. Our factors are at right angles, which is to say uncorrelated. At this point, assuming we have extracted those two factors using Principal Factors, the position of the axes is in the arbitrary orientation of maximal variance. The loadings are now the coordinates of the variables on the axes formed by the factors. The predicted correlation between the two variables is

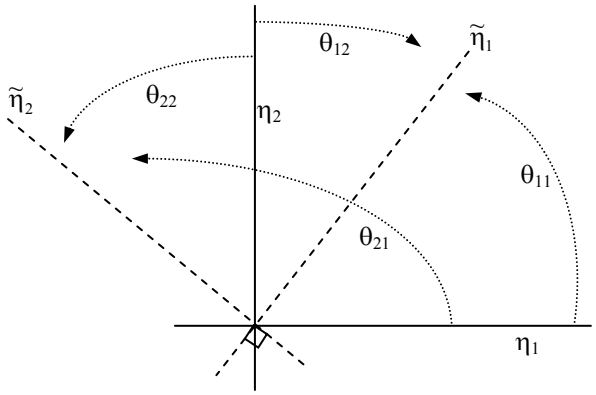
$$\hat{r}_{12} = \sqrt{h_1^2}\sqrt{h_2^2}\cos\theta. \tag{11.8}$$

In the next figure, we complicate things somewhat by having four variables appear.



All variables load on all original axes, η_1 and η_2. However, the loadings or coordinates on the new axes, $\tilde{\eta}_1$ and $\tilde{\eta}_2$ will be different. Two of the variables will load heavily on $\tilde{\eta}_1$ while the other

two will load on $\tilde{\eta}_2$. The cross loadings will be minimal which creates a much simpler $\mathbf{\Lambda}$ matrix in which the interpretation of the η's will be facilitated. In order to rotate the original axes into the new positions, we will need a bit of trigonometry. Below we have labeled all of the angles between each original axis and each new one:



We can construct the orthonormal rotation matrix **C** such that

$$\mathbf{C} = \begin{bmatrix} \cos\theta_{11} & \cos\theta_{12} \\ \cos\theta_{21} & \cos\theta_{22} \end{bmatrix}$$

and even though θ_{12} is "reversed", since cos θ = cos (360 - θ), it comes out the same. Note that the first subscript refers to the new axis and the second to the old. This concept works in spaces of arbitrary dimensionality. So how do we pick the angle of rotation? What constitutes a good orientation of the axes with the variables? What we are looking for is called *simple structure*. This is an idea due to Thurstone [summarized in Thurstone (1935)] who came up with three principles.

1. Each row of $\mathbf{\Lambda}$ should have at least one zero.
2. Each column of $\mathbf{\Lambda}$ should have at least m zeroes.
3. For every pair of columns of $\mathbf{\Lambda}$ there should be at least m variables with zeroes in one column but not in the other.

The most famous implementation of rotation to simple structure is Kaiser's *Varimax procedure* that maximizes the variance of the squared loadings within each column. The original formula, sometimes still called raw varimax, is to pick the rotation that maximizes the variance of the squared loadings in each column j of $\mathbf{\Lambda}$

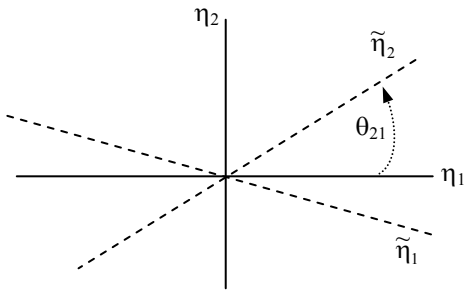
$$\left[\sum_{i}^{p} (\lambda_{ij}^2)^2 - \frac{\left(\sum_{i}^{p} \lambda_{ij}^2 \right)^2}{p} \right] \frac{1}{p} \qquad (11.9)$$

The formula widely used today (see Harman, 1976, pp. 290-1) weights each factor by the inverse of its total communality, but conceptually it follows the lines of the above equation. Other

approaches maximize the variance within each row (*Quartimax*), or equally between rows and columns (*Equimax*).

11.6 *Oblique Rotation*

Of course nothing guarantees that the factors that we see in marketing will be orthogonal. In order to create a rotation like the one pictured below,



we could use a transformation matrix

$$\mathbf{C} = \begin{bmatrix} \cos\theta_{11} & \cos\theta_{12} & \cdots & \cos\theta_{1m} \\ \cos\theta_{21} & \cos\theta_{22} & \cdots & \cos\theta_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \cos\theta_{m1} & \cos\theta_{m2} & \cdots & \cos\theta_{mm} \end{bmatrix}$$

which would take the old axes into a new set. Note that in this case $\mathbf{CC}' \neq \mathbf{I}$. The elements of $\mathbf{C}$ are direction cosines, and the sum of cross-products of direction cosines gives the cosine of the angle between the two vectors which according to Equation (11.8) is the same thing as a correlation. Thus we have for the correlations between the new factors:

$$\widetilde{\mathbf{\Psi}} = \mathbf{CC}' \tag{11.10}$$

The new loadings, in $\widetilde{\mathbf{\Lambda}}$, can be inferred from the fact that since

$$\hat{\mathbf{R}} = \mathbf{\Lambda\Lambda}' + \mathbf{\Theta}$$

it must also be the case that

$$\mathrm{R} = \widetilde{\mathbf{\Lambda}}\mathbf{CC}'\widetilde{\mathbf{\Lambda}}' + \mathbf{\Theta}$$

so that obviously $\mathbf{\Lambda} = \widetilde{\mathbf{\Lambda}}\mathbf{C}$, which then implies further that $\mathbf{C}^{-1}\mathbf{\Lambda} = \widetilde{\mathbf{\Lambda}}$. When factors are orthogonal, and we have standardized both the variables and the factors to be z-scores, the loadings in the $\mathbf{\Lambda}$ matrix can also be interpreted as correlations between the variables and the factors. When we have non-orthogonal factors, this is no longer so. We can, however, calculate these correlations, known as the *factor structure*, using

$$\mathbf{S} = \widetilde{\mathbf{\Lambda}}\mathbf{\Psi}. \qquad (11.11)$$

There are a number of analytic techniques available to perform oblique rotation including *Oblimax*, *Quartimin*, *Oblimin* and *Promax*.

References

Harman, Harry H. (1976) *Modern Factor Analysis. Third Edition.* Chicago: University of Chicago Press.

Jöreskog, Karl Gustav (1967) Some Contributions to Maximum Likelihood Factor Analysis. Psychometrika. 32 (December), 443-82.

Lawley, D.N. and A. E. Maxwell (1963) *Factor Analysis as a Statistical Method*. London: Butterworth.

Thurstone, L. L. (1935) *The Vectors of Mind*. Chicago: University of Chicago Press.

Section IV: Consumer Judgment and Choice

Chapter 12: Judgment and Choice

Prerequisites: Chapter 5, Sections 3.9, 3.10, 6.8

12.1 *Historical Antecedents*

In the 19th century Gustav Fechner attempted to understand how it is that humans perceive their world. The simplest place to start was by asking how it is that we perceive basic physical quantities such as the heaviness of a block of wood, the brightness of a light, or the loudness of a tone. He thought that there were three important elements behind the sequence by which the process operates:

(1) The external physical environment, which we will denote n
(2) Brain activity, which we will denote m, and
(3) Conscious perception, which we will denote s.

Fechner believed that the relationship between (2) and (3) was inaccessible to science, and that anyway, they were just two different ways of looking at the same phenomenon. On the other hand, the relationship between (1) and (2) was part of physics, or perhaps physiology. Here, he concluded that there was some sort of one-to-one correspondence. He decided that he would investigate the relationship between (1) and (3), and, some would argue, by doing so created the science that we call psychology. He was concerned therefore with the way that simple physical stimuli come to be perceived. He proposed the following law, now known as *Fechner's Law*:

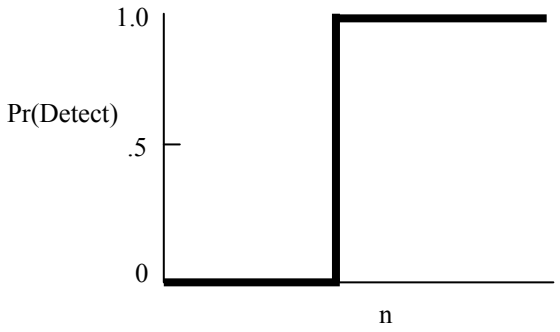
$$s = c \ln [n / n_0] \tag{12.1}$$

where s has been previously defined as the conscious perception of the loudness, brightness, or heaviness in question; and n the actual physical value of the stimulus. The constant c summarizes the sensitivity of the sense in question, while n_0 is the *absolute threshold*. The absolute threshold is the lowest limit of perception. For example, if we are talking about sounds, n_0 would be the softest sound detectable. The fact that Fechner used a log function is particularly meaningful. We can relate this to a variety of concepts, such as the economic notion of diminishing returns. The function predicts that proportional changes are equally important. In other words, if I am holding a one ounce block and I add 1/10th of an ounce of additional weight, this creates the same amount of perceived change as if I had a 1 pound block and I add 1/10th of a pound. This notion was later empirically verified by Weber who discovered that the size of a just noticeable difference was proportion to n,

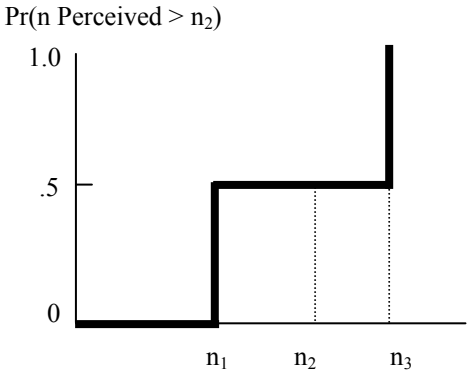
$$\Delta n = kn \tag{12.2}$$

where k quantifies the sensitivity of the sense for the observer.

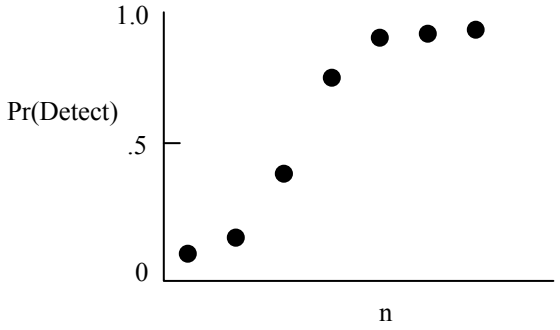
We now continue this historical review with the notion of absolute detection. We will say that the physical stimulus is measured in units of n, for example seconds, kilograms, centimeters, foot-candles, and so forth. In the 19th century it was imagined that there was a threshold, above which perception of the stimulus began, and below which there was no perception. Assuming we are dealing with brightness, it was assumed that as n increased, the conscious perception of the light popped suddenly into existence:



The position where this occurred was called the absolute threshold. A related experiment might have subjects compare two lights, and to make a judgment as to which was brighter. Then the question became one of *difference thresholds*, that is, a point above which the comparison light would be perceived of as identical and below which it would be perceived as dimmer, and another point, above which the comparison would be seen as brighter. The situation is pictured below.



We would say that the upward JND (Just Noticeable Difference) would be the interval $n_3 - n_2$ and the downward JND would be $n_2 - n_1$. Things did not turn out like the graphs pictured above, however. In fact, empirical data for the probability of detection revealed a much smoother function. An idealized example is given below:



How can we account for this?

12.2 *A Simple Model for Detecting Something*

Here we propose a simple model that says the psychological effect of a stimulus i is

$$s_i = \bar{s}_i + e_i \tag{12.3}$$

where $\bar{s}_i$ is the impact on the sense organ of the observer and e_i is random noise, perhaps added by the nervous system, the senses or by distraction. Let us assume further that, as in Section 4.2,

$$e_i \sim N(0, \sigma^2) \quad \text{so that}$$

$$s_i \sim N(\bar{s}_i, \sigma^2)\,. \tag{12.4}$$

Now, assume that there actually is a fixed threshold so that the subject detects the stimulus if $s_i \geq s_0$, i. e. the threshold is located at s_0. More formally we can write that

$$\Pr[\text{Detect stimulus i}] = \hat{p}_i = \Pr[s_i \geq s_0]\,. \tag{12.5}$$

At this point we need to establish a zero point for the psychological continuum, s, that we have created. It would be convenient if we set $s_0 = 0$. This psychological continuum is of course no more than an interval scale, and so its origin is arbitrary. We might as well place the zero point at a convenient place. In that case, we have

$$\hat{p}_i = \frac{1}{\sqrt{2\pi}\sigma}\int_0^{+\infty} \exp\left[-(s_i - \bar{s}_i)^2 / 2\sigma^2\right] ds_i\,. \tag{12.6}$$

Now we define $z = \dfrac{s_i - \bar{s}_i}{\sigma}$. In that case $dz/ds_i = 1/\sigma$ or $dz = ds_i/\sigma$. This will allow us to change the variable of integration, or in simple terms, switch everything to a standardized, z-score notation. This is shown below:

$$\hat{p}_i = \frac{1}{\sqrt{2\pi}\sigma}\int_0^{+\infty} \exp\left[-(s_i - \bar{s}_i)^2 / 2\sigma^2\right] ds_i$$

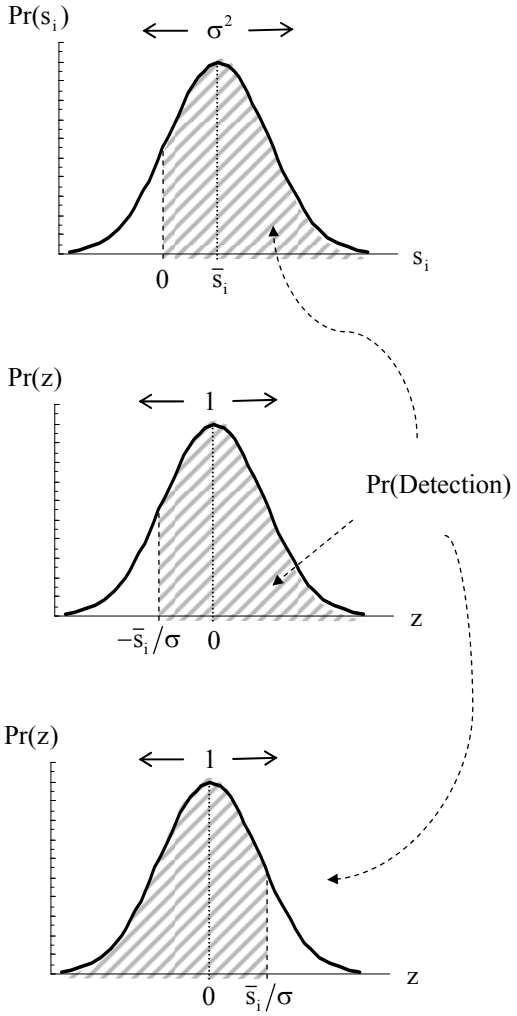
$$= \frac{1}{\sqrt{2\pi}}\int_{\frac{0-\bar{s}_i}{\sigma}}^{+\infty} \exp\left[\frac{-z^2}{2}\right] dz.$$

In words, as we go from the first line above to the last, we change from an "s-score" to a standardized z-score. In the first and second lines the integration begins at 0, but in the third line we have standardized so that we have subtracted the mean (from 0) and divided by σ. One last little change and we will have a very compact way to represent this probability. Since the normal distribution is symmetric, the area from +z to +∞ is identical to the area between -∞ and –z. In

terms of the equation above, the area between $\frac{0-\bar{s}_i}{\sigma}$ and $+\infty$ is then the same as that between $-\infty$ and $\frac{\bar{s}_i}{\sigma}$. We can therefore rewrite our detection probability as

$$\begin{aligned}\hat{p}_i &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\bar{s}_i}{\sigma}} \exp\left[\frac{-z^2}{2}\right] dz \\ &= \Phi[\bar{s}_i / \sigma],\end{aligned} \tag{12.7}$$

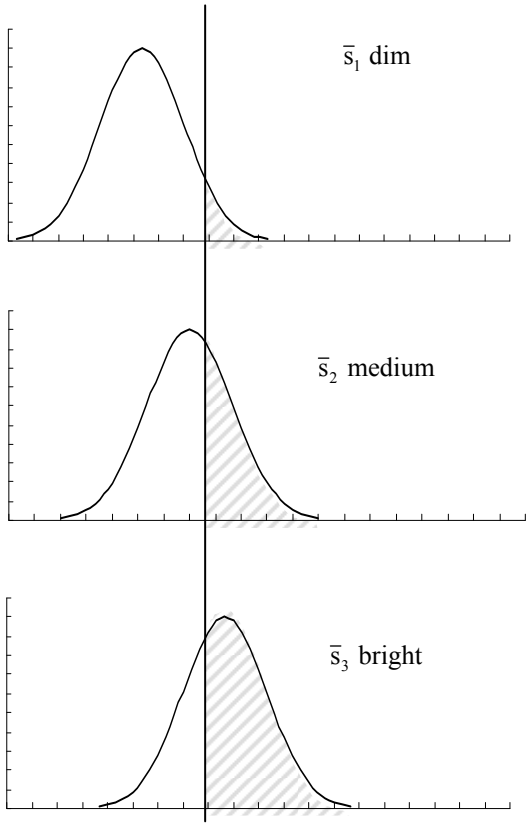
where $\Phi(\cdot)$ is the standard normal distribution function [see Equation (4.14)]. A graphical representation of all of this appears below.



We can now summarize two points; one general and one particular to the detection problem at hand. In general, we might remember that for any random variable, call it a, for which a ~ N [E(a), V(a)], then

$$\Pr [a \geq 0] = \Phi [E(a) / \sqrt{V(a)}] . \qquad (12.8)$$

In this particular case, s_i is playing the role of a, with $\bar{s}_i$ being E(a) and σ^2 being the V(a). And why do detection data not look like a step function? According to this model, they should look like a normal ogive. As the physical stimulus is varied, lets say by making it brighter and therefore easier to detect, $\bar{s}_i$ becomes larger and more and more of the distribution of s_i ends up being to the right of the threshold. This is illustrated in the figure below; with the shaded area representing the probability of detection of a light at three different intensities: dim, medium and bright.



12.3 *Thurstone's Law of Comparative Judgment*

In the previous section we have discussed how people can detect something such as a dim light in a darkened room, a slight noise in an otherwise silent studio, or a small amount of a particular smell. That experimental situation is called absolute judgment, and we modeled it by positing the existence of a fixed threshold plus normal random noise. Now let's contemplate how people compare two objects, for example, which of two wooden blocks are heavier, a procedure known as comparative judgment. In 1927 L. L. Thurstone published a paper in which he specified a model

for the comparative process, generalizing the work that had gone on before by extending his analysis to stimuli that did not have a specific physical referent. His chosen example was "the excellence of handwriting specimens." This sort of example must stand alone, in the sense that we cannot rely on some sort of physical measure to help us quantify excellence. To Thurstone, that did not really matter. He simply proposed that for any property for which we can compare two objects, there is some psychological continuum. And the process by which we react differently to the several comparison objects is just called the "discriminal process."

We should not let this slightly anachronistic language throw us off. Thurstone's contribution was fundamental and highly applicable to 21st century marketing. Suppose I ask you to compare two soft drinks and to tell me which one you prefer. This is the situation that Thurstone addressed. We can use his method to create interval scale values for each of the compared brands, even though we are only collecting ordinal data: which of the two brands is preferred by each subject. This is the essence of psychological scaling – use the weakest possible assumptions (i. e. people can make ordinal judgments of their preferences) and still end up with interval level parameters. In the case of preference judgments, these parameters are usually called *utilities*, based on the economic theory of rational man.

To create an interval scale, Thurstone borrowed a data collection technique called *paired comparisons*. In paired comparisons, a subject makes a judgment on each unique pair of brands. For example, with four brands; A, B, C and D the subject compares A and B, AC, AD, BC, BD and CD. In general there are $q = \frac{t(t-1)}{2}$ unique pairs among t brands. An additional point should be added here. For one, it turns out that just looking at pairs is not the most efficient way to scale the t brands. Despite this, the mathematics behind Thurstone's Law is very instructive.

Lets look at a miniature example of paired comparison data. Consider the table below where a typical entry represents the probability that the row brand is chosen over the column brand.

	A	B	C
A	-	.6	.7
B	.4	-	.2
C	.3	.8	-

Each table entry gives the Pr[Row brand is chosen over the Column brand]. Such a table is sometimes called antisymmetric, as element i, j and element j, i must sum to 1.0. As such, we can use the q non-redundant pairs that appear in the lower triangular portion of the table as input to the model

Another point is that there are two different ways to collect data. If the brands are relatively confusable, you can collect data using a single subject. Otherwise, the proportions that appear in the table are aggregated over a sample of individuals.

As before, we will assume that the process of judgment of a particular brand, such as brand i, leads to an output, call it s_i. We further hypothesize that for brand i

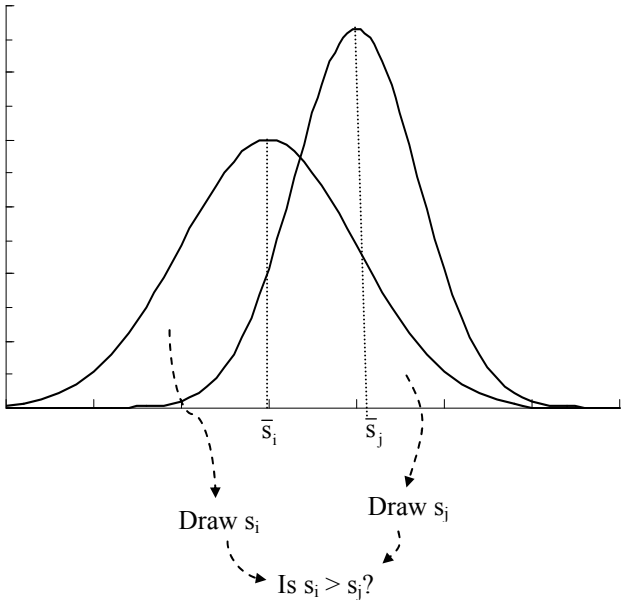
$$s_i = \bar{s}_i + e_i . \tag{12.9}$$

Similarly to what we did before in Equation (12.4), we further assume that

$$e_i \sim N(0, \sigma_i^2) \text{ with} \tag{12.10}$$

$$\text{Cov}(e_i, e_j) = \sigma_{ij} = r\sigma_i\sigma_j. \tag{12.11}$$

We hypothesize that brand i is chosen over brand j whenever $s_i > s_j$. This situation, which as shown below, bears a certain resemblance to a two sample *t*-test:



Now we can say that the probability that brand i is chosen over brand j

$$p_{ij} = \Pr(s_i > s_j) = \Pr(s_i - s_j > 0).$$

So how will we derive that probability? Turning back a bit in this chapter, recall Equation (12.8) which gave us an expression for the Pr(a > 0), namely $\Phi[E(a) / \sqrt{V(a)}]$, assuming that a is a normal variate. In the current case, the role of a is being played by $s_i - s_j$ and so we need to figure out $E(s_i - s_j)$ and $V(s_i - s_j)$. The expectation is simple.

$$E(s_i - s_j) = E\left[(\bar{s}_i + e_i) + (\bar{s}_j - e_j)\right]$$

$$= \bar{s}_i - \bar{s}_j,$$

since by our previous assumption the $E(e_i) = E(e_j) = 0$, and according to Theorem (4.4), the expectation of sum is the sum of the expectations. As far as the variance goes, we can use Theorem (4.9) to yield

$$V(s_i - s_j) = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_i^2 & \sigma_{ij} \\ \sigma_{ij} & \sigma_j^2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$= \sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}$$

$$= \sigma_i^2 + \sigma_j^2 - 2r\sigma_i\sigma_j .$$

At this point we have all of the pieces that we need to figure out the probability that one brand is chosen over the other. It is

$$\hat{p}_{ij} = \Pr(s_i > s_j) = \Phi\left[(\bar{s}_i - \bar{s}_j) \Big/ \sqrt{\sigma_i^2 + \sigma_j^2 - 2r\sigma_i\sigma_j} \right]. \tag{12.12}$$

Thurstone imagined a variety of cases for this derivation. In Case I, one subject provides all of the data as we have mentioned before. In Case II, each subject judges each pair once and the probabilities are built up across a sample of different responses. In Case III, r is assumed to be 0 (or 1, it doesn't matter), and in Cases IV and V all of the variances are equal – exactly in Case V and approximately in Case IV.

12.4 *Estimation of the Parameters in Thurstone's Case III: Least Squares and ML*

We will continue assuming Case III, meaning that each brand can have a different variance, but the correlations or the covariances of the brands are identical. By convention we tie down the metric of the discriminal dimension, s, by setting $\bar{s}_1 = 0$ and $\sigma_1^2 = 1$. We will now look at four methods to estimate the $2(t-1)$ unknown parameters in the model, namely, the values $\bar{s}_2, \bar{s}_3, \cdots, \bar{s}_t, \sigma_2^2, \sigma_3^2, \cdots, \sigma_t^2$. These methods are unweighted nonlinear least squares, weighted nonlinear least squares, modified minimum χ^2 and maximum likelihood. Unweighted nonlinear least squares begins with the observation that with the model,

$$\Pr(s_i > s_j) = \Phi\left[(\bar{s}_i - \bar{s}_j) \Big/ \sqrt{\sigma_i^2 + \sigma_j^2} \right],$$

we can use the inverse normal distribution function, $\Phi^{-1}(\cdot)$ on both sides. To understand what Φ^{-1} does, lets remember what the Φ function does – remember that Φ is the standard normal distribution function. For example, $\Phi(1.96) = .975$, and $\Phi(0) = .5$. If Φ takes a z score and gives you the probability of observing that score or less, Φ^{-1} takes a probability and gives you a z score. So $\Phi^{-1}(.975) = 1.96$, for example. What this means is that if we transform our choice probabilities into z scores, we can fit them with a model that looks like

$$\Phi^{-1}[\Pr(s_i > s_j)] = \Phi^{-1}\left\{ \Phi\left[(\bar{s}_i - \bar{s}_j) \Big/ \sqrt{\sigma_i^2 + \sigma_j^2} \right] \right\} \tag{12.13}$$

$$\hat{z}_{ij} = (\bar{s}_i - \bar{s}_j) \Big/ \sqrt{\sigma_i^2 + \sigma_j^2} ,$$

where $\hat{z}_{ij}$ is the predicted z score that corresponds to the choice that brand i is chosen over brand j. Now of course we have a string of such z scores, one for each of the q unique pairs,

$$\hat{z}_{12} = (\bar{s}_1 - \bar{s}_2) \Big/ \sqrt{\sigma_1^2 + \sigma_2^2}$$

$$\hat{z}_{13} = (\bar{s}_1 - \bar{s}_3) \Big/ \sqrt{\sigma_1^2 + \sigma_3^2}$$

$$\cdots = \cdots$$

$$\hat{z}_{(t-1)t} = (\bar{s}_{t-1} - \bar{s}_t) \Big/ \sqrt{\sigma_{t-1}^2 + \sigma_t^2} \ .$$

In unweighted nonlinear least squares we will have as a goal the minimization of the following objective function –

$$f = \sum_{i=1}^{t-1} \sum_{j=i+1}^{t} (z_{ij} - \hat{z}_{ij})^2 \tag{12.14}$$

where the summation is over all q unique pairs of brands. This technique is called unweighted because it does not make any special assumptions about the errors of prediction, in particular, assuming that they are equal or homogeneous. In general, this assumption is not tenable when we are dealing with probabilities, but this method is quick and dirty and works rather well. We can use nonlinear optimization (see Section 3.9) to pick the various $\bar{s}_i$ and σ_i^2 values which are unknown a priori and must be estimated from the sample. We do this by picking starting values for each of the unknowns and then evaluating the vector of the derivative of the objective function with respect to each of those unknowns. We want to set this derivative vector to the null vector as below,

$$\begin{bmatrix} \partial f/\partial \bar{s}_1 \\ \partial f/\partial \bar{s}_2 \\ \cdots \\ \partial f/\partial \bar{s}_t \\ \hline \partial f/\partial \sigma_1^2 \\ \partial f/\partial \sigma_2^2 \\ \cdots \\ \partial f/\partial \sigma_t^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ \hline 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}, \tag{12.15}$$

but we must do this iteratively, beginning with starting values and using these to evaluate the derivative. The derivative, or the slope, lets us know which way is "down", and we step off in that direction a given distance to come up with new, improved estimates. This process is repeated until the derivative is zero, meaning that we are at the bottom of the objective function, f.

The next approach also relies on nonlinear optimization and is called *weighted nonlinear least squares*, or in this case, it is also known as *Minimum Pearson* χ^2, since we will be minimizing the

classic Pearson Chi Square formula. We will not be transforming the data using $\Phi^{-1}(\cdot)$. Instead, we will leave everything as is, using the model formula

$$\hat{p}_{ij} = \Phi\left[(\bar{s}_i - \bar{s}_j)\Big/\sqrt{\sigma_i^2 + \sigma_j^2}\right].$$

Our goal is to pick the $\bar{s}_i$ and the σ_i^2 so as to minimize

$$\hat{\chi}^2 = \sum_i^t \sum_{j \neq i}^t \frac{(np_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} \tag{12.16}$$

which the reader should recognize as the formula for the Pearson Chi Square with $n\hat{p}_{ij}$ a different way of writing the expected frequency for cell i, j. Note that in the above formula, the summation is over all off-diagonal cells and that $p_{ji} = 1 - p_{ij}$ and of course $\hat{p}_{ji} = 1 - \hat{p}_{ij}$. As an alternative, we can utilize matrix notation to write the objective function. This will make clear the fact that minimum Pearson Chi Square is a GLS procedure as discussed in Section 6.8, although in the current case our model is nonlinear. Now define

$$\mathbf{p}' = \lfloor p_{12} \quad p_{13} \quad \cdots \quad p_{(t-1)t} \rfloor$$

and

$$\hat{\mathbf{p}}' = \lfloor \hat{p}_{12} \quad \hat{p}_{13} \quad \cdots \quad \hat{p}_{(t-1)t} \rfloor.$$

Note also that for each element in **p**,

$$V(p_{ij}) = V[p_{ij} - \hat{p}_{ij}] = \frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{n} \tag{12.17}$$

where n is the number of observations upon which the value p_{ij} is based. Using this information, we can create a diagonal matrix **V**, placing each of the terms $\hat{p}_{ij}(1 - \hat{p}_{ij})/n$ on the diagonal of **V** in the same order that we placed the pair choice probabilities in **p** and $\hat{\mathbf{p}}$. In that case we can say

$$V(\mathbf{p}) = \mathbf{V}. \tag{12.18}$$

Now we will minimize

$$\hat{\chi}^2 = (\mathbf{p} - \hat{\mathbf{p}})' \mathbf{V}^{-1} (\mathbf{p} - \hat{\mathbf{p}}) \tag{12.19}$$

which is equivalent to the previous equation for Chi Square, and which is a special case of Equation (6.23). This technique is called weighted nonlinear least squares so as to distinguish it from ordinary, or unweighted, least squares. Also, remember that the elements in $\hat{\mathbf{p}}$, that is, the predicted pair choice probabilities, are nonlinear functions of the unknowns, the $\bar{s}_i$ and σ_i^2. For this reason we would use the nonlinear optimization methods of Section 3.9 here as well.

A third method we have of estimating the unknown parameters in the Thurstone model is called *Modified Minimum* χ^2 or sometimes *Logit* χ^2. In this case the objective function differs only slightly from the previous case, substituting the observed data for the expectation or prediction in the denominator:

$$\hat{\chi}^2 = \sum_{i}^{t} \sum_{j \neq i}^{t} \frac{(np_{ij} - n\hat{p}_{ij})^2}{np_{ij}} . \tag{12.20}$$

This tends to simplify the derivatives and the calculations somewhat, but perhaps is not as necessary as it once was when computer time was more expensive than it is today.

Before we turn to Maximum Likelihood Estimation, it could be noted here that we might also use a *Generalized Nonlinear Least Squares* approach that takes into account the covariances between different pairs (Christoffersson 1975, p. 29)

$$\mathrm{Cov}(p_{ij}, p_{kl}) = \frac{p_{ijkl} - p_{ij} p_{kl}}{n} \tag{12.21}$$

where p_{ijkl} is the probability that a subject chose i over j and k over l. These covariances could be used in the off-diagonal elements of **V**.

Finally, we turn to *Maximum Likelihood* estimation of the unknowns. Here the goal is to pick the $\bar{s}_i$ and the σ_i^2 so as to maximize the likelihood of the sample. To begin, we define

$$f_{ij} = np_{ij}, \text{ that is, since}$$

$$p_{ij} = \frac{f_{ij}}{n},$$

i. e. the f_{ij} are the frequencies with which brand i is chosen over brand j. We also note that

$$p_{ji} = 1 - p_{ij} = \frac{n - f_{ij}}{n} .$$

We can now proceed to define the likelihood of the sample under the model as

$$\ell_0 = \prod_{i=1}^{t-1} \prod_{j=i+1}^{t} \hat{p}_{ij}^{f_{ij}} (1 - \hat{p}_{ij})^{n - f_{ij}} . \tag{12.22}$$

Note that with the two multiplication operators, the subscripts i and j run through each unique pair such that j > i. The log likelihood has its maximum at the same place as the likelihood. Taking logs on both sides leads to

$$\ln(\ell_0) = L_0 = \sum_{i=1}^{t-1} \sum_{j=i+1}^{t} f_{ij} \ln \hat{p}_{ij} + (n - f_{ij}) \ln(1 - \hat{p}_{ij}) \tag{12.23}$$

which is much easier to deal with, being additive in form rather than multiplicative. Note here that we have used the rule of logarithms given in Equation (3.1), and also the rule from Equation (3.3).

The expression L_0 gives the log likelihood under the model, assuming that the model holds. The probability of the data under the general alternative that the pattern of frequencies is arbitrary is

$$\ell_A = \prod_{i=1}^{t-1} \prod_{j=i+1}^{t} p_{ij}^{f_{ij}} (1 - p_{ij}^{n-f_{ij}}) . \tag{12.24}$$

Analogously to L_0, define L_A as $\ln(\ell_A)$. In that case

$$\hat{\chi}^2 = -2 \ln \frac{\ell_0}{\ell_A} = 2[L_A - L_0] . \tag{12.25}$$

Now we would need to figure out the derivatives of $\hat{\chi}^2$ with respect to each of the unknown parameters, the $\bar{s}_i$ and σ_i^2, and drive those derivatives to zero using nonlinear optimization as discussed in Section 3.9. When we reach that point we have our parameter estimates.

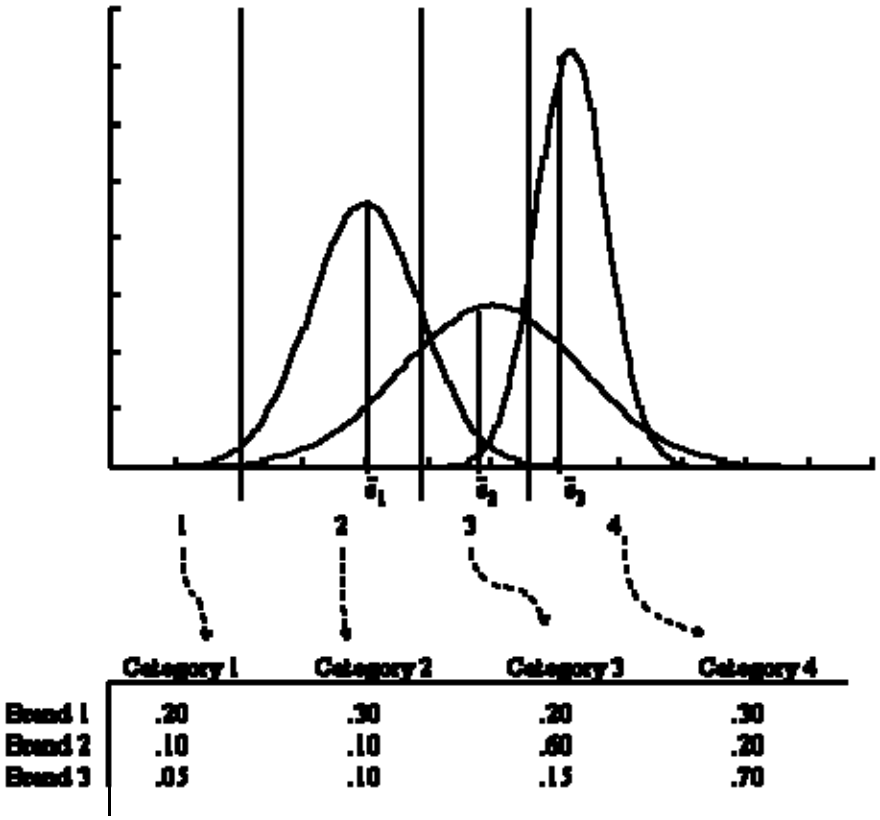
Note that for all of our estimation schemes; unweighted least squares, weighted least squares, modified minimum Chi Square, and Maximum Likelihood; we have q independent probabilities [t (t – 1) / 2] and 2 (t – 1) free parameters. The model therefore has q – 2 (t – 1) degrees of freedom.

12.5 *The Law of Categorical Judgment*

In addition to paired comparison data, Thurstone also contemplated absolute judgments, that is, when subjects assign ordered categories to objects without reference to other objects. For example, we might have a series of brands being rated on a scale as below,

Like it a lot – Like it a little bit – Not crazy about it – Hate it
[ ] [ ] [ ] [ ]

which is a simplified (and I hope marginally whimsical) version of the ubiquitous category rating scale used in thousands of marketing research projects a year. We assume that the psychological continuum is divided into four areas by three thresholds or cutoffs. In general, with a J point scale we would have J – 1 thresholds. We will begin with the probability that a subject uses category j for brand i. We can visualize our data as below:



The probabilities shown above represent the probability that a particular brand is rated with a particular category. However, we need to cumulate those probabilities from left to right in order to have data for our model. The *cumulated probabilities* would look like the table below.

Brand 1	.20	.50	.70	1.00
Brand 2	.10	.20	.80	1.00
Brand 3	.05	.15	.30	1.00

Define the jth cutoff as c_j. We set $c_0 = -\infty$ and $c_J = +\infty$. We can then estimate values for $c_1, c_2, \cdots, c_{J-1}$. These cumulated probabilities are worthy to be called the p_{ij} and they represent the probability that brand i is judged in category j or less, which is to say, to the left of cutoff j. Our model is that each brand has a perceptual impact on the subject given by

$$s_i = \bar{s}_i + e_i \text{ with}$$

$$e_i \sim N(0, \sigma^2) .$$

In that case

$$\hat{p}_{ij} = \Pr[s_i < c_j] = \Pr[c_j - s_i > 0] . \qquad (12.26)$$

But we have already seen a number of equations that look just like this! The probability that a normal random variable is greater than zero, Equation (12.8), has previously been used in the Law of Comparative Judgment. That probability, in this case of absolute judgment, is given by

$$\hat{p}_{ij} = \Phi\left[c_j - \bar{s}_i / \sigma_i\right]. \tag{12.27}$$

The importance of how to model categorical questionnaire items should be emphasized here. Such items are often used in factor analysis and structural equation models (Chapters 9, 10 and 11) under the assumption that the observed categorical ratings are normal. On the face of it, that would seem highly unlikely given that one of the assumptions of the normal distribution is that the variable is continuous and runs from $-\infty$ to $+\infty$! In the Law of Categorical Judgment, however, the variables s_i behave exactly that way. What's more, one can actually calculate the correlation between two Thurstone variables using what is known as a *polychoric correlation* and model those rather than Pearson type correlations.

12.6 *The Theory of Signal Detectability*

The final model to be covered in this chapter is another Thurstone-like model, but one invented long after Thurstone's 1927 paper. In World War II, Navy scientists began to study sonar signals, and more germane to marketing, they began to study the technician's response to sonar signals. Much later, models for human signal detection came to be applied to consumers trying to detect real ads that they had seen before, interspersed with distractor ads never shown to those consumers.

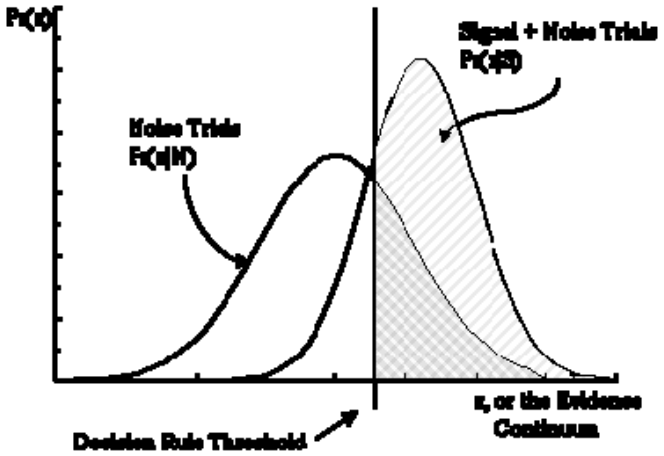
The theory of signal detectability (TSD) starts with the idea that a detection task has two distinct components. First, there is the actual sensory discrimination process, the resolving power if you will, of the human memory or the human senses being put to the test. This is related to our physiology, our sensitivity as receivers of the signal in question, and the signal-to-noise ratio. Second, there is a response decision involved. This is not so much a sensory issue as a cognitive one. It is related to bias, expectation, payoffs and losses, and motivation. For example, if you think you hear a submarine and it turns out you are wrong, the Captain may make you peel a crate of potatoes down in the mess hall. However, if you don't think that the sound you heard was a submarine and it turns out to have been one, you and the Captain will both find yourselves in Davy Jones' Locker, if you don't mind the nautical allusion. Given a particular ability to actually detect the sign of a sub, you might be biased towards making the first error and not making the second one. The TSD is designed to separate this response bias from your actual ability to detect subs.

Returning to our group of consumers being asked about ads they have seen, there are a number of ways to collect data. Assume that they have seen a set of ads. You are now showing them a series of ads which include the ads that they have seen along with some new ones that were never shown. Obviously, not including distractor ads is a little bit like giving a True/False test with no false items. You can ask them to say Yes or No; I saw that ad or I didn't. This is known as the *Yes/No Procedure*. You can also ask them on a ratings scale that might run from "Very Sure I Have Not Seen This Ad" on the left to "Very Sure I Have Seen This Ad" on the right. This is known as the *Ratings Procedure*. Finally, you can give them a sheet of paper with one previously exposed ad on it, and n - 1 other ads never before seen. Their task would be to pick the remembered ad from among the n alternatives, a procedure known as *n-alternative forced choice*, or n-afc for short. These procedures, and TSD, can be used for various sorts of judgments: Same/Different, Detect/No Detect, Old/New, and so forth. At this point, we will begin discussing the Yes/No task. The target ad that the consumer has seen will be called the signal, while the

distractor ads will be called the noise. We can summarize consumer response in the following table:

		Response	
		S	N
Reality	S	Hit	Miss
	N	False Alarm	Correct Rejection

Here the probability of a Hit plus the probability of a Miss sum to 1, as do the False Alarm and Correct Rejection rates. The consequence of a Yes/No trial is a value on the evidence continuum. The Subject must decide from which distribution it arose: the noise distribution or the distribution that includes the signal. We can picture the evidence distribution below.



The x axis is the consumer's readout of the evidence to the consumer that the current trial contains an ad that they did indeed see. However, for whatever reason, due to the similarity between some target and some distractor ads, or other factors that could affect the consumer's memory, some of the distractor ads also invoke a relatively high degree of familiarity. The subject's task is difficult if the two distributions overlap, as they do in the figure. The difference in the means of the two distributions is called d′. The area to the right of the threshold for the Signal + Noise distribution, represented by lines angling from the lower left to the upper right, gives you the probability of a Hit, that is the Hite rate or HR. The area to the right of the Noise distribution gives you the False Alarm rate, or FAR. In the Figure, this is indicated by the double cross-hatched area. For noise trials we have

$$x_n = \bar{x} + e$$

and for signal + noise trials

$$x_s = \bar{x} + e + d'$$

where the parameter d′ represents the difference between the two distributions. We will assume that

$$e \sim N(0, \sigma^2)$$

and we fix $\overline{x} = 0$ and $\sigma^2 = 1$. Define the cutoff as c. Then

$$HR = Pr\,[Yes \mid Signal] = Pr(x_s > c) \tag{12.28}$$

$$= Pr\,(x_s - c > 0).$$

From Theorem (4.4) we can show that

$$E(x_s - c) = d' - c$$

and from Equation (4.8) that

$$V(x_s - c) = \sigma^2 = 1$$

so that

$$HR = \Phi(d' - c) \tag{12.29}$$

from Equation (12.8). As far as noise trials go,

$$FAR = Pr\,[Yes \mid Noise] = Pr(x_n > c) \tag{12.30}$$

$$= Pr\,[x_n - c > 0]\ .$$

Since

$$E(x_n - c) = -c$$

and

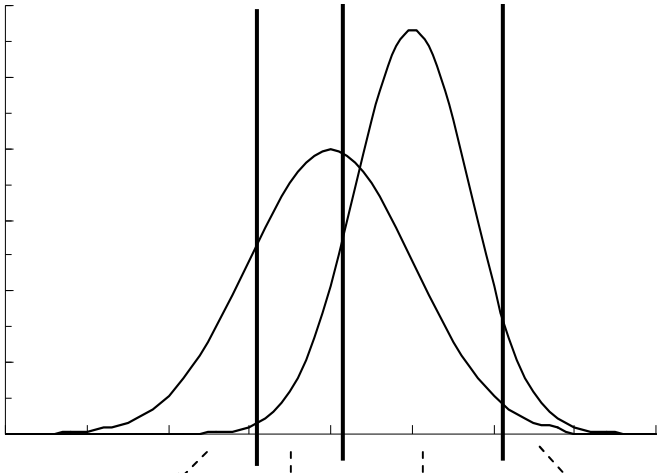
$$V(x_n - c) = \sigma^2 = 1,$$

we deduce that the

$$FAR = \Phi\,(-c)\ . \tag{12.31}$$

We can therefore transform our two independent data points, the HR and the FAR, into two TSD parameters, d′ and c. We can not test the model since we have as many parameters as independent data points. In order to improve upon this situation, we now turn to the Ratings procedure.

With ratings, we use confidence judgments to supplement the simple Yes/No decision of the consumer. The picture of what is going on under the ratings approach appears below;

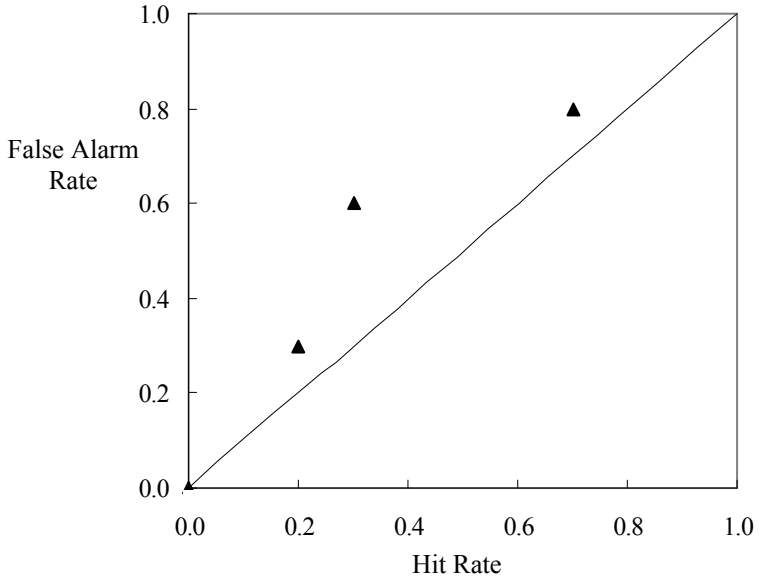


Signal	.30	.30	.20	.20
Noise	.20	.10	.40	.30

Just as we did in Section 12.5 with the Law of Categorical Judgment, we cumulate this table, which results in a stimulus-response table looking like

Signal	.30	.60	.80	1.00
Noise	.20	.30	.70	1.00

Each of the J-1 cutoffs, the c_j, defines a Hit Rate (HR_j) and a False Alarm Rate (FAR_j). Plotting them yields what is known as a Receiver Operating Characteristic, or ROC. Our pretend example is plotted below:



The shape of the ROC curve reveals the shape of the distributions of the signal and the noise. If we used z score coordinates instead of probabilities, the ROC should appear as a straight line. This suggests that we could fit the ROC using unweighted least squares. We will follow up on that idea shortly, but for now, let us review the model. For the Hit Rate for cutoff j we have

$$HR_j = \Pr[x_s - c_j > 0]$$

$$= \Phi\,[(d' - c_j) / \sigma_s] \tag{12.32}$$

while for noise trials we have

$$FAR_j = \Pr[x_n - c_j > 0]$$

$$= \Phi\,(- c_j) \tag{12.33}$$

Now we have $2\cdot(J - 1)$ probabilities with only $J + 1$ parameters: $d', \sigma_s^2, c_1, c_2, \cdots, c_{J-1}$. Of course, we could use weighted least squares or maximum likelihood. Or we could plot the ROC using z scores and fit a line. In that case, the equation of the line would be

$$\hat{Z}_{HR} = d' + \frac{\sigma_n}{\sigma_s} Z_{FAR}$$

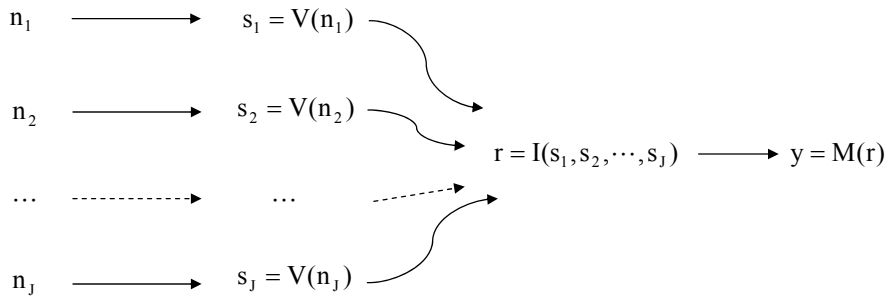
$$= d' + \frac{1}{\sigma_s} Z_{FAR}\ .$$

We close this chapter with just a word about the n-afc procedure. You can run this technique either sequentially or simultaneously. In either case, the consumer is instructed to pick exactly

one out of the n alternatives presented. There are no criteria or cutoffs in play in this procedure. According to the TSD, the percentage correct can be predicted from the area under the ROC curve.

12.7 *Functional Measurement*

We wrap up this chapter with a quick overview of what is known as functional measurement. We started the chapter talking about the relationship between the physical world and the mental impressions of that world. To round out the picture, after the sense impressions are transformed into internal stimuli, those stimuli may be combined, manipulated, evaluated, elaborated or integrated by the consumer into some sort of covert response. Then, this covert response is transformed into an observable behavior and voila, we have data to look at! A diagram will facilitate the explanation of the process:



On the left, the inputs are transformed into mental events, we can call them discriminal values by the function V(·). In the case of physical input, V(·) is a psychophysical function. In the case of abstract input, we can think of V(·) as a valuation function. Then, the psychological or subjective values are integrated by some psychological process, call it I(·), to produce a psychological response. This might be a reaction to an expensive vacation package that goes to a desired location, or a sense of familiarity evoked by an ad. Finally, the psychomotor function M(·) transforms the mind's response into some overt act. This could be the action of putting an item in the shopping cart, or checking off a certain box of a certain questionnaire item. With the help of conjoint measurement, certain experimental outcomes allow all three functions to be ascertained.

References

Christoffersson, Anders (1975) Factor Analysis of Dichotomized Variables. *Psychometrika*. 40(1), 5-32.

Psychophysics

Batchelor, R.A. (1986) "The Psychophysics of Inflation," *Journal of Economic Psychology,* 7 (September), 269-90.

Monroe, Kent B. (1977) "Objective and Subjective Contextual Influences on Price Perception." In Arch G. Woodside, Jagdish Sheth, and Peter D. Bennet (Eds.) *Consumer and Industrial Buying Behavior*. NY: Elsevier-North Holland.

Sinn, Hans-Werner (1985) "Psychophysical Laws in Risk Theory," *Journal of Economic Psychology,* 6, 185-206.

Kamen, Joseph M. and Robert J. Toman (1970) "Psychophysics of Prices," *Journal of Marketing Research,* 7 (February), 27-35.

Gabor, Andre, Clive W. J. Granger and Anthony P. Sowter (1971) "Comments on 'Psychophysics of Prices', *Journal of Marketing Research,* 8 (May), 251-2.

Comparative Judgement

Thurstone, L.L. (1927) "A Law of Comparative Judgement," *Psychological Review*, 38, 368-89.

Freiden, Jon B. and Douglas S. Bible (1982) "The Home Purchase Process: Measurement of Evaluative Criteria through Pairwise Measures," *Journal of the Academy of Marketing Science,* 10 (Fall, 359-76

Signal Detection and Categorical Judgement

Cradit, J. Dennis., Armen Taschian and Charles Hofacker (1994) "Signal Detection Theory and Single Observation Designs: Methods and Indices for Advertising Recognition Testing," *Journal of Marketing Research,* (February), 117-27.

Parducci, Allen (1963) "Category Judgement: A Range-Frequency Model," *Psychological Review*, 72 (6), 407-18.

Tashchian, Armen J., J. Dennis White, and Sukgoo Pak (1988) "Signal Detection Analysis and Advertising Recognition: An Introduction Measurement and Interpretation Issues," *Journal of Marketing Research* 25 (November), 397-404.

Srinivasan, V. and Amiya K. Basu (1989) "The Metric Quality of Ordered Categorical Data," M*arketing Science*, 8 (Summer), 205-30.

Functional Measurement

Anderson, Norman H. (1982) "Cognitive Algebra and Social Psychophysics" in Bernd Wegener (Ed.) *Social Attitudes and Psychophysical Measurement,* Hillsdale, NJ: Lawrence Erlbaum.

Levin, Irwin P., Richard D. Johnson and Patricia J.Deldin (1985) "Framing Effects in Judgement Tasks with Varying Amounts of Information," *Organization Behavior and Human Decision Processes,* 36 (December), 362-77.

Levin, Irwin P. (1985) "How Changes in Price and Salary Affect Economic Satisfaction: Information Integration Models and Inference Processes," *Journal of Economic Psychology*, 6 (June), 143-55.

Johnson, Richard D. and Irwin P. Levin (1985) "More than Meets the Eye: The Effect of Missing Information of Purchase Evaluations," *Journal of Consumer Research,* 12 (September), 169-77.

White J. Dennis and Elise L. Truly (1989) "Price-Quality Integration and Warranty Evaluation: A Preliminary Test of Alternative Models f or Risk Assessment," *Journal of Business Research*, 19, 109-25.

Birnbaum, Michael H. (1982) "Controversies in Psychological Measurement," in Bernd Wegener (Ed.) *Social Attitudes and Psychophysical Measurement,* Hillsdale, NJ: Lawrence Erlbaum.

Lynch, John G., Jr. (1985) "Uniqueness Issues in the Decompositional Modeling of Multiattribute Overall Evaluations: An Information Integration Perspective," *Journal of Marketing Research*, 22 (February), 1-19

Chapter 13: Random Utility Models

Prerequisites: Sections 12.1 - 12.4

13.1 *Some Terminology and a Simple Example*

The subject of this chapter is a type of model known as a Random Utility Model, or RUM. RUMs are very widely applied marketing models, especially to the sales of frequently purchased consumer packaged goods; in other words; the kind of stuff you see in a supermarket. All of the models in this chapter logically follow from Thurstone's Law of Comparative Judgment that we covered in Chapter 12. However, in this chapter we will consider the situation in which consumers pick one brand from a set of more than two brands, and we will also contemplate distributions other than the normal. We can summarize the assumptions of Thurstone's Law, and of the models in this chapter, as follows:

Assumption one is that choice is a *discrete event*. What this means is that choice is all-or-nothing. The consumer, as a rule, cannot leave the supermarket with .3432 cans of Coke and .6568 cans of Pepsi. They will tend to leave with 1 can of their chosen brand, and 0 cans of their not chosen brand. Thus choice is not a continuous dependent variable.

Assumption two is that the attraction or utility towards a brand varies across individuals as a random variable. In Thurstone's Law, we called this the discriminal dispersion and we assumed it was normal. By using the term utility, we are being consistent with economic theory. We also fequently use the term *attraction*, we are being consistent with the retailing literature. In any case, assumption two is all about the word "random" in the label random utility model.

The last assumption is that the consumer chooses the brand with the highest utility. This makes our consumer an economically rational being. Thank goodness.

In general, we will be concentrating on the class of RUMs known as *logit models*. These are models that make a distributional assumption different than the normal and lead to much simpler calculations. In the next sections we will be introduced to the logit model in all its glory. But before that happens, here is a list of other important terms that will come into play:

Dichotomous dependent variable – any dependent variable capable of taking on exactly two discrete values.

Polytomous dependent variable – any dependent variable capable of taking on exactly $J > 2$ discrete values.

Income type independent variable – a variable that varies over consumers. A logit model incorporating only income type variables is sometimes be called a *polytomous logit model*.

Price type independent variable – a variable that varies over consumers and brands. A logit model with at least one of these is often called a *conditional logit model*. We might note here however, that there is no difference in the way we treat price and income variables if we are looking at a dichotomous dependent variable. The difference only comes into play when there are three or more possible choices.

Aggregate data – data that have been summarized for each unique combination of the independent variables. To keep things simple, let us say we have just one independent variable; coupon value; and that there are exactly four different values. For each coupon value, we might count up how

many people buy the product (that is, use the coupon) and how many do not. The choice probabilities for each of the four coupon values constitute the data analyzed as the dependent variable. We would obviously have four data points, each point being two numbers: the choice probability and the value of the coupon. We can estimate this sort of data using either Generalized Least Squares or Maximum Likelihood.

Disaggregate data – raw data consisting of individual choices. It is possible that each observation has a unique combination of values on the independent variables. Maybe there are hundreds of different coupon values and hundreds of different possible prices. Each data point might come from a single individual, with a one signifying that that person bought the product, and a zero signifying that that person did not buy. Disaggregate data can only be analyzed by ML.

We are going to start with a simple example involving retail choice. In the small southern city of Rome, Alabama, there is a hypothetical food store that carries hard to find Italian items. A sample of individuals was asked, "Do you shop at the Negozio?" We define the dependent variable such that

$$y_i = \begin{cases} 1 \text{ if Yes} \\ 0 \text{ if No} \end{cases} \tag{13.1}$$

for person i. We can also define x_i as the distance between person i's residence and the Negozio. Of course, we will need to also define e_i as a random, independent error. We could use the linear model of Chapter 5 to fit this model. In that case we would have

$$y_i = \beta_0 + x_i\beta_1 + e_i \tag{13.2}$$

$$\hat{y}_i \equiv E(y_i) = \beta_0 + x_i\beta_1 . \tag{13.3}$$

Now we are going to define the probability that individual i chooses (has chosen) the store and the complement of this probability. For the former, we will use the notation p_{i1} and for the latter p_{i2}. Given this notation, we can say that the predicted choice probabilities are

$$\hat{p}_{i1} = \Pr[y_i = 1] = \Pr[\text{Yes}] \quad \text{and} \tag{13.4}$$

$$\hat{p}_{i2} = \Pr[y_i = 0] = \Pr[\text{No}] . \tag{13.5}$$

It should be clear that $\hat{p}_{i2} = 1 - \hat{p}_{i1}$. It must also be the case, given the definition of what we mean by expectation that

$$E(y_i) = (1)\hat{p}_{i1} + (0)\hat{p}_{i2}$$

$$= \hat{p}_{i1} .$$

Combining this result with Equation (13.3), we conclude that

$$\hat{p}_{i1} = \beta_0 + x_i\beta_1 . \tag{13.6}$$

There are two problems with this conclusion. First, a choice probability, really; any probability; has to obey the rule

$$0 \le \hat{p}_{i1} \le 1 \tag{13.7}$$

but there is no requirement that ordinary least squares estimation will produce predicted values between 0 and 1. In other words, OLS may produce *logically inconsistent* choice probabilities. A second important feature of probabilities is that

$$\hat{p}_{i1} + \hat{p}_{i2} = 1 \tag{13.8}$$

but again, we are not guaranteed that regression will produce complementary probabilities that add up to 1. In other words, the predicted values are not *sum constrained*. There is also a third problem. With OLS regression we make the Guass-Markov assumption [Equation (5.16)] in order to perform hypothesis testing. Specifically, in regression we generally assume $e_i \sim N(0, \sigma_i^2)$ with $\sigma_i^2 = \sigma^2$ for all i, that is; $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. But in the model we are now examining, there are exactly two possible values for e_i –

$$e_i = \begin{cases} 1 - (\beta_0 + x_i\beta_1) & \text{for } y_i = 1 \\ 0 - (\beta_0 + x_i\beta_1) & \text{for } y_i = 0. \end{cases} \tag{13.9}$$

By the definition of variance [see Equation (4.7)], we have

$$V(e_i) = E[e_i - E(e_i)]^2 \tag{13.10}$$

but since $E(e_i) = 0$, the expression above simplifies to $V(e_i) = E(e_i^2)$. And combining Equations (13.9) and (13.10) we see that

$$E(e_i^2) = \hat{p}_{i1}(1 - \beta_0 - x_i\beta_1)^2 + \hat{p}_{i2}(-\beta_0 - x_i\beta_1)^2 .$$

Note that since e_i is discrete, we use Equation (4.2) for its expectation. Combining the equation above with Equation (13.6) implies

$$\begin{aligned} E(e_i^2) &= \hat{p}_{i1}(1 - \hat{p}_{i1})^2 + (1 - \hat{p}_{i1})(-\hat{p}_{i1})^2 \\ &= \hat{p}_{i1}(1 - \hat{p}_{i1}) \\ &= (\beta_0 + x_i\beta_1)(1 - \beta_0 - x_i\beta_1) . \end{aligned} \tag{13.11}$$

But now we have a problem. The formula for the variance of the error has the independent variable on the right hand side. What's more, that independent variable has the subscript i hanging off it. How can the variance of e_i be the same for all i when it depends on x_i? It cannot – we have heteroskedasticity of error variance. Our OLS parameter estimates might be unbiased and consistent, but they are not efficient. Standard errors and significance tests do not hold. Although by definition, OLS produces the smalleset sum of squared error that can be, we have now uncovered three problems with using it for choice data: logical inconsistency, the lack of the sum constraint, and heteroskedasticity. Some simply find it inelegant to use a procedure capable of predicting a probability of less than zero or more than one.

There are a number of ways to fix these problems. You could at least take care of the logical inconsistency by using the linear probability model. This model simply forces $\hat{y}_i$ to 0 and 1 whenever it shows up outside the range:

$$\hat{p}_{i1} = \begin{cases} 0 & \text{if } \beta_0 + x_i\beta_1 \le 0 \\ \beta_0 + x_i\beta_1 & \text{if } 0 < \beta_0 + x_i\beta_1 < 1 \\ 1 & \text{if } \beta_0 + x_i\beta_1 \ge 1. \end{cases}$$

A second more theoretically grounded model is the *Probit model*. The probit model uses the same assumptions of the Thurstone model as presented in Chapter 12 namely that the utility of each of the choice options is normally distributed. In that case, we have

$$\begin{aligned} \hat{p}_{i1} &= F_p(\beta_0 + x_i\beta_1) = \Phi(\beta_0 + x_i\beta_1) \\ &= \int_{-\infty}^{\beta_0 + x_i\beta_1} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z^2}{2}\right] dz . \end{aligned} \tag{13.12}$$

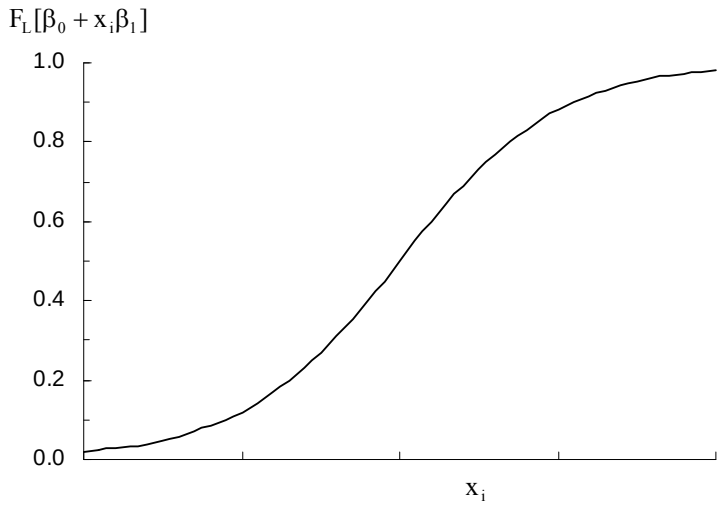
We could linearize the model by applying the PROBability Inverse Transform, or PROBIT transform (i.e. Φ^{-1}) and see the meaning of the name of this technique, as well as use unweighted least squares on the resulting z scores

$$\Phi^{-1}\left[\hat{p}_{i1}\right] = \hat{z}_{i1} = \beta_0 + x_i\beta_1 .$$

Unweighted least squares would not solve the third problem, namely heteroskedasticity. There are a variety of other estimation schemes for probit regression that would deal with this problem, but now we turn our attention to a very widely used model for choice data, the logit model. Note that in Equation (13.12) the appearance of the function F_p. Another version of this F function might be based on a transformation other than the normal or probit. This is illustrated below:

$$\hat{p}_{i1} = F_L(\beta_0 + x_i\beta_1) = \frac{e^{\beta_0 + x_i\beta_1}}{1 + e^{\beta_0 + x_i\beta_1}} \quad . \tag{13.13}$$

F_L is called the logistic function and so the model is sometimes called *logistic regression*. A visual representation is given below:



The logistic function is highly similar to the normal ogive. There are some important differences between it and the normal when there are more than two choice objects, but we will get to that topic later. For now, you might note that you can interpret the sign of the β in much the same way that you can in ordinary regression. A positive β implies that the choice probability goes up as x goes up. When dealing with this function, we can make the notation cleaner by defining $u_i = \beta_0 + x_i\beta_1$ so that

$$\hat{p}_{i1} = \frac{e^{u_i}}{1+e^{u_i}}.$$

Now, multiply both sides by e^{-u_i} / e^{-u_i}

$$\hat{p}_{i1} = \frac{e^{u_i}}{1+e^{u_i}} \cdot \frac{e^{-u_i}}{e^{-u_i}}$$

$$= \frac{1}{1+e^{-u_i}}$$

which shows another way to write the model. In general, we will use the previous version,

$$\hat{p}_{i1} = \frac{e^{u_i}}{1+e^{u_i}}. \qquad (13.14)$$

As such, lets look at the probability that the respondent does not go to the store. That is

$$\hat{p}_{i2} = 1 - \hat{p}_{i1} = \frac{1+e^{u_i}}{1+e^{u_i}} - \frac{e^{u_i}}{1+e^{u_i}}$$

$$= \frac{1}{1+e^{u_i}}. \qquad (13.15)$$

Now look at the expression for $\hat{p}_{i1}$ in Equation (13.14) and for $\hat{p}_{i2}$ in Equation (13.15). In effect we have a/(a+b) for the one and b/(a+b) for the other, with 1 and e^{u_i} playing the roles of a and b. The logistic model is a special case of Bell, Keeney and Little's (1975) *Market Share Theorem* and what Kotler (1984) once called the *Fundamental Theorem of Market Share*. We can make this theorem more general by using the following notation:

$$\hat{p}_{ij} = \frac{a_{ij}}{\sum_{m}^{J} a_{im}}.$$

In our case, there are J = 2 brands, $a_{i1} = e^{u_i}$ and $a_{i2} = e^0 = 1$.

The logit model is not a linear model but it can be linearized. Repeating the model,

$$\hat{p}_{i1} = \frac{e^{u_i}}{1+e^{u_i}}$$

and multiplying both sides by $1/\hat{p}_{i2}$, the reciprocal of Equation (13.15), yields

$$\hat{p}_{i1}/\hat{p}_{i2} = \frac{e^{u_i}}{1+e^{u_i}} \cdot \frac{1+e^{u_i}}{1}$$

$$= e^{u_i}.$$

Now, we can take logs to get

$$\ln(\hat{p}_{i1}/\hat{p}_{i2}) = u_i = \beta_0 + x_i\beta_1.$$

The left hand side is called a *logit*. You can transform your choice probabilities into logits and fit a linear model using unweighted least squares. This, at least, solves both the logical consistency issue and the lack of sum constraint when OLS regression is applied to raw probabilities. It does not deal with the issue of efficiency, however. For that we will need to contemplate weighted least squares or maximum likelihood.

13.2 *Aggregate Data*

Imagine that we have a table of data, a table of different groups really. Our table might look like the one below, which shows N populations and the frequency of Yes's and No's within each population:

	Response		
Population	Yes ($y_i = 1$)	No ($y_i = 0$)	**x**
1	f_{11}	f_{12}	x_1
2	f_{21}	f_{22}	x_2
…	…	…	…
i	f_{i1}	f_{i2}	x_i
…	…	…	…
N	f_{N1}	f_{N2}	x_N

In the table, f_{i1} is the frequency with which members of group i say Yes, or simply put, the number of people living at distance x_i from the Negozio who go to that store. In what follows, it will be useful to define

$$n_i = f_{i1} + f_{i2}$$

$$p_{i1} = f_{i1} / n_i$$

$$\hat{\ell}_{i,12} = \ln(\hat{p}_{i1}/\hat{p}_{i2})$$

and analogously,

$$\ell_{i,12} = \ln(p_{i1}/p_{i2}) .$$

Of course, the distinction between $\ell_{i,12}$ and $\hat{\ell}_{i,12}$ is important. The first one is the observed logit and the second one is the logit as predicted by the model. Even when the model holds in the population studied, sampling error will see to it that they are not identical. To make an analogy to regression, we can say

$$\ell_{i,12} = \beta_0 + x_i\beta_1 + (\ell_{i,12} - \hat{\ell}_{i,12}) \ .$$

Without proof, let me claim that

$$E(\ell_{i,12}) = \hat{\ell}_{i,12} \tag{13.16}$$

and that

$$\begin{aligned} V(\ell_{i,12}) &= V(\ell_{i,12} - \hat{\ell}_{i,12}) \\ &= \frac{1}{n_i \hat{p}_{i1} \hat{p}_{i2}} . \end{aligned} \tag{13.17}$$

In summary, what this means is that in our model,

$$\ell_{i,12} = \beta_0 + x_i\beta_1 + (\ell_{12/i} - \hat{\ell}_{i,12}),$$

the error term in parentheses has

$$E(\ell_{i,12} - \hat{\ell}_{i,12}) = 0 \text{ and}$$

$$V(\ell_{i,12} - \hat{\ell}_{i,12}) = \frac{1}{n_i \hat{p}_{i1} \hat{p}_{i2}}.$$

What's more, it can be shown [this is related to but not the same as Equation (6.2)] as $n_i \to \infty$

$$\ell_{i,12} - \hat{\ell}_{i,12} \sim N\left[0,\ 1/n_i \hat{p}_{i1} \hat{p}_{i2}\right] \tag{13.18}$$

and in fact the approximation to the normal is already quite close by the time $n_i \geq 30$.

13.3 *Weighted Least Squares and Aggregate Data*

Under ordinary circumstances, $E(y_i - \beta_0 + x_i\beta)$ has the constant variance σ^2, and we minimize, as in Equation (5.21),

$$SS_{Error} = \sum_i (y_i - \beta_0 - x_i\beta_1)^2.$$

The residual in our case, that is the term in parentheses above, has variance $1/n_i \hat{p}_{i1} \hat{p}_{i2}$ which is clearly not a constant, since the subscript i appears in the term. In three places! We can, however, use this knowledge to stabilize the variance. We will create a set of weights consisting of the reciprocal of the variance of each observation. Specifically, we define

$$w_i = n_i \hat{p}_{i1} \hat{p}_{i2}$$

as the weights that we will use in the weighted least square (WLS) formula SS_{Error} formula below

$$SS_{Error} = \sum_i w_i (y_i - \beta_0 - x_i\beta_1)^2. \tag{13.19}$$

Here we might note that the weights serve to emphasize or de-emphasize the influence of a particular observation depending on its sampling variance. The higher the variance, the less influence the observation has in the determination of the SS_{Error}.

At this time we are going to shift into matrix notation so as to come up with a more general expression for WLS. Lets say that we have one independent variable, as before, consisting of travel distance to our shop in Rome, AL. Call that variable $\mathbf{x}_{\cdot 1}$. A second independent variable might be the family income of each respondent, $\mathbf{x}_{\cdot 2}$. Then

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \cdots & \cdots & \cdots \\ 1 & x_{i1} & x_{i2} \\ \cdots & \cdots & \cdots \\ 1 & x_{N1} & x_{N2} \end{bmatrix} = \begin{bmatrix} \mathbf{x}'_{1\cdot} \\ \mathbf{x}'_{2\cdot} \\ \cdots \\ \mathbf{x}'_{i\cdot} \\ \cdots \\ \mathbf{x}'_{N\cdot} \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

Note that the **X** matrix has N rows, with an upper case N used to maintain a distinction between the number of populations, and n_i, the number of observations within each population i. Now we can write our model as

$$\begin{aligned} \hat{p}_{i1} &= \frac{e^{\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2}}{1 + e^{\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2}} \\ &= \frac{e^{\mathbf{x}'_{i\cdot}\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'_{i\cdot}\boldsymbol{\beta}}} \end{aligned} \tag{13.21}$$

or using the logit expression,

$$\hat{\ell}_{i,12} = \ln \frac{\hat{p}_{i1}}{\hat{p}_{i2}} = \mathbf{x}'_{i\cdot}\boldsymbol{\beta}\,. \tag{13.22}$$

This second way of expressing the model is convenient for estimation using the linear model. To do so, we begin by stacking the predicted and observed logits from each of the N populations into the vectors

$$\hat{\boldsymbol{\ell}}' = \begin{bmatrix} \hat{\ell}_{1,12} & \hat{\ell}_{2,12} & \cdots & \hat{\ell}_{N,12} \end{bmatrix}$$

$$\boldsymbol{\ell}' = \begin{bmatrix} \ell_{1,12} & \ell_{2,12} & \cdots & \ell_{N,12} \end{bmatrix}.$$

The model is then

$$\hat{\boldsymbol{\ell}} = \mathbf{X}\boldsymbol{\beta} \tag{13.23}$$

Now, we take the variances for each term $\ell_{i,12} - \hat{\ell}_{i,12}$ and place them into the covariance matrix **V** as diagonal elements:

$$\mathbf{V} = \begin{bmatrix} 1/n_1\hat{p}_{11}\hat{p}_{12} & 0 & \cdots & 0 \\ 0 & 1/n_2\hat{p}_{21}\hat{p}_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1/n_N\hat{p}_{N1}\hat{p}_{N2} \end{bmatrix}.$$

Also note that we can relate the elements of this matrix to the previous scalar notation in Equation (13.19) because

$$\mathbf{V}^{-1} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_N \end{bmatrix}.$$

In matrix terms, our objective function is

$$\begin{aligned} f &= (\boldsymbol{\ell} - \hat{\boldsymbol{\ell}})'\mathbf{V}^{-1}(\boldsymbol{\ell} - \hat{\boldsymbol{\ell}}) \\ &= (\boldsymbol{\ell} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\boldsymbol{\ell} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{13.24}$$

The f function, at its minimum, is distributed as χ^2 when the model holds in the population. Thus, it serves as a test of the null hypothesis that the model is correct. This is basically the same approach we used in Equation (12.19), with Minimum Pearson χ^2. If we were to replace the $\hat{p}$'s in the **V** matrix with p's, we would have Modified Minimum χ^2. When we set $\partial f/\partial\boldsymbol{\beta} = \mathbf{0}$ we find

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\ell}$$

as the GLS parameter estimates. Since $V(\boldsymbol{\ell}) = V(\boldsymbol{\ell} - \hat{\boldsymbol{\ell}}) = \mathbf{V}$ we further find that

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= \left\{[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\right\}\mathbf{V}\left\{[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}\right\}' \\ &= [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]. \end{aligned}$$

Following the same line of reasoning that we used in Section 6.8 (and also Section 17.4), we can use the above matrix for confidence intervals or to test hypotheses of the form

$$H_0: \beta_j = 0$$

or more generally

$$H_0: \mathbf{a}'\boldsymbol{\beta} - c = 0$$

and create the usual *t*-statistic with the denominator being formed by the scalar

$$\mathbf{a}'\,(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{a}, \text{ i.e.}$$

$$\hat{t} = \frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - \mathrm{c}}{\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{a}}}.$$

For multiple degree of freedom hypotheses of the form

$$\mathrm{H}_0\text{: } \mathbf{A}\boldsymbol{\beta} - \mathbf{c} = \mathbf{0}$$

we use

$$\mathrm{SS}_{\mathrm{H}} = (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{A}]^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}),$$

and for error,

$$\mathrm{SS}_{\mathrm{Error}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathrm{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

13.4 *Maximum Likelihood and Disaggregate Data*

With disaggregate data, we have household level observations. For the time being we return to the relatively simple case of a single independent variable, our distance measure from household i to the store. It is quite possible that each household has a unique value on this variable, especially if it is measured as a continuous variable. In addition, for each household we have a 1 if that household goes to the store and we have a 0 otherwise. Modifying our sample size notation once again, lets say we have N households altogether, with N_1 of them having said "Yes" and being scored with a '1' on the dependent variable, and N_2 of them having said "No." The model for the choice probability stays the same as before, we have just returned to the situation of a single independent variable for now,

$$\hat{\mathrm{p}}_{\mathrm{i1}} = \frac{\mathrm{e}^{\beta_0 + \mathrm{x}_{\mathrm{i1}}\beta_1}}{1 + \mathrm{e}^{\beta_0 + \mathrm{x}_{\mathrm{i1}}\beta_1}}$$

$$= \frac{\mathrm{e}^{\mathbf{x}_{\mathrm{i}\cdot}'\boldsymbol{\beta}}}{1 + \mathrm{e}^{\mathbf{x}_{\mathrm{i}\cdot}'\boldsymbol{\beta}}},$$

but our objective function will be quite different. To begin creating the objective function, we might consider sorting the data into two piles: in the first pile we place the N_1 households saying "Yes" and in the second, the remainder who have said "No." We note that under the model, the probability of observing our N_1 Yes's and the rest of the sample with its No's, is

$$\ell_0 = \prod_{\mathrm{i}=1}^{\mathrm{N}_1} \hat{\mathrm{p}}_{\mathrm{i1i}} \prod_{\mathrm{i}=\mathrm{N}_1+1}^{\mathrm{N}} \hat{\mathrm{p}}_{\mathrm{i2}} \qquad (13.25)$$

assuming that each observation is independent of all the others. This notation emphasizes the fact that there are two piles of observations: the first which consists of households going to the store, and the second consisting of those who do not frequent the place. Another way to write the likelihood is perhaps more clever, and relies on the fact that we have decided to score $y_i = 1$ if household i buys from the store and $y_i = 0$ if it does not. Rewriting ℓ_0 we have

$$\ell_0 = \prod_{i=1}^{N_1} (\hat{p}_{i1})^{y_i} \, (\hat{p}_{i2})^{1-y_i} \tag{13.26}$$

which takes advantage of the fact that for any value a, $a^1 = a$ while $a^0 = 1$. This second form avoids having to sort the observations. However, returning to Equation (13.25), we can flesh out the predicted choice probabilities. When we do that, the likelihood is seen as

$$\ell_0 = \prod_{i=1}^{N_1} \left(\frac{e^{\beta_0 + x_{i1}\beta_1}}{1 + e^{\beta_0 + x_{i1}\beta_1}} \right) \prod_{i=N_1+1}^{N} \left(\frac{1}{1 + e^{\beta_0 + x_{i1}\beta_1}} \right). \tag{13.27}$$

Since the maximum of the likelihood occurs at the same place as the maximum of the log likelihood, we will take logs and get

$$\begin{aligned} L_0 &= \ln(\ell_0) \\ &= \sum_{i=1}^{N_1} (\beta_0 + x_i\beta_1) - \sum_{i=1}^{N} \ln\left(1 + e^{\beta_0 + x_i\beta_1}\right). \end{aligned} \tag{13.28}$$

To go from the previous expression for ℓ_0 in Equation (13.27) to the second line of Equation (13.28) for L_0 immediately above requires that you notice the denominator is identical for both $\hat{p}_{i1}$ and $\hat{p}_{i2}$, and that ln(1) = 0. That explains why the first summation in Equation (13.28) goes to N_1 and the second goes all the way to N. We now wish to set

$$\frac{\partial L_0}{\partial \beta_0} = \frac{\partial L_0}{\partial \beta_1} = 0$$

and solve for $\hat{\beta}_0$ and $\hat{\beta}_1$ using nonlinear optimization as is discussed in Section 3.9. To that end, the second order derivatives are quite useful. These provide additional information about the search direction. But what's more, they can be used to figure out the covariances and variances of the ML parameter estimates, which allows us to do hypothesis testing. For example, lets start with $\partial L_0/\partial \beta_0$, and think of it as a function of the value of β_1. How does $\partial L_0/\partial \beta_0$ change as β_1 changes? The limit of the slope of $\partial L_0/\partial \beta_0$ (treated as a "dependent variable" in the calculus sense) on β_0 (treated as the "independent variable") is the second order derivative and it may be written

$$\frac{\partial}{\partial \beta_0}\left(\frac{\partial L_0}{\partial \beta_1}\right) \equiv \frac{\partial^2 L_0}{\partial \beta_0 \partial \beta_1} = \frac{\partial^2 L_0}{\partial \beta_1 \partial \beta_0}.$$

Think of this as element h_{12} and h_{21} in the symmetric **H** matrix, called *the Hessian*. Element 1, 1 is

$$h_{11} = \frac{\partial}{\partial \beta_0}\left(\frac{\partial L_0}{\partial \beta_0}\right) \equiv \frac{\partial^2 L_0}{\partial \beta_0 \partial \beta_0} = \frac{\partial L_0}{\partial \beta_0 \partial \beta_0}$$

and of course element 2, 2 would be defined analogously. Minus the expectation of the Hessian is called the *Information Matrix*, i. e. –E(**H**). Finally, the inverse of the information matrix gives us the variance-covariance matrix of the unknowns, which is to say

$$V(\hat{\boldsymbol{\beta}}) = [-E(\mathbf{H})]^{-1}.$$

We can now test hypotheses using this matrix to provide the denominator of the *t*-statistic. Note that the Hessian is square and symmetric, and it will have one row (and one column) for each unknown parameter.

A final observation, before we start thinking about what happens if we have three choice options as opposed to only two, is that we can create an R^2 like statistic by comparing the log likelihood of the model, with the log likelihood of a model that consists only of β_0, that is, it has no real independent variables. This is illustrated below:

$$\rho^2 = 1 - \frac{L_0}{L_0^*}$$

where L_0^* is the likelihood under the model with just an intercept.

13.5 *Three or More Choice Options*

The situation with three or more brands, or three or more store choices, or Web links, etc., is rather more complicated than the two option case. Of course, we can say that

$$p_{i1} + p_{i2} + p_{i3} = 1$$

so at least we know something about the situation. However, there are now three potential logits: $\ln(p_{i1}/p_{i3})$, $\ln(p_{i2}/p_{i3})$ and $\ln(p_{i1}/p_{i2})$. But

$$\ln\frac{p_{i1}}{p_{i2}} = \ln\frac{p_{i1}}{p_{i3}} - \ln\frac{p_{i2}}{p_{i3}}$$

so one logit is redundant in the same sense that one of the three choice probabilities is not independent of the other two: if you know two of the probabilities you can figure out the third by subtracting the total of the other two from 1. With J brands, we will create J – 1 generalized logits. It is traditional to use the last brand, often a store or generic brand, as the denominator. The full model, called the *Multinomal Logit model* or MNL model is given below:

$$\hat{p}_{ij} = \frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}_j}}{\sum_m^J e^{\mathbf{x}'_{mi}\boldsymbol{\beta}_m}} \tag{13.29}$$

where p_{ij} is the choice probability for brand j (j = 1, 2, ···, J) for case i. In this context i could either index populations, as would be the case with aggregate data, or individuals, which would be

the case with disaggregate data. The vector $\mathbf{x}'_{ij}$ provides values for the independent variables for brand j, observation i, while the vector $\boldsymbol{\beta}_j$ contains the unknown parameters for each independent variable for brand j. We can express the model as a special case of the Fundamental Theorem of Market Share,

$$\hat{p}_{ij} = \frac{a_{ij}}{\sum_{m}^{J} a_{im}} \quad \text{with}$$

$$a_{ij} = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}_j).$$

By tradition, we set the attraction for the last brand, brand J, equal to 1, i. e. $a_{iJ} = 1$ for all i, and thus $\boldsymbol{\beta}'_J = [0 \quad 0 \quad \cdots \quad 0]$. For ML estimation we pick elements of the β_r vectors to maximize

$$\ell_0 = \prod_{i=1}^{N_1} \hat{p}_{i1} \prod_{i=N_1+1}^{N_2} \hat{p}_{i2} \cdots \prod_{i=N_{J-1}+1}^{N} \hat{p}_{iJ}$$

where we have sorted the cases into J piles corresponding to the choice option picked by that individual.

13.6 *A Transportation Example of the MNL Model*

The following example is inspired by, but not identical to, an actual dataset reported in Currim (1985), who considered the choice faced by household i between getting to work by car (1), bus (2), or using the metro (3). Our explanatory variables are

I_i	Income of household i
C_i^j	Cost (price) of alternative j for household i
CAV_i	Cars per driver for household i
BTR_i	Bus transfers required for member of household i to get to work via the bus

One possible model for this situation might be

$$\ln \frac{\hat{p}_{i1}}{\hat{p}_{i3}} = \alpha_1 + I_i\beta_1 + (C_i^1 - C_i^3)\beta_3 + CAV_i\beta_4$$

$$\ln \frac{\hat{p}_{i2}}{\hat{p}_{i3}} = \alpha_2 + I_i\beta_2 + (C_i^2 - C_i^3)\beta_3 + BTR_i\beta_5.$$

Now we will have an opportunity to put into play some of the terminology we looked at in the beginning of the chapter but have not used up to now. Lets look at the role of income in this model. Income is constant across the choices that a family can make, but in the two logits, income has a different coefficient (β_1 and β_2). The quality of the choice option might vary, and so income may well contribute to families preferring choice 1 over choice 3, but it may lead to families preferring choice 3 over choice 2.

The price variable, C_i^j, varies across choice options as well as households. For one particular family, a car trip may be \$4.00 (including depreciation), a bus trip might be \$1.00 and a trip on the Metro could be \$1.50. But while C_i^j varies, the coefficient β_3 is constant. Such a variable is sometimes called *generic*. McFadden calls this sort of structure the *conditional logit* model. It is also known as the *simple effects model*.

The variables CAV_i and BTR_i only apply to one alternative. Thus they are called *Alternative Specific Variables* (ASVs). The α_j are also alternative specific variables. To be specific, they are alternative specific constants (ASCs). You might imagine a MNL model with only alternative specific constants . This would be quite similar to a Thurstone model, such as the Comparative Judgment model discussed in Section 12.3, only in this case we have a distribution other than the normal. In fact, in current context the ASCs function as a sort of error term. They represent the attraction towards the brand that exists independently of any measured assets such as its price, etc.

For GLS estimation it makes sense to create a single linear equation for the logits. That equation would look like this:

$$\begin{bmatrix} \hat{\ell}_{1,13} \\ \hat{\ell}_{2,13} \\ \cdots \\ \hat{\ell}_{N,13} \\ \hat{\ell}_{1,23} \\ \hat{\ell}_{2,23} \\ \cdots \\ \hat{\ell}_{N,23} \end{bmatrix} = \begin{bmatrix} 1 & 0 & I_1 & 0 & C_1^1 - C_1^3 & CAV_1 & 0 \\ 1 & 0 & I_2 & 0 & C_2^1 - C_2^3 & CAV_2 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & I_N & 0 & C_N^1 - C_N^3 & CAV_N & 0 \\ 0 & 1 & 0 & I_1 & C_1^2 - C_1^3 & 0 & BTR_1 \\ 0 & 1 & 0 & I_2 & C_2^2 - C_2^3 & 0 & BTR_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & I_N & C_N^2 - C_N^3 & 0 & BTR_N \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}$$

$$\hat{\boldsymbol{\ell}} = \mathbf{X}\boldsymbol{\beta}\,. \tag{13.30}$$

On the other hand, the best way to represent the model if we were going to do ML estimation is to show it as the nonlinear equation for the choice probabilities as

$$\hat{p}_{ij} = \frac{e^{\mathbf{x}_{ij}'\boldsymbol{\beta}_j}}{\sum_m^J e^{\mathbf{x}_{im}'\boldsymbol{\beta}_j}}\,. \tag{13.31}$$

13.7 *Other Choice Models*

There are a variety of related alternative forms for choice models, but for each model discussed in this section, and more generally in this chapter with the clear exception of the probit model of Section 13.10, we will be assuming the Fundamental Theorem,

$$\hat{p}_{ij} = \frac{a_{ij}}{\sum_m^J a_{im}}\,.$$

We will be assuming that we have k marketing instruments, meaning that we have a set of marketing variables perhaps including price, advertising effort, distribution effort, or some product attributes. The conditional or simple effects MNL is

$$a_{ij} = \exp(\sum_{k} x_{ijk}\beta_{k}) . \tag{13.32}$$

This model assumes that each marketing instrument has its own β coefficient, but each brand's marketing efforts have the same result for marketing instrument k. In other words, there is a β coefficient for each marketing instrument (price, place, etc.), but these are constant across the J brands.

Another type of simple effects model has been championed by Cooper and Nakanishi (1988). It is called the *Multiplicative Competitive Interaction* (MCI) model and looks like

$$a_{ij} = \prod_{k} x_{ijk}^{\beta_{k}} . \tag{13.33}$$

The MCI model follows in the footsteps of the economic Cobb-Douglas function of Equation (16.3), often used for demand equations for continuous dependent variables.

We can also have *differential effects* models in which the impact of each brands varies. Perhaps one of the brands is better than some of the others at leveraging its marketing efforts so it receives more benefit per dollar spent on advertising, to use that instrument as an example. There is a version of the differential effects model for the MNL,

$$a_{ij} = \exp(\sum_{k} x_{ijk}\beta_{jk}) \tag{13.34}$$

and for the MCI:

$$a_{ij} = \prod_{k} x_{ijk}^{\beta_{jk}} . \tag{13.35}$$

You will note that the beta coefficients have a subscript for the brand in the above two models. Finally, there is the *full extended model*. In the case of the MNL, this has been called the *universal logit model*:

$$a_{ij} = \exp(\sum_{m}^{J}\sum_{k} x_{imk}\beta_{mjk}) . \tag{13.36}$$

There is also a fully extended MCI model,

$$a_{ij} = \prod_{m}\prod_{k} x_{imk}^{\beta_{mjk}} . \tag{13.37}$$

These include asymmetric cross effects of one brand on another.

13.8 *Elasticities and the MNL Model*

How does our share change when we change the value of a marketing instrument? Lets assume that we have a model with only one marketing instrument; price. In line with Section 16.1, we define the *price elasticity of market share* for brand j as

$$e_{ij} = \frac{\partial \hat{p}_{ij}}{\partial x_{ij}} \cdot \frac{x_{ij}}{\hat{p}_{ij}}$$

for observation i. According to the generic or simple-effects model,

$$a_{ij} = \exp\left[\alpha_j + (x_{ij} - x_{iJ})\beta\right] \text{ or}$$

$$a_{ij} = \exp(\alpha_j + x_{ij}\beta).$$

In order to figure out the elasticity, we must start with the derivative,

$$\frac{\partial \hat{p}_{ij}}{\partial x_{ij}} = \frac{\partial}{\partial x_{ij}}\left[a_{ij}\left(\sum_m a_{im}\right)^{-1}\right].$$

In addition to the power rule and chain rule of the calculus (see Section 3.3), we need to note that

$$de^a/da = e^a$$

and

$$\frac{de^{f(x)}}{dx} = e^{f(x)} \cdot f'(x).$$

In that case

$$\frac{d\hat{p}_{ij}}{dx_{ij}} = -a_{ij}\left(\sum_m a_{im}\right)^{-2} a_{ij}\beta + a_{ij}\beta\left(\sum_m a_{im}\right)^{-1}$$

$$= -\hat{p}_{ij}^2\beta + \hat{p}_{ij}\beta$$

$$= \beta\hat{p}_{ij}(1 - \hat{p}_{ij}),$$

so that the elasticity is then

$$e_{ij} = \beta x_{ij}(1 - \hat{p}_{ij}). \qquad (13.38)$$

Since x_{ij} appears in the expression for the elasticity, the elasticity is not constant and instead changes along the price-share curve. The elasticity is also inversely proportional to the share

which makes sense – the higher your share already, the harder it is to drive it closer to one by dropping prices even more.

For the simple effects MCI where $a_{ij} = x_{ij}^{\beta}$, we have

$$e_{ij} = \beta(1 - \hat{P}_{ij}).$$

In contrast to the MNL model, the MCI model produces constant elasticities much like the Cobb-Douglas function does for continuous dependent variables.

Marketing scientists are often interested in the cross elasticity for brand j with respect to some other brand j′. This quantity summarizes the extent to which the share of j depends on the prices set by the brand management of j′. This reveals the nature and amount of competition among the brands in the choice set. By definition, the price *cross elasticity of share* for brand j with respect to brand j′ is

$$e_{i,jj'} = \frac{\partial \hat{p}_{ij}}{\partial x_{ij'}} \cdot \frac{x_{ij'}}{\hat{p}_{ij}}. \tag{13.39}$$

The derivative for the simple effects MNL is

$$\frac{\partial \hat{p}_{ij}}{\partial x_{ij'}} = -\hat{p}_{ij}\hat{p}_{ij'}\beta$$

which makes the cross elasticity

$$e_{i,jj'} = -\hat{p}_{ij'}x_{ij'}\beta. \tag{13.40}$$

Since no j subscript appears on the right hand side, only j′, brand j′ has the same impact on all other brands. This impact is proportional to the share of j. For the differential effects model, i. e. $a_{ij} = \exp(\alpha_j + x_{ij}\beta_j)$,

$$e_{i,jj'} = -\hat{p}_{ij'}x_{ij'}\beta_{j'},$$

each brand exerts a different pressure, but that pressure is the same on all the other brands.

That brand j′ should exert the same pressure on all brands flies in the face of common sense. We often think that some brands compete more with certain other brands and less with others. This common sense notion is part of what is known as *Independence of Irrelevant Alternatives*.

13.9 *Independence of Irrelevant Alternatives*

Independence of Irrelevant Alternatives, or *IIA* as it is lovingly known, refers to the tendency of the Fundamental Theorem to model competition in a very symmetric way. We will now discuss the issue of asymmetric competition. Imagine that the transportation needs of a certain city are served by two companies: The Blue Bus Company and the Yellow Cab Company. Imagine further that these two companies split the market in half with each getting a market share of 50%.

What would happen if a new competitor arrives, namely, the Red Bus Company. The Fundamental Theorem would have us believe that the new market shares will be 1/3rd each. Does this seem realistic to you?

The universal logit model can handle asymmetric competition. Technically speaking, however, it is actually not a RUM! The only other model in this chapter to be able to deal with the problem of IIA is presented next.

13.10 *The Polytomous Probit Model*

Again we will be concerned with the market share of brand j out of J different brands. The utility of each brand is normally distributed over the consumers in the market. Each individual picks the utility that is largest. We will define our model as

$$\mathbf{y} = \mathbf{B}\mathbf{x} + \boldsymbol{\varepsilon}$$

where **y** is the J by 1 random vector of utilities described above, **B** is J by k and **x** is k by 1. This model can include income or price type variables in **x**. Their presence determines the appearance of **B** which, like in covariance structure models discussed in Chapter 10, or the ML MNL models discussed earlier in this chapter, can have zeroes in various positions. The random input vector can be characterized by noting that

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) .$$

The share for brand j is

$$\hat{p}_j = \Pr[y_j > y_{j'}]$$

$$= \Pr[y_j - y_{j'} > 0] \quad \text{for all } j' \neq j.$$

Now define $v_{j'}^{(j)} = y_j - y_{j'}$. Now we simply rewrite the expression for the share of brand j as

$$\hat{p}_j = \Pr[v_{j'}^{(j)} > 0] \quad \text{for all } j' \neq j.$$

For the next step, we will place all of the $v_{j'}^{(j)}$ for each brand $j' \neq j$ into the vector $\mathbf{v}^{(j)}$. The action of subtracting all of the other brands from brand j is obviously a linear operation. We will illustrate this operation using brand 1 as our example. Define the J - 1 by J matrix

$$\mathbf{M}^{(1)} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix}$$

for brand j = 1. As we can see, this **M** matrix differences all of the other rows from the first row of any postmultiplying matrix. So in particular,

$$\mathbf{v}^{(1)} = \mathbf{M}^{(1)}\mathbf{y}$$

and in general for brand j

$$\mathbf{v}^{(j)} = \mathbf{M}^{(j)}\mathbf{y} .$$

Of course, Theorem (4.5) and Theorem (4.9) show us that

$$E[\mathbf{v}^{(j)}] \equiv \hat{\mathbf{v}}^{(j)} = \mathbf{M}^{(j)}\hat{\mathbf{y}} = \mathbf{M}^{(j)}\mathbf{B}\mathbf{x} \text{ and}$$

$$V[\mathbf{v}^{(j)}] \equiv \mathbf{\Sigma}^{(j)} = \mathbf{M}^{(j)}\mathbf{\Sigma}\mathbf{M}^{(j)'} .$$

Therefore according to the multivariate normal distribution, presented in Equation (4.17), the share for brand j is

$$\hat{P}_j = \frac{1}{(2\pi)^{N/2}|\mathbf{\Sigma}^{(j)}|^{1/2}} \int_0^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} \exp\left[-(\mathbf{v}^{(j)} - \hat{\mathbf{v}}^{(j)})'\mathbf{\Sigma}^{(j)^{-1}}(\mathbf{v}^{(j)} - \hat{\mathbf{v}}^{(j)})/2\right] dv_{J-1}^{(j)}\, dv_{J-2}^{(j)} \cdots dv_1^{(j)} .$$

Share is equal to the probability that the utility for brand j exceeds the utility for all other brands, $j' \neq j$.

References

Cooper, Lee G and Masao Nakanishi (1988) *Market-Share Analysis*. Boston: Kluwer.

Kotler, Philip (1984) *Marketing Management Fifth Edition*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

McFadden, Daniel (1986) "The Choice Theory Approach to Market Research," *Marketing Science* 5 (Fall), 275-97.

Logit Models

Green, Paul E., Frank J. Carmone and David P. Wachspress (1977) "On the Analysis of Qualitative Data in Marketing Research," *Journal of Marketing Research*, 14 (February), 52-9.

Malhotra, Noresh K. (1984) "The Use of Linear Logit Models in Marketing Research," *Journal of Marketing Research,* 21 (February), 20-31.

Gensch, Dennis H. and Wilfred W. Recker (1979) "The Multinomial Multiattribute Logit Choice Model," *Journal of Marketing Research*, 16 (February), 124-32.

Flath, David and E.W. Leonard (1979) "A comparison of Two Logit Models in the Analysis of Qualitative Marketing Data." *Journal of Marketing Research* 16, (November), 533-8.

Bunche, David S. and Richard R. Batsell (1989) " A Monte Carlo Comparison of Estimators for the Multinomial Logit Model," *Journal of Marketing Research,* 26 (February), 56-68.

Probit Models

Malhotra, Naresh K. (1983) "A Threshold Model of Store Choice," *Journal of Retailing,* 59, (Summer), 3-21.

Hofacker, Charles F. (1990) "Derivation of Covariance Probit Elasticities," *Management Science,* 36 (April), 500-4.

Currim, Imran S., (1982) "Predictive Testing of Consumer Choice Models Not Subject to Independence of Irrelevant Alternatives, "*Journal of Marketing Research,* 21 (November), 20-31.

Kamakura, Wagner A. and Rajendra Srivastava (1984) "Predicting Choice Shares under Conditions of Brand Interdependence," *Journal of Marketing Research,* 21 (November), 20-31.

Nested Logit Models and Extensions of the MNL Mode

Louviere, Jordan J. and George Woodworth (1983) "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data," *Journal of Marketing Research*, 20 (November), 350-67.

Moore, William L. and Donald R. Lehman (1989) "A Paired Comparison Nested Logit Model of Individual Preference Structures," *Journal of Marketing Research¸26 (November), 420-8.*

Elasticities and Competition

Mahajan, Vijay, Paul E. Green and Stephen M. Goldberg (1982) " A Conjoint Model for Measuring Self-and Cross-Price/Demand Relationships," 19 (August), 334-42.

Gupta, Sunil (1988) "Impact of Sales Promotions on When, What and How Much to Buy," *Journal of Marketing Research*, 25 (November), 342-55.

Noncompensatory Choice

Manrai, Ajay K. and Prabhakant Sinha (1989) "Elimination by Cutoffs," *Marketing Science,* 8 (Spring), 133-52.

Rotondo, John (1986) "Price as an Aspect of Choice in EBA," *Marketing Science* , 5 (Fall), 391-402.

Johnson, Eric J. and Robert J. Meyer (1984) "Compensatory Choice Models of Noncompensatory Processes: The Effect of Varying Context," *Journal of Consumer Research* 11 (June), 528-41.

Grether, David and Louis Wilde (1984) "An Analysis of Conjunctive Choice: Theory and Experiments," *Journal of Consumer Research*, 10 (March) 373-85

Chapter 14: Nonmetric Scaling

Prerequisites: Chapter 7, Section 3.9

14.1 *Additive Conjoint Measurement*

In Chapters 5 through 7 we look at the classical statistical models from which we get our *t*-test, ANOVA, and of course, regression. For example, in a factorial design, we generally assume that our dependent variable is measured at the interval or ratio level, and we test to see if the cell means combine in an additive way, or if instead, we need to include interaction terms. Nonmetric additive conjoint measurement turns this reasoning exactly on its head. In this section we will assume the additivity in order to learn something about the level of measurement of the dependent variable. Or, we can make very weak assumptions about this measurement - i. e. that its merely ordinal - and still capture the main effects of the factorial design.

In this section we will use a linear model, and for the most part we will assume that we have a factorial design. For example, a consumer may be looking at a series of vacation packages. For now, lets just say that each package has five destinations and four prices leading to 20 different packages altogether. Our data might consist of a ranking (or possibly rating if there are not too many ties) of the 20 vacation packages. In any case, we will assume that our data are ordinal. Here are the steps that we will go through.

Step 0. Initialize a second version of the dependent variable,

$$\mathbf{y}^* = \mathbf{y}$$

where $\mathbf{y}$ is the original ordinal-scaled 20 by 1 column vector of the rankings for each of the 20 packages. As we will see, the $\mathbf{y}^*$ vector will be the optimally scaled version of the data.

Step 1. Use least squares to fit an additive model. Our usual notation would have us fit a model looking like

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \hat{\mathbf{y}}^* + \mathbf{e}$$

where $\mathbf{X}$ is a *design matrix* containing, lets say, effect coding (see Section 7.2) for main effects only and the $\boldsymbol{\beta}$ vector contains those effects. We can also use scalar notation that allows us to keep track of data from individual cells in the two way design. For example, looking at destination i and price j we might have

$$y^*_{ij} = \alpha_i + \beta_j + e_{ij} = \hat{y}^*_{ij} + e_{ij}$$

where α_i represents the effect of being in row i (destination i) and β_j captures the effect of being in column j of the design, that is the column with price j. Note that since the $\hat{y}^*_{ij}$ are optimally rescaled anyway we do not have to worry about the grand mean. We can just absorb that in the scaling. Step 1 is a least squares step where we pick the α_i and the β_j (or if you prefer the matrix notation, the elements of the $\boldsymbol{\beta}$ vector) so as to minimize the sum of squared error. But now we are going to go into a second least squares step.

Step 2. Find the monotone transformation that would improve the fit of the above model as much as possible. We will be focusing on this step in this section, but for now, we can say that in Step 1

we modified the parameters to fit the dependent variable, but in this step we are modifying the dependent variable to better fit the additive model. Symbolically we can say that the new values of the y_{ij}^{*} will be

$$y_{ij}^{*} = H_m[y_{ij}, \hat{y}_{ij}^{*}].$$

This says that we will be modifying the optimally scaled dependent values (y^{*}), based on the ordinal data (y) and the linear model of the optimally scaled dependent variable ($\hat{y}^{*}$). The function H_m is monotone, which means that the optimally rescaled dependent variables have to be in the same order as the original ordinal data. Ordinal means that only the order of the numbers is invariant, and the rescaled dependent variable honors that order, meaning that it is equivalent to the original.

Step 3. Refit the additive model.

Step 4. If the model fits OK, stop. Otherwise go back to Step 2 and repeat.

The procedure alternates between two least squares steps with one (Step 1 above) fitting the linear model minimizing the sum of squared error, and the other (Step 2 above) fitting the data minimizing a sum of squares called *STRESS*. We are thus alternating least squares steps, and this technique is part of a family of techniques that are called *Alternating Least Squares* or ALS. We will discuss STRESS in just a little bit. Now let us jump into Step 2 in more detail, a step that we call *optimal scaling*.

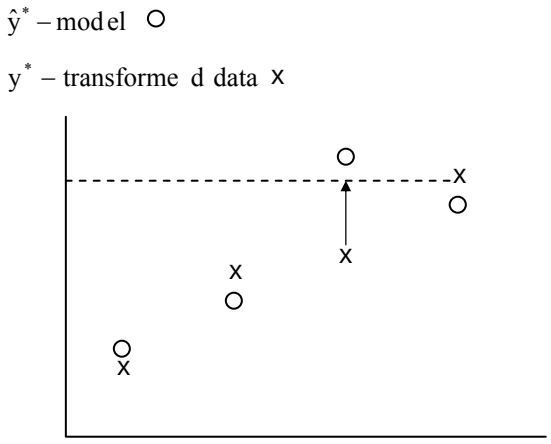
For the time being, to make our life easier, we will assume that there are no ties in the data. If we sort the data, and revert back to one subscript that refers to each datum's sort sequence, the original ordinal data would look like

$$y_1 < y_2 < \cdots < y_n . \tag{14.1}$$

In that case, to honor or maintain this order, we impose on the optimally scaled values the following constraints:

$$y_1^{*} \le y_2^{*} \le \cdots \le y_n^{*} . \tag{14.2}$$

We can picture the situation by looking at what is known as a *Shepard Diagram*, named after the Stanford psychologist Roger Shepard,



The x's represent the optimally rescaled data, and the 0's are the predictions of that data from the additive model. In particular, focus on the third data point. Due to the monotone restrictions, it cannot pass the fourth data point. It can come right up to its value and tie it however, since Equation (14.2) allows for equality. The program will move the y^* value as close to its predicted value as it can without violating the monotone constraints. In analytic terms, we minimize the following quantity subject to the inequalities above:

$$\text{STRESS} = \sqrt{\frac{\sum_i (y_i^* - \hat{y}_i^*)^2}{\sum_i (\hat{y}_i^* - \bar{\hat{y}}^*)^2}} \tag{14.3}$$

where $\bar{\hat{y}}^*$ is the average of the $\hat{y}_i^*$. In effect, the denominator normalizes the value of STRESS. It is the numerator that is where the action is. But remember, the formula is subject to the series of inequalities given in Equation (14.2).

There are two ways to handle ties in the data. The primary approach occurs when $y_i = y_j$ implies $y_i^* \leq y_j^*$. Here we are treating the data as fundamentally continuous with thresholds. The secondary approach is when when $y_i = y_j$ implies $y_i^* = y_j^*$. Here we treat the data as truly discrete, and we fully honor equalities.

The output from this procedure consists of the y_{ij}^*, called the *utilities*, and the α_i and the β_j, called *part-worths*. These can all be used to simulate various market conditions.

14.2 *Multidimensional Scaling*

Many of the models in this and other sections represent choice situations. Which brand do you like more and which do you like less? *Multidimensional Scaling*, often abbreviated MDS, is designed to get at the consumer's perception of the brands rather than their preferences for them. Later on we will bring preference back into the model, but for now we will focus on the way that the consumer sees the brands. We will also focus on nonmetric MDS, meaning that we will assume that the data are ordinal. Nonetheless, we will be able to fit a model with interval scaled parameters, just as we did in the section on conjoint measurement.

The MDS data collection procedure is one of the least obtrusive methods that exist in the world of marketing research. The respondent's job is to rank (or rate) pairs of brands as to how similar they are. We might imagine a simplified experimental design with three brands A, B and C. The respondent will tell us which of the three possible pairs, AB, AC, BC, are the most similar. Then which pair is the next most similar, and so forth until all of the pairs are ranked as to their similarity.

MDS uses a geometric model for *similarity* or *proximity judgments*. Brands judged highly similar, according to the model, are represented near each other in a *perceptual space*. Conversely, brands judged dissimilar find themselves distant in this perceptual space of r dimensions. Later we will get back to the dimensionality of the space. Now, lets think about the similarity judgment between brand i and brand j, and call it d_{ij}. The optimally rescaled data will be called d_{ij}^* while the predicted rescaled data, that is predicted from the distance model, will be called $\hat{d}_{ij}^*$. As before we will be using Alternating Least Squares. The perceptual space reveals the aspects of the brands that the consumer considers salient when looking at those brands. The steps in the algorithm are

Step 0. Initialize $d_{ij}^* = d_{ij}$.

Step 1. Fit the distance model $\hat{d}_{ij}^*$ to the d_{ij}^*.

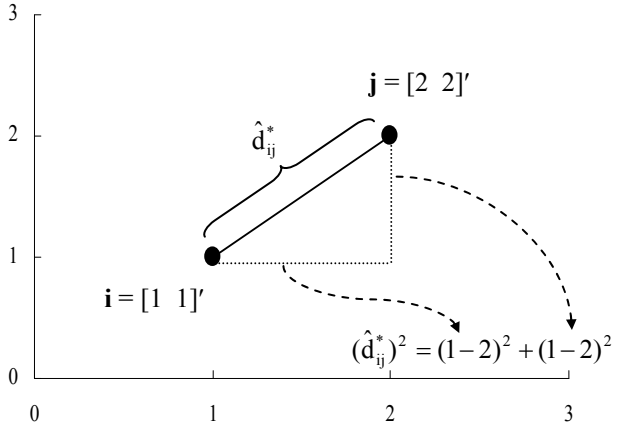
Step 2. Optimally rescale the d_{ij}^* to the $\hat{d}_{ij}^*$ honoring the order of the d_{ij}.

Step 3. Quit if done or go back to Step 1.

As before, in Step 2 we will be minimizing STRESS. However, at this point it would be wise to look at the distance model used in Step 1. We will be modeling the proximities as distances,

$$\hat{d}_{ij}^* = \sqrt{\sum_{m}^{r} (x_{im} - x_{jm})^2}. \tag{14.4}$$

Some of you may remember this equation from a high school geometry course. It is the *Euclidean distance* between points i and j in a space of r dimensions. The parameter x_{im} represents the coordinate for brand i on the m^{th} dimension. Assuming that r = 2, we might look at a graph of the situation:



In the case pictured, the distance between i and j is $\sqrt{2}$.

The flexibility of MDS can hardly be overstated. There are at least three categories of methods that allow us to capture similarity or proximity:

Direct

Ask for pairwise ratings or rankings
Have respondents sort objects into categories
Pick the pair of pairs most similar (Method of tetrads)
For each member of a trio, indicate which other brand it is most similar to (Method of triads)

Attribute Based

Calculate correlations over measures
Calculate distances over measures

Behavioral

Traffic volume, phone calls, trade or migration between two cities, regions, etc.
Switching proportions between brands
Confusability
Cross elasticities
Percent agreement, Chi Square, other measures of association

14.3 *Other Distance Models*

In addition to the classic Euclidean formula, other formulae qualify as distances, which are also called *metrics*. More accurately, we might use the word metric. Four axioms must be satisfied for a set of numbers to qualify as a metric:

Identity $\hat{d}^*_{ii} = 0,$ (14.5)

Non-negativity $\hat{d}^*_{ij} \geq 0,$ (14.6)

Symmetry $\hat{d}^*_{ij} = \hat{d}^*_{ji}$, and (14.7)

Triangle inequality $$\hat{d}^*_{ij} + \hat{d}^*_{jk} \geq \hat{d}^*_{ik} . \quad (14.8)$$

Of course the already noted classic Euclidean Distance equation,

$$\hat{d}^*_{ij} = \sqrt{\sum_m^r (x_{im} - x_{jm})^2},$$

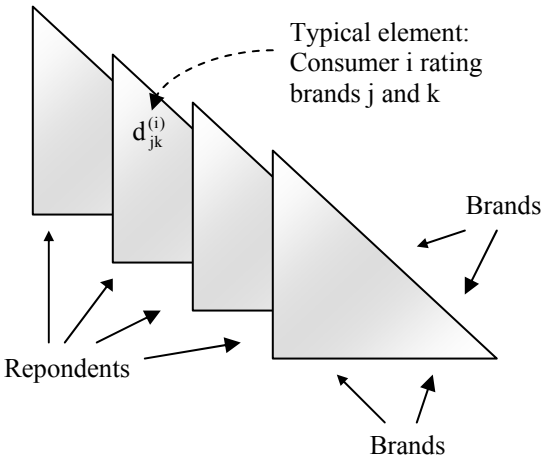
satisfies all four requirements. But other distance models are possible. For example, there is a very flexible formula called the *Generalized Minkowski Metric* in which

$$\hat{d}^*_{ij} = \left[\sum_m^r | x_{im} - x_{jm} |^a \right]^{1/a} . \quad (14.9)$$

When a = 2 you get the classic Euclidean formula. When a = 1 you get a metric known as the *City Block Metric*. This is the distance between two places where all angles have to be 90° and the triangle inequality holds as an equality. This metric is often used when the objects being scaled are perceptually decomposable. As $a \to \infty$ you get the *supremum metric* in which respondents only notice the biggest difference.

14.4 *Individual Differences in Perception*

Up to this point in this chapter, we have been looking at two way single mode data. What this means is that we have a data matrix with rows and columns and thus it is said to be *two way data*. There is just a single mode, however, since both ways of the matrix are indexed by brands. Now we will look at what happens when we have *three way data* representing *two modes*. The second mode will be individual subjects. A diagram for this sort of data is given below:



A typical element in the dataset would be $d^{(i)}_{jk}$, the similarity judgment for brands j and k made by subject i. If the numbers from each matrix are not comparable, as they would be if each subject was engaging in rank-ordering, the data are called *matrix conditional*.

Data such as these can be analyzed using the *Weighted Euclidean Model*, also known as the *INDSCAL model* (INDividual Differences SCALing). That model is given now:

$$\hat{d}_{jk}^{(i)*} = \left[\sum_{m}^{r} w_{im} (x_{jm} - x_{km})^2 \right]^{1/2}, \tag{14.10}$$

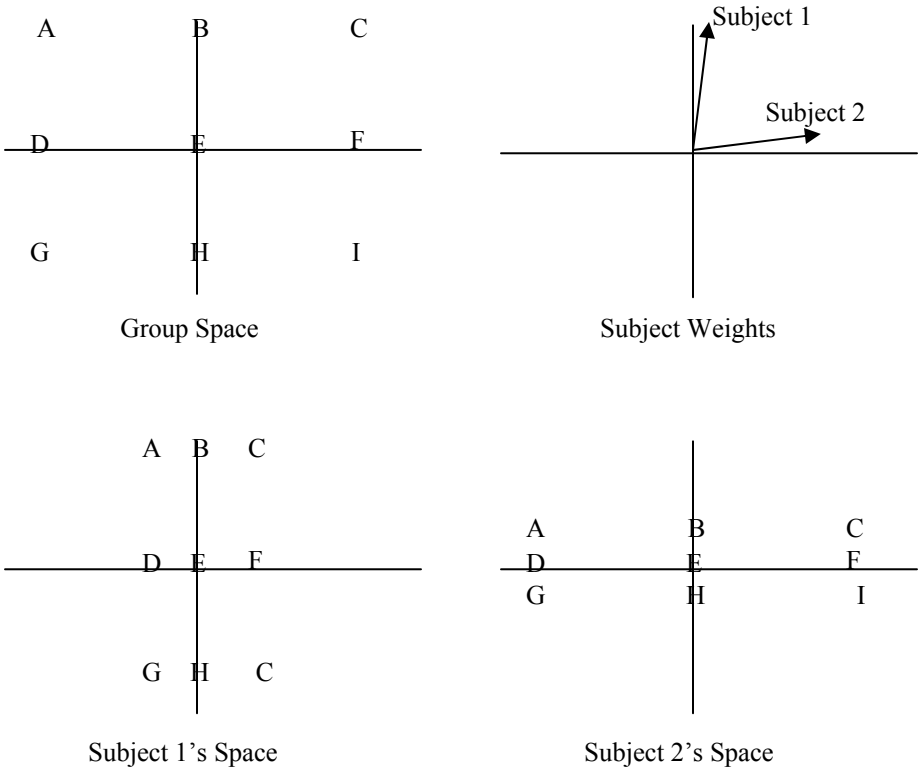
with the added element of the weights, the w_{im}, which represent the importance placed on dimension m by individual i. The coordinates, the x_{jk}, are still coordinates in this model but they are the coordinates of the brand in the *group space*. Each individual has their own coordinates which create a Euclidean space in which the axes have been stretched or shrunk. The coordinate for individual i would be

$$y_{jm}^{(i)} = x_{jm} \cdot w_{im}^{1/2} \tag{14.11}$$

so that

$$\hat{d}_{jk}^{(i)*} = \left[\sum_{m}^{r} (y_{jm}^{(i)} - y_{km}^{(i)})^2 \right]^{1/2} \tag{14.12}$$

using these "customized" coordinates. A diagram of how all this looks appears below:



Subject 1, who has a high weight for dimension 2 and a very small weight for dimension 1 has a space where brands that are separated on the second dimension are very dissimilar, but brands

whose only difference lies along dimension 1 (for example, brands A, D and G), seem quite similar to this person. Subject 2 has the opposite pattern.

The INDSCAL model can be conveniently represented in matrix notation. Place the coordinates for brand j in the vector

$$\mathbf{x}'_{j\cdot} = [\,x_{j1} \quad x_{j2} \quad \cdots \quad x_{jm} \quad \cdots \quad x_{jr}\,] .$$

Here, the dot subscript reduction operator on the symbol $\mathbf{x}'_{j\cdot}$ comes from Equation (1.2), and basically is used to hold the place of the second subscript, the one for the dimensions. We also put the subject weights on the diagonal of the matrix $\mathbf{W}^{(i)}$ as

$$\mathbf{W} = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_{ir} \end{bmatrix} .$$

Then the INDSCAL model is

$$\hat{d}_{jk}^{(i)*} = \left[\mathbf{x}'_{j\cdot}\mathbf{W}^{(i)}\mathbf{x}_{k\cdot}\right]^{1/2}$$

and the original, unweighted Euclidean model is a special case where $\mathbf{W}^{(i)} = \mathbf{I}$ for all i, or in other words, where

$$\hat{d}_{jk}^{*} = \left[\mathbf{x}'_{j\cdot}\mathbf{x}_{k\cdot}\right]^{1/2} .$$

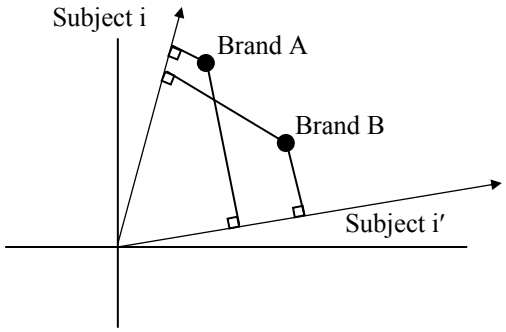
14.5 *Preference Models: The Vector Model*

In this model, we will be representing not just which brands are similar to which others, but which brands the consumers like the best. The vector model can be estimated from a variety of data types, but here we will assume that we have similarity judgments and preferences rankings or ratings.

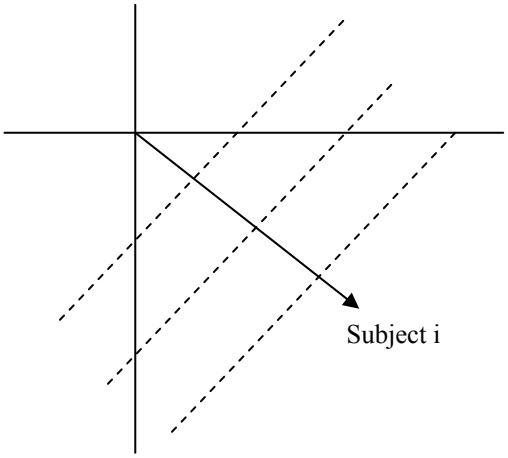
In the Vector model, the brands are represented as points in the perceptual space, as before. Each brand has a set of coordinates on the r dimensions in the perceptual space,

$$\mathbf{x}'_{j\cdot} = [\,x_{i1} \quad x_{i2} \quad \cdots \quad x_{im} \quad \cdots \quad x_{ir}\,] .$$

But now, each subject is also represented in the space, by a vector. The projection of the brand on that vector determines the preference for it. The situation is illustrated below:



Subject i prefers Brand A to Brand B, as the projection for A exceeds the value for B on that subjects *preference vector*. Subject i′, on the other hand, prefers B to A as that person's projections line up in the opposite order. Note that the projections of the brands onto each of the subject vectors occur at right angles to those subject vectors. It is instructive to look at *isopreference contours* for a particular subject. A subject will be indifferent between any two brands sitting on the same isopreference contour since both have equal appeal. These contours are graphed below:



According to the model, our Subject i would be completely indifferent between any brands that would appear on the same dashed line. However, the subject would prefer a brand on a line farther from the origin to one on a line closer in.

In order to express these ideas mathematically, we need to be able to identify each consumer's vector. It will be convenient to pick a point on the vector at a distance of 1 unit from the origin. Doing so, we then have

$$\mathbf{y}'_{i\cdot} = \begin{bmatrix} y_{i1} & y_{i2} & \cdots & y_{im} & \cdots & y_{ir} \end{bmatrix}$$

and since the distance from this point to the origin is 1, we have

$$\sqrt{\mathbf{y}'_{i\cdot}\mathbf{y}_{i\cdot}} = \sqrt{\sum_{m}^{r} y_{im}^2} = 1.$$

The preference of person i for brand j, which is the projection of the brand's point onto the subject's preference vector so as to create a 90° angle with that vector, is given by

$$\hat{s}_{ij} = \frac{\sum_{m}^{r} y_{im} x_{jm}}{\left[\sum_{m}^{r} y_{im}^2\right]^{1/2}}$$

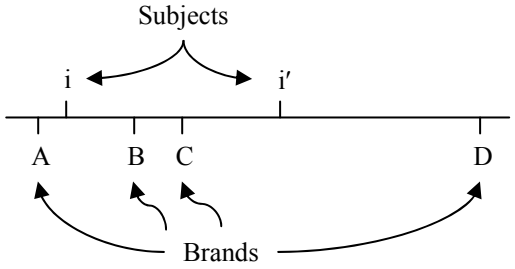
but since the denominator is 1, we have simply

$$\hat{s}_{ij} = \sum_{m}^{r} y_{im} x_{jm} = \mathbf{y}'_{i\cdot} \mathbf{x}_{j\cdot} . \tag{14.13}$$

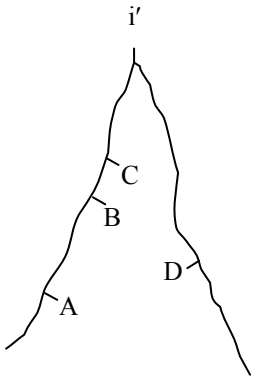
The vector model holds for many different product attributes. For example, for price, less is always better. For miles per gallon, more is preferred to less. But there are some product attributes for which the vector model makes ridiculous predictions. For example, it is quite possible that I would like a car that is larger than a sub-compact. But does this mean I would always want a larger and larger car? The vector model predicts that I would prefer a 2 mile long car to a sub-compact. This brings to mind the story of the porridge, which might be too hot, it might be too cold, or it might be perfect. To model perceptual attributes that act like this requires that we turn to the notion of an ideal point.

14.6 *Preference Models: The Ideal Point Model*

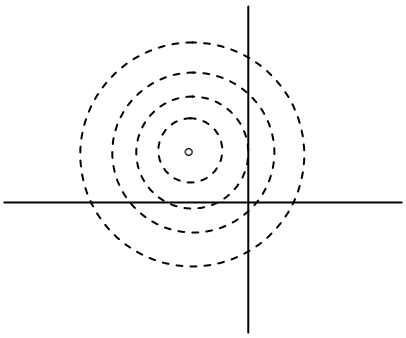
In the vector model, we represent individuals as a direction in the perceptual space. In the ideal point model, individuals, as well as brands, become points. In that sense we have a *joint space*. The situation is illustrated below with a one dimensional joint space.



We have picked two hypothetical respondents: i and i′. In the ideal point model, the closer a brand is to you, the more you like it. Thus i has the preference sequence: A-B-C-D while i′ prefers the brands in the following order: C-B-D-A. If the dimension were like a string, the preference of subject i′ could be determined by picking up that string at his or her ideal point:

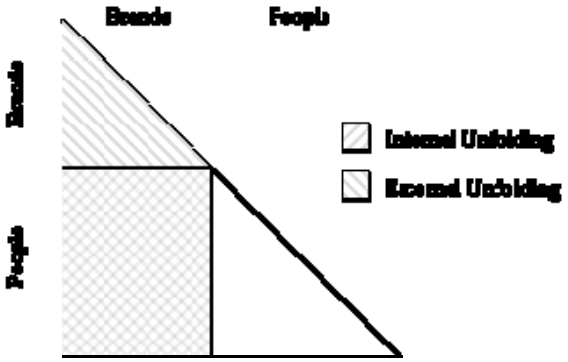


In fact, using this technique we actually perform the opposite mathematical operation. Given a set of consumers' preference rankings, we unfold the string, or more generally the r-dimensional space, to deduce the underlying position of the brands and the individuals' ideal points. In fact, this technique is sometimes known as *unfolding*. Looking at a two dimensional space, we can see the isopreference contours for the unfolding model:



Again, the subject whose ideal point is located at the center of those concentric circles, will be indifferent between any pair of brands appearing on the same circle. A brand on an inner circle will be preferred to a brand on a more outer circle.

There are two ways of collecting data for this model, and in fact for the vector model described above. You can collect internal data, which means that each individual rates or ranks their preference towards each brand, or you can combine that with perceptual ratings or rankings of the similarity of the brands. The situation is represented by the data matrix below.



Here we placed people and brands in the same lower triangular matrix. *Internal unfolding* utilizes just the people × brands rectangular part of this matrix, while *external unfolding* adds the brand × brand information.

The ideal point model is a distance model, so the formula for the preference of subject i for brand j is just the distance between subject i's ideal point and brand j's position in the joint space,

$$\begin{aligned}\hat{s}_{ij} &= \left[\sum_{m}^{r}(y_{im} - x_{jm})^2\right]^{1/2} \\ &= \left[(\mathbf{y}_{i\cdot} - \mathbf{x}_{j\cdot})'(\mathbf{y}_{i\cdot} - \mathbf{x}_{j\cdot})\right]^{1/2}.\end{aligned} \tag{14.14}$$

We can use metric or alternating least squares versions of this technique, and it is possible to have a version where there are individual differences in perception of the joint space, as we had with the INDSCAL model.

References

Conjoint Measurement

Green, Paul E. and V. Srinivasan (1978) "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research*, 5 (September), 103-21.

Green, Paul E. and V. Srinivasan (1990) "Conjoint Analysis in Marketing: New Developments with Implication for Research and Practice," *Journal of Marketing,* 54 (October), 3-19.

Jaccard, James, David Brinberg and Lee J. Ackerman (1986) "Assessing Attribute Importance: A Comparison of Six Methods," *Journal of Consumer Research,* 12 (March), 463-8.

Validity and Conjoint Measurement

Green, Paul E., Wayne S. DeSarbe and Pradeep K. Kedia (1980) "On the Insensitivity of Brand-Choice Simulations to Attribute Importance Weights," *Decision Sciences,* 11, 439-50.

Green, Paul E., Kristiaan Helsen and Bruce Shandler (1988) "Conjoint Internal Validity under Alternative Profile Presentations," *Journal of Consumer Research,* 15 (December), 392-7.

Green, Paul E., Abba M. Kreiger, and Pradep Bansal (1988) "Complete Unacceptable Levels in Conjoint Analysis: A Cautionary Note," *Journal of Marketing Research*, 25 (August), 293-300.

Hagerty, Michael R. (1986) "The Cost of Simplifying Preference Models," M*arketing Science,* 5 (Fall), 298-319.

Moore, William L. and Morris B. Holbrook (1990) "Conjoint Analysis on Object with Environmentally Correlated Attributes: The Questionable Importance of Representative Design," *Journal of Consumer Research* 16 (March), 490-7.

Reibstein, David, John E.G. Bateson and William Boulding (1988) "Conjoint Analysis Reliability: Empirical Findings," *Marketing Science*, 7 (Summer), 271-86.

Safizadeh, M. Hossein (1989) "The Internal Validity of the Trade-off Method of Conjoint Analysis," *Decision Sciences* 20 (Summer), 451-61.

Teas R. Kenneth (1985) "An Analysis of the Temporal Stability and Structural Reliability of Metric Conjoint Analysis Procedures," *Journal of the Academy of Marketing Science,* 13 (Winter/Spring), 122-42.

Wittink, Dick R. and Phillipe Cattin (1981) "Alternative Estimation Methods for Conjoint Analysis: A Monte Carlo Study," *Journal of Marketing Research,* 18 (February), 101-6

Conjoint: Choice Simulation and Aggregation

Green, Paul E. and Abba M. Krieger (1988) "Choice Rules and Sensitivity Analysis in Conjoint Simulators," *Journal of the Academy of Marketing Science*, 16 (Spring), 114-27.

Kamakura, Wagner A. (1988) "A Least Squares Procedure for Benefit Segmentation with Conjoint Experiments," *Journal of Marketing Research,* 25 (May), 157-67.

Wiley, James B. and James T. Low (1983) "A Monte Carlo Simulation of Two Approaches for Aggregation Conjoint Data,"*Journal of Marketing Research,* 20 (November), 405-16.

Conjoint: Extensions

Green Paul E. (1984) "Hybrid Models for Conjoint Analysis: An Expository Review," *Journal of Marketing Research,* 21 (May), 155-69.

Tantiwong, Duang Tip and Peter C. Wilton (1985) "Understanding Food Store Preferences among the Elderly Using Hybrid Conjoint Measurement Models," *Journal of Retailing,* 61 (Winter), 35-64.

Green Paul E., J. Douglas Carroll and Stephen M. Goldberg (1981) "A General Approach to Product Design Optimization Via Conjoint Analysis," *Journal of Marketing,* 45 (Summer), 17-37.

Green, Paul E. and Abba M. Krieger (1985) "Models and Heuristics for Product Line Selection," *Marketing Science* 4 (Winter), 1-19

Teas R. Kenneth (1987) "Magnitude Scaling of the Dependent Variable in Decompositional Multiattribute Preference Models," *Journal of the Academy of Marketing Science,* 15 (Fall), 64-73.

MDS in Marketing Research

Cooper, Lee G. (1983) "A Review of Multidimensional Scaling in Marketing Research," *Applied Psychological Measurement*, 7 (Fall), 427-50

Green, Paul E. and Frank J. Carmone (1969) "Multidimensional Scaling: An Introduction and comparison of Nonmetric Unfolding Techniques," *Journal of Marketing Research*, 6 (August), 330-41.

Lehmann, Donald R. (1972) "Judged Similarity and Brand-Switching Data as Similarity Measures," *Journal of Marketing Research,* 9 (August), 331-4.

Perreault, William D. and Forrest W. Young (1980) "Alternating Lease Squares Optimal Scaling: Analysis of Nonmetric Data in Marketing Research," *Journal of Marketing Research,* 17 (February), 1-13.

Moore, William L. and Morris B. Holbrook (1982) "On the Predictive Validity of Joint Space Models in Consumer Evaluations of New Concepts," *Journal of Consumer Research*, 9 (September), 206-10

MacKay, David B. and Joseph L. Zinnes (1986) "A Probabilistic Model for the Multidimensional Scaling of Proximity and Preference Data," *Marketing Science,* 5 (Fall), 324-44.

Schiffman, Susan S., M. Lance Reynolds and Forrest W. Young (1981) *Introduction to Multidimensional Scaling*. New York: Academic.

Shocker, Allen D. and V. Srinivasan (1974) "A Consumer Based Methodology for the Identification of New Product Ideas," *Management Science,* Series B, 20 (February), 921-37.

MDS Models: Applications and Extensions

DeSarbo Wayne S. and Vithala R. Rao (1986) "A Constrained Unfolding Methodology for Product Positioning," *Marketing Science,* 5 (Winter), 1-19.

Gavish, Bezale, Dan Horsky and Kizhanatham Srikanth (1983) "An Approach to the Optimal Positioning of a New Product," *Management Science*, 29 (November), 1277-97.

Green, Paul E. and Abba M. Krieger (1989) "Recent Contributions to Optimal Product Positioning and Buyer Segmentation," *European Journal of Operational Research*, 41 (July), 127-41.

Harshman, Richard A., Paul E. Green, Yorman Wind and Margaret E. Lundy (1982) "A Model for the Analysis of Asymmetric Data in Marketing Research," *Marketing Science*, 1 (Spring), 205-42.

Holbrook, Morris B. and Duglas V. Holloway (1984) "Marketing Strategy and the Structure of Aggregate, Segment-specific, and Differential Preferences," *Journal of Marketing,* 48 (Winter), 62-7.

Sudharshan, D., Jerrod H. May and Allan D. Shocker (1987) "A Simulation Comparison of Methods for New Product Location," *Marketing Science* 6 (Spring), 182-201.

Chapter 15: Stochastic Choice

Prerequisites: Chapter 5, Sections 3.9, 3.10

15.1 *Key Terminology*

The topic of this chapter is a set of choice models that deal with consumer behavior over time. We will begin by looking at data that tabulates what consumers do on two sequential purchase occasions. Do they buy the same brand twice, or do they switch from one brand to another? Later in the chapter we will look at the number of times a particular brand has been purchased, a type of data often called *purchase-incidence* data.

In some cases, we will assume that the population being studied is homogeneous. This is tantamount to the Gauss-Markov assumption [presented in Equation (5.16)] that we typically make in the general linear model, that is, that each observation can be described by the same parameter. In other cases, we may assume that the population being studied is heterogeneous with that parameter taking on different values. The parameter may itself follow some sort of distribution, often called a *mixing distribution*.

There is a different sort of homogeneity-heterogeneity distinction that comes up in models dealing with data collected over time. Regardless as to whether each unit, browser, consumer or household in the population can be described by the same parameter, is it possible that the parameter can change over time? A parameter that remains invariant across time periods is generally referred to as being *stationary* rather than homogeneous. More formally, we would define stationarity for a parameter θ such that

$$\theta_t = \theta_{t'} = \theta \text{ for all } t, t' = 1, 2, \cdots, T. \tag{15.1}$$

That terminology out of the way, let us now turn to the brand switching matrix which contains the key raw data for the models of the next few sections.

15.2 *The Brand Switching Matrix*

In what follows we will assume that we have three brands; call them A, B and C. Of course this terminology should not obscure the generality of the type of data we will be discussing. The three brands might actually be three Web sites, for example. In any case, in this section for each household we will be looking at a series of observations across T time periods: $y_1, y_2, \cdots, y_t, \cdots y_T$. We might admit here that the y_t values should also have a subscript for household, but that is dropped for notational convenience. You can think of the value y_t as being randomly selected from some population of households. For now we will look at T = 2 purchase occasions and organize the data from these two occasions in a two way contingency table that might look a lot like the one below:

		Purchase Occasion Two			
		A	B	C	
Purchase	A	10	5	10	25
Occasion	B	8	12	5	25
One	C	10	10	30	50

The table tells us that, for example, 10 households bought brand A on week one and then bought it again on week 2. On the other hand, of the 25 households who bought brand A on week one, 5 of

them switched to brand B on the second purchase occasion. It will be useful to be clear on different sorts of probabilities that can be formed from raw data such as these. An example of a *joint probability* would be the probability that a household in the sample bought A on week (occasion) one *and* then B on week 2, in other words Pr(y_1 = A and y_2 = B). We can also write this as Pr(A, B). Making the notation a bit more general, let us define Pr(j, k) as the joint probability that brand j is chosen on the first occasion and k on the second. From the table we can see that Pr(A, B) = 5/100 since 5 families from the sample of 100 families did just that.

A *marginal probability* gives the summary of a row or a column. For example, what is the probability of buying brand A on week one? The answer is 25/100, as 25 out of 100 families did that, and that figure also happens to be the market share for brand A on week one. As such we might use the letter m and notate that value $m_A^{(1)}$. Alternatively we could also use an expression like Pr(A), where it is understood we are talking about week one.

Finally, a *conditional probability* involves subsetting the table in some way. A conditional probability looks at the odds of something happening within that subset of the table. We might ask, given that a family bought A on week one, what is the conditional probability that they would turn around and buy B on week two? In other words, what is Pr(y_2 = B | y_1 = A)? A vertical bar is traditionally used to indicate a conditional probability. Here the numerator differs from the joint probability. You can think of this as the probability of B conditional on A, or *given* A. In either case, Pr(B | A) = 5/25, as there are 25 families who bought brand A on week one, and of these, 5 bought B on the next occasion. Again we could make the notation a bit more general by referring to Pr(k | j), or p_{jk}, as the conditional probability that brand k is chosen on the next occasion given that j was chosen on the previous occasion. While the notation p_{jk} will be used to refer to Pr(k | j), this probability is actually in position j, k of the *transition matrix*, illustrated below.

We might note that

$$m_j^1 = \sum_k \Pr(j,k), \tag{15.2}$$

$$\Pr(k \mid j) = \frac{\Pr(j,k)}{\Pr(j)} \text{ and that} \tag{15.3}$$

$$\sum_k \Pr(k \mid j) = 1. \tag{15.4}$$

In all three cases above the summation over the index k is taken to mean over all J brands in the study that appear in the switching matrix. Here the value m_j^1 is the share for brand j on week 1.

15.3 *The Zero-Order Homogeneous Bernoulli Model*

In this section we will once again be looking exactly two purchase occasions, i. e. T = 2. We begin by contemplating exactly two brands, A and B, and we will look at this situation with a particularly simple model. The zero-order homogeneous Bernoulli model assumes that on any purchase occasion the probability that A is bought is p. Here are the joint probabilities:

		Occasion Two	
		A	B
Occasion One	A	p^2	$p(1-p)$
	B	$(1-p)\,p$	$(1-p)^2$

For example, looking at the joint probability Pr(A, A), according to the model we have two independent events, each one occurring with a probability of p. The probability of two independent events can be calculated by multiplication. That the two events are independent is one of the strongest assumptions of the model. In effect, it assumes no feedback from one purchase event to the next. In other words, this model is zero-order just like a series of coin flips. Recall that with a fair coin, regardless of how many heads in a row have come up, the probability of a head on the next toss is still exactly .5.

The joint probability of any string of purchases can be calculated from multiplication as in

$$\Pr(A, B, A, A, B, \cdots) = p \cdot (1-p) \cdot p \cdot p\,(1-p) \cdot \cdots$$

The probability of r purchases of A out of T occasions would be

$$\binom{T}{r} p^r (1-p)^{T-r} \tag{15.5}$$

where the notation $\binom{T}{r}$ refers to the number of combinations of T things taken r at a time and is given by

$$\binom{T}{r} = \frac{T!}{r!(T-r)!}$$

and $T! = T \cdot (T-1) \cdot (T-2) \cdots 1$. The conditional probabilities can also be displayed in the same occasion-by-occasion format. When displayed as below, the table is called a transition matrix.

		Occasion Two	
		A	B
Occasion One	A	p	(1 - p)
	B	p	(1 - p)

The elements of the transition matrix, for example Pr(k | j), the probability that k is chosen given that j was chosen previously, are notated p_{jk} since that conditional probability arises from row j and column k.

15.4 *Population Heterogeneity and The Zero-Order Bernoulli Model*

Lets say that the value of p is itself a random variable, rather than a fixed parameter that describes the population of households, but there is still no feedback from one occasion to the next. On the surface it seems that this should imply, just as in a series of coin flips, that the next flip should not depend on what happens in any previous flips, right? It turns out the population heterogeneity and the lack of stationarity over time have similar implications in switching data. To get a handle on

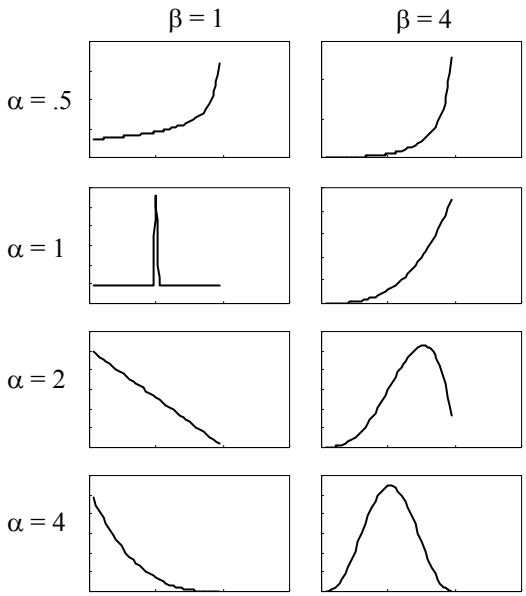
the nature of the heterogeneity of the value of p, we typically use the *Beta distribution* (Lilien and Kotler 1983), where

$$\Pr(p) = c_1 p^{\alpha-1} (1 - p)^{\beta-1} \tag{15.6}$$

The constant c_1 is a place holder that needs to be there to make sure that the distribution integrates to 1, i. e. it must be the case that

$$\int_{-\infty}^{\infty} \Pr(p)dp = 1$$

because Pr(p) is a density function (see Section 4.2). The two parameters of this distribution, α and β, control the shape of it. As compared to the normal, a wide variety of shapes are possible! Some idealized examples are pictured below in a graph that shows the Pr(p) on each of the y-axes:



Given some value of p, the likelihood of r purchases out of T occasions (Lilien and Kotler 1983) is

$$\Pr(r, T \mid p) = c_2 p^r (1 - p)^{T-r}. \tag{15.7}$$

The constant c_2 is a place holder for $\binom{T}{r}$, which does not figure into the derivation that follows. At this time it is appropriate to invoke the name of the Reverend Thomas Bayes, given that his name is attached to a simple theorem that connects two different sorts of conditional probabilities. For any two events, a and b, we know that by definition

$$\Pr(a \mid b) = \frac{\Pr(a, b)}{\Pr(b)}$$

but also that

$$\Pr(b \mid a) = \frac{\Pr(a, b)}{\Pr(a)}.$$

This suggests that there are two ways to write Pr(a, b),

$$\Pr(a, b) = \Pr(a \mid b) \cdot \Pr(b) = \Pr(b \mid a) \cdot \Pr(a),$$

which, when set equal to each other yields

$$\Pr(a \mid b) \cdot \Pr(b) = \Pr(b \mid a) \cdot \Pr(a)$$

$$\Pr(a \mid b) = \frac{\Pr(b \mid a) \cdot \Pr(a)}{\Pr(b)}.$$

From his theorem we can deduce that

$$\Pr(p \mid r, T) = \frac{\Pr(r, T \mid p) \Pr(p)}{\Pr(r, T)}. \tag{15.8}$$

In the numerator of the right hand side we see the likelihood of the data given the model from Equation (15.7), i.e. Pr(r, T | p). The density for p, assumed to be beta distributed, is also in the numerator. This is usually called the *prior distribution*, or sometimes just the *priors*. The left hand side also has a name, the *posterior probability*. It is the posterior probability of choice on the next occasion given a history of r purchases out of T occasions. If we define c_3 as 1/ Pr(r, n), then the posterior probability can be rewritten as

$$\begin{aligned} \Pr(p \mid r, T) &= c_3 \cdot \Pr(r, T \mid p) \cdot \Pr(p) \\ &= c_4 \cdot p^r (1-p)^{T-r} \cdot p^{\alpha-1} (1-p)^{\beta-1} \end{aligned} \tag{15.9}$$

which means that the posterior probability looks like a beta distribution with parameters $\alpha^* = \alpha + r$ and $\beta^* = \beta + T - r$. The upshot is that even though there is no memory or purchase feedback in this model, the posterior probability makes it look like there is. But the reason for this is that the population is not homogeneous. If we collect up all the households for which no one bought A, we probably have a group for whom p is lower than average. Dividing the sample of households in this way makes it look like there is contagion - a bunch of B's in a row lead to a higher probability of another B, not another flip of the coin.

We can estimate the choice parameter, p, using (Lilien and Kotler 1983)

$$\hat{p} = E(p \mid r, T) = \frac{\alpha + r}{\alpha + \beta + T}. \tag{15.10}$$

For example, for T = 3 we could look at eight possible triples that could occur with two brands and three weeks; AAA, AAB, ABA, ABB, BAA, BAB, BBA, BBB. The value of r is 0 for triple BBB. According to the model, the prediction for all those with three purchases of Brand B in a row would be

$$\hat{p} = E(p \mid 0, 3) = \frac{\alpha}{\alpha + \beta + 3}.$$

For r = 1 and T = 3 we could have ABB, BAB and BBA. All three sequences lead to the same estimate on trial 4,

$$\hat{p} = E(p \mid 1, 3) = \frac{\alpha + 1}{\alpha + \beta + 3}$$

As you can see, we can derive values for the choice probabilities, that is, values of p, on week 4. These probabilities arise from the more fundamental parameters that underlie the distribution of p, namely α and β, which are the unknowns and as such must be estimated from the sample. We could certainly minimize Pearson Chi Square across the eight data points from the triples. According to Minimum Pearson Chi Square, we pick values of α and β in such a way as to make

$$\hat{\chi}^2 = \sum_{j}^{8} \frac{(p_j - \hat{p}_j)^2}{\hat{p}_j} \tag{15.11}$$

as small as possible. We could also use modified minimum Chi Square or Maximum Likelihood. To do any of these we would need to determine the derivates of the objective function and drive them to zero,

$$\frac{\partial \hat{\chi}^2}{\partial \alpha} = \frac{\partial \hat{\chi}^2}{\partial \beta} = 0,$$

using the methods described in Section 3.9. As there are eight triplets from three weeks worth of purchases, and two unknowns, the model can be tested against Chi Square on 6 degrees of freedom.

15.5 *Markov Chains*

Now we will look at models that assume homogeneity across consumers or households, but not zero memory. In fact, a defining aspect of a Markov chain is that the system has memory that goes back one time period. If we define y_t as the brand chosen on occasion t, this memory can be described as

$$\Pr(y_t = j \mid y_{t-1}, y_{t-2}, \cdots, y_0) = \Pr(y_t = j \mid y_{t-1}). \tag{15.12}$$

We also assume stationarity which can be interpreted as the statement below:

$$\Pr(y_t = j \mid y_{t-1}) = \Pr(y_{t'} = j \mid y_{t'-1})$$

for all t, t′ and j.

A Markov chain is characterized by a transition matrix and an initial state vector. The transition matrix consists of the conditional probabilities Pr(k | j) such that $\sum_k \Pr(k \mid j) = 1$. A sample transition matrix is presented below:

		Occasion t + 1	
		A	B
Occasion t	A	.7	.3
	B	.5	.5

For example, in the lower left hand corner we see Pr(A | B) which is element 2,1 (p_{21}) in the table and is equal to .5. The second characterizing feature of a Markov chain is the initial vector which represents the market shares at time zero. A typical element would be $\{m_j^{(0)}\}$ which is the market share for brand j at time 0. In that case we can define the J by 1 vector of shares as

$$\mathbf{m}^{(0)} = [m_1^{(0)} \quad m_2^{(o)} \quad \cdots \quad m_J^{(0)}]$$

Given a transition matrix and an initial state, we can now predict the market shares for any time period. For example, looking at brand k, we might ask what will the share of brand k be after one week. To do this, we can use the *Law of Total Probability*. After time 0 there are J things that could have happened, that is to say there are J ways for k to be picked at time 1. A purchaser of brand 1 could have switched to k, a purchaser of brand 2 could have switched to k, and so forth until we reach the last brand, brand J. This is illustrated below:

$$m_k^{(1)} = \underbrace{Pr(k \mid 1)}_{\text{Pr (buy k given previous purchase of brand 1)}} \cdot \underbrace{m_1^{(0)}}_{\text{Pr (bought 1 previously)}} + Pr(k \mid 2) \cdot m_2^{(0)} + \cdots + Pr(k \mid J) \cdot m_J^{(0)}$$

We can use a slightly more elegant notation to say the same thing as

$$m_k^{(1)} = \sum_j^J p_{jk} m_j^{(0)} .$$

Here note that the law of total probability has us running down the rows of the **P** matrix, that is, running through all the ways that event k can happen at time t + 1. We can also express all of the market shares at one time using linear algebra,

$$[\mathbf{m}^{(1)}]' = [\mathbf{m}^{(0)}]'\mathbf{P}$$

$$[\mathbf{m}^{(2)}]' = [\mathbf{m}^{(1)}]'\mathbf{P} = [\mathbf{m}^{(1)}]' = [\mathbf{m}^{(0)}]'\mathbf{PP}$$

$$[\mathbf{m}^{(3)}]' = [\mathbf{m}^{(2)}]'\mathbf{P} = [\mathbf{m}^{(1)}]' = [\mathbf{m}^{(0)}]'\mathbf{PPP}$$

$$\cdots = \cdots = \cdots = \cdots$$

$$[\mathbf{m}^{(t)}]' = [\mathbf{m}^{(0)}]'\mathbf{P}^t . \tag{15.13}$$

We frequently assume an equilibrium such that the share vector no longer changes and estimate the elements of **P** from panel data. These elements themselves may be modeled with a smaller number of parameters that reflect the fundamental marketing concepts that are driving the data.

Recall that in the zero-order homogeneous Bernoulli model the transition matrix took on the appearance:

$$\begin{bmatrix} p & 1-p \\ p & 1-p \end{bmatrix}.$$

Here remember that the rows represent the state of the market at time t while the columns are the states at time t + 1. Element p_{jk} is the conditional probability, Pr(k | j).

Something we might call the *Superior-Inferior model* has a transition matrix

$$\begin{bmatrix} 1 & 0 \\ p & 1-p \end{bmatrix}.$$

No one who ever tries the first brand goes back to the second. One of the two states is an absorbing state - eventually the whole market will end up there.

In the *Variety-Seeking model* the propensity to buy a brand again is reduced by some fraction v:

$$\begin{bmatrix} p - vp & - \\ - & (1-p) - v(1-p) \end{bmatrix}$$

You will note that since $\sum_{k}^{J} p_{jk} = 1$, we can figure out one column by subtraction. Also note what happens as v goes from 0 to 1. The closer v gets to 0, the closer the model resembles the Bernoulli.

How would we estimate the parameters v and p? We could look at the 8 triples that are possible, AAA, AAB, ···, BBB. Each one has a prediction from the model. For example, for AAA we would have

$$\Pr(AAA) = (p - vp)^3.$$

We would have 8 data points, and two unknowns, and we could just use Minimum Pearson Chi Square, Maximum Likelihood, or other methods as described in Section 12.4 as well as for the logit model in Sections 13.3 and 13.4.

15.6 *Learning Models*

The models of this section are part of *Linear Operator Theory*, which was originally applied in Marketing to learning about frozen orange juice by Kuehn. Here we are going to assume that we have but one brand of interest, that is either purchased or not:

$$y_t = \begin{cases} 1 & \text{if the brand of interest is bought at time t} \\ 0 & \text{if the brand of interest is not bought at time t} \end{cases}$$

If our brand is purchased at time t - 1, we apply the acceptance operator and presumably learning occurs:

$$p_t = \alpha_1 + \beta_1 p_{t-1}$$

while on the other hand, if the brand is rejected, we apply the rejection operator:

$$p_t = \alpha_2 + \beta_2 p_{t-1}.$$

Consider what happens when a loyal consumer repeatedly buys our brand,

$$p_1 = \alpha_1 + \beta_1 p_0$$

$$p_2 = \alpha_1 + \beta_1 [\alpha_1 + \beta_1 p_0]$$

$$.... = ...$$

or working recursively backwards from time t

$$p_t = \alpha_1 + \beta_1 p_{t-1}$$

$$p_t = \alpha_1 + \beta_1 [\alpha_1 + \beta_1 p_{t-2}]$$

$$p_t = \alpha_1 + \beta_1 [\alpha_1 + \beta_1 (\alpha_1 + \beta_1 p_{t-3})]$$

Now, multiplying out this last version we have

$$p_t = \alpha_1 + \beta_1 \alpha_1 + \beta_1^2 \alpha_1 + \beta_1^3 p_{t-3}.$$

Eventually, we note that a pattern emerges so that we have

$$\begin{aligned} p_t &= \alpha_1 + \beta_1 \alpha_1 + \beta_1^2 \alpha_1 + \beta_1^3 \alpha_1 + \beta_1^4 \alpha_1 + \cdots \\ &= \alpha_1 [1 + \beta_1 + \beta_1^2 + \beta_1^3 + \cdots]. \end{aligned} \tag{15.14}$$

The term in the brackets is an infinite series, but that does not mean that it is equal to infinity. Call it b

$$\begin{aligned} b &= 1 + \beta_1 + \beta_1^2 + \beta_1^3 + \cdots \\ &= \sum_{i=0}^{\infty} \beta_1^i. \end{aligned} \tag{15.15}$$

For Equation (15.15) to work requires that $0 \le \beta_1 < 1$. If we multiply Equation (15.15) by β_1 we get

$$\beta_1 b = \beta_1 + \beta_1^2 + \beta_1^3 + \cdots \tag{15.16}$$

Subtract Equation (15.16) from Equation (15.15) above and the difference is 1:

$$b - \beta_1 \cdot b = 1$$

$$b(1 - \beta_1) = 1$$

$$b = \frac{1}{1-\beta_1}. \tag{15.17}$$

Combining Equation (15.14) and Equation (15.17), we can conclude that if the brand is always purchased, the probability will approach

$$p_t = \frac{\alpha_1}{1-\beta_1}, \tag{15.18}$$

a phenomenon known as incomplete habit formation. In this Linear Operator Theory, if $\beta_1 = \beta_2 = 0$ then we end up with a transition matrix just like the one shown below:

$$\begin{bmatrix} \alpha_1 & 1-\alpha_1 \\ \alpha_2 & 1-\alpha_2 \end{bmatrix}$$

which is a zero-order Bernoulli model!

15.7 *Purchase Incidence*

The main exemplar of a purchase-incidence model uses the Negative Binomial Distribution or NBD. In order to lead into that, however, we will start with two simpler models, the first of which is the binomial, named after the terms in the expansion of

$$(q + p)^T$$

with $q = 1 - p$. Term number $r + 1$ is $q^{T-r} p^r$ which we have already seen used in the expression for the probability of T things taken r at a time in Equations (15.5) and (15.7):

$$\binom{T}{r} p^r (1-p)^{T-r}$$

The term $\binom{T}{r}$ gives the number of ways out T to have r "successes" while $p^r(1-p)^{T-r}$ is the probability of each one of those ways. Now, consider a table from a panel of n households, with each household being categorized in terms of how many purchases of our brand that they have executed during the T week study period:

r	Number of Households
0	f_0
1	f_1

2	f_2
...	...
	f_T
Total	n

with a typical entry being f_r, which gives the number of households with r purchases during the study period. These are the data that we will attempt to account for with the model. The binomial model states simply that the probability of a purchase by any household on any week is p. We can estimate p using a particularly simple method called the *method of moments*. It is the case that

$$\bar{x} = E(r) = pT \tag{15.19}$$

gives the average number of purchases across households, or in other words, the average number of purchases per household. Solving for p we have simply

$$p = \frac{\bar{x}}{T}.$$

For example, if the average household purchase 2 items out of 4 occasions, then p = 2/4 = .5. According to the binomial model, we could substitute .5 for p in the formula

$$\hat{f}_r = T \cdot \binom{T}{r} (1-p)^{T-r} p^r. \tag{15.20}$$

We could test the model with a Chi Square that compares the predicted and observed frequencies of households with 1, 2, ···, T purchases.

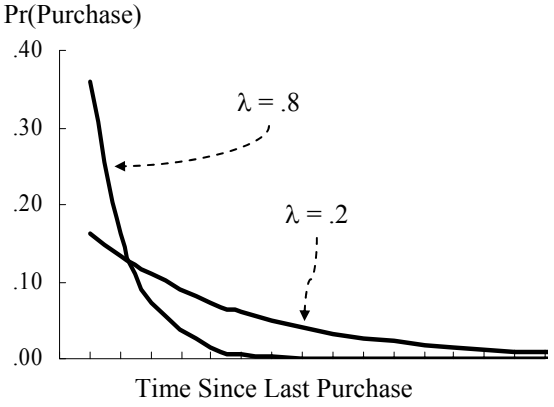
The Poisson model arises from the Binomial by letting $T \to \infty$ and $p \to 0$ but holding $Tp = \lambda$. The model originated from studies of the deaths of Prussian soldiers from kicks to the head by horses, apparently a worrisome occupational hazard. The number of Army corps with one death, two deaths, and so forth, was tabulated. The Poisson model asserts that

$$\hat{f}_r = n \cdot e^{-\lambda} \frac{\lambda^r}{r!}. \tag{15.21}$$

Fortunately for Prussian soldiers, the Poisson, which means fish in French but is actually named after it's inventor, is considered a distribution for "rare" events. The model assumes that there are a large number of small time periods with a small, but constant purchase probability in any time period. This is no doubt more realistic than the Binomial model, but unfortunately the Poisson makes an odd prediction about the probability that t time periods pass between one purchase occasion and the next

$$Pr(t) = \lambda e^{-\lambda t} \tag{15.22}$$

which is a special case of the *exponential distribution*. That this assumption is not in keeping with the reality of shopping can be seen in the graph below that looks at the relationship between time elapsed and the probability of a purchase:



The Poisson has the interesting property that its mean is equal to its variance,

$$E(r) = \bar{x} = \lambda$$

$$V(r) = s^2 = \lambda.$$

We could easily use the Method of Moments to estimate λ, and of course we also have at our disposal Minimum χ^2, which would require that we compare the f_r and $\hat{f}_r$ values, Maximum Likelihood, and so forth.

15.8 *The Negative Binomial Distribution Model*

The NBD model is named from the terms in the expansion of $(q - p)^{-r}$. The distribution can arise in a number of ways. For example, it could represent the probability that T trials will be needed for r successes. In effect, it is a binomial where the number of coin tosses is itself the random outcome. It could also represent a Poisson distribution with a contagion process such that the Poisson parameter λ changes over time. Another possible mechanism that leads to the NBD is where we have a Poisson model but the λ values is distributed across households according to the gamma distribution. The gamma is part of the general family of distributions that includes the Chi Square as a special case. According to the NBD model, the number of households purchasing the brand under study is

$$\hat{f}_r = n \cdot \left(\frac{k}{k+m}\right)^k \left(\frac{m}{k+m}\right)^m \frac{\Gamma(k+r)}{\Gamma(k)\,r!}. \tag{15.23}$$

The gamma function, Γ(·), not to be confused with the gamma distribution, acts like a factorial operator (the ! symbol) for non-integral arguments. For integral q, Γ(q) = (q + 1)!. In general,

$$\Gamma(q) = \int_0^\infty x^{q-1} e^{-x} dx. \tag{15.24}$$

Here we might note certain similarities between the Binomial model in Equation (15.20) and the Negative Binomial in Equation (15.23). In the latter, the role of p is played by k/(k + m) while 1-

p is analogous to m/(k + m). As before, we will be estimating k and m according to the method of moments, or using ML or Minimum Chi Square.

Here we might note certain similarities between the Binomial model in Equation (15.20) and the Negative Binomial in Equation (15.23). In the latter, the role of p is played by $k/(k+m)$ while 1 - p is analogous to $m/(k+m)$. As before, we will be estimating k and m according to the method of moments, or using ML or Minimum Chi Square.

References

Kahn, Barbara E., Manohar U. Kalwani and Donald G. Morrison (1988) "Niching Versus Change -of-Pace Brands: Purchase Frequencies and Penetration Rates to Infer Brand Positionings," *Journal of Marketing Research*, 25 (November), 384-90.

Kahn, Barbara E., Manohar U. Kalwani and Donald G. Morrison (1984) "Measuring Variety-Seeking and Reinforcement Behaviors Using Panel Data," *Journal of Marketing Research* 23 (May), 89-100.

Lilien, Gary L. and Philip Kotler (1983) *Marketing Decision Making*. New York: Harper and Row.

Schmittlein, David C., Albert C. Bemmaor and Donald G. Morrison (1985) "Why Does the NBD Model Work? Robustness in Representing Product Purchases and Imperfectly Recorded Purchases," *Marketing Science* 4 (Summer), 255-56.

Wheat, Rita D. and Donald G. Morrison (1990) "Assessing Purchase Timing Models: Whether or Not is Preferable to When," *Marketing Science,* 9 (Spring), 162-70.

Section V: Economics and Econometrics

Chapter 16: Microeconomics

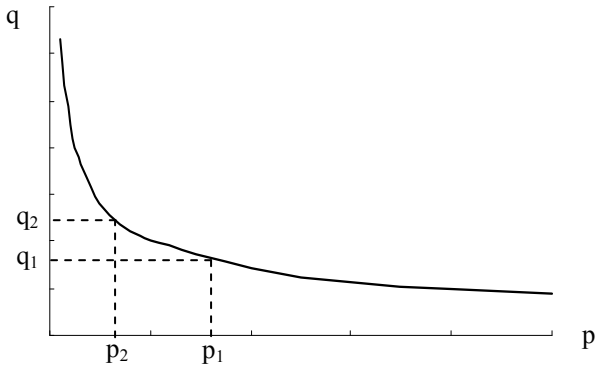
Prerequisites: Chapter 5

16.1 *The Notion of Elasticity*

In this section we will be making four key assumptions about demand for products and services.

(1) Consumers maximize utility,
(2) Consumers have full knowledge of all relevant market conditions,
(3) Sellers maximize short-term profit and
(4) Demand is infinitely divisible.

Imagine that we are looking at the relationship between price and quantity demanded. In the figure below, we have highlighted two price points, p_1 and p_2, and the two corresponding demand points.



The elasticity represents the percentage change in the quantity demanded, which is represented on the y axis of the graph above, divided by the percentage change in the price demanded, which shows up on the x axis. We will be using the symbol e to refer to the elasticity. In that case we have

$$e = \frac{q_1 - q_2/q_1}{p_1 - p_2/p_1} = \frac{\Delta q/q}{\Delta p/p} = \frac{\Delta q}{\Delta p} \cdot \frac{q}{p}.$$

If we assume that Δq and Δp can be made arbitrarily small, meaning that price and demand are both infinitely divisible, we can the take the following step

$$e = \frac{\Delta q}{\Delta p} \cdot \frac{q}{p}$$

$$= \frac{dq}{dp} \cdot \frac{q}{p} \qquad (16.1)$$

where dq/dp is the derivative of q with respect to p. By time honored tradition, if $-1 \le e \le 0$ we say that the demand is inelastic. On the other hand, if $e < -1$ we say that demand is elastic.

In Chapter 5 we studied the linear model which in this context would be

$$q_i = \beta_0 + p_i \beta_1. \tag{16.2}$$

Dropping the subscript i, and since $dq/dp = \beta_1$ (if you wish you can review Section 3.2), obviously under the linear model in Equation (16.2) it is then the case that

$$e = \beta_1 \frac{p}{q}.$$

Continuing to assume that the model of Equation (16.2) holds, we can now substitute $\beta_0 + p\beta_1$ for q to give us

$$e = \beta_1 \frac{p}{\beta_0 + p\beta_1}.$$

A linear demand function creates a situation in which the elasticity depends on p. Now lets try (as we did in Section 7.6) a quadratic function,

$$q_i = \beta_0 + p_i \beta_1 + p_i^2 \beta_2.$$

Then, again dropping the subscript i, $dq/dp = \beta_1 + 2p\beta_2$ and therefore, according to Equation (16.1) we have

$$e = (\beta_1 + 2p\beta_2)\frac{p}{q}.$$

OK, now we are ready for the Cobb-Douglas function that models demand as

$$q_i = \beta_0 p_i^{\beta_1} \text{ or} \tag{16.3}$$

$$\ln q_i = \ln\beta_0 + \beta_1 \ln p_i. \tag{16.4}$$

Note that while the model is nonlinear, we can easily estimate it using OLS because it can be linearized by taking the log of the independent and dependent variables. When we estimate Equation (16.4) we get the same value of interest, β_1, that we see in Equation (16.3). The derivative is

$$\frac{dq}{dp} = \beta_1 \beta_0 p^{\beta_1 - 1}$$

so that

$$e = \beta_1 \beta_0 p^{\beta_1 - 1} \frac{p}{q}.$$

Again, we see q in the equation so we substitute the model in Equation (16.3) to get

$$e = \frac{\beta_1 \beta_0 p^{\beta_1 - 1} p^1}{\beta_0 p^{\beta_1}}.$$

Note that p has been written as p^1 just so that we can use the rule of Equation (3.7) and end up with p^{β_1} in the numerator. Everything cancels, except for a single term and we have, for the Cobb-Douglas function,

$$e = \beta_1 \tag{16.5}$$

which shows us that the Cobb-Douglas is a constant elasticity model, meaning that the elasticity stays the same all along the x-axis of price.

16.2 *Optimizing the Pricing Decision*

In this section we are not going to assume either the Cobb-Douglas function of Equation (16.3) or any other particular demand function. Instead, we leave it that sales are a function of price, i.e. $q = f(p)$. But we are not optimizing demand, we are interested in optimizing profit which requires that we take costs into account. Lets say that costs are a function of quantity demanded, i. e. $c = g(q)$. In summary, we wish to make

$$\rho = \text{revenue - cost} = pq - g(q) \tag{16.6}$$

as large as possible. The breakeven point occurs when

$$pq = g(q)$$

$$p = \frac{g(q)}{q}$$

as revenue is equal to cost at that point. But we don't want to just break even, instead we want $d\rho/p = 0$ as this would be the point at which the function is at an extreme point. We have

$$\frac{d\rho}{dp} = \frac{dp \cdot q}{dp} + \frac{dg(q)}{dp} = 0$$

since the sum of the derivatives are equal to the derivative of the sum [Equation (3.12)]. When the above equation is at zero,

$$\frac{dp \cdot q}{dp} = \frac{dg(q)}{dp} .$$

Rewriting the Equation (16.6), we see that

$$\rho = p \cdot f(p) - g[f(p)]$$

meaning that according to the chain rule [Equation (3.14)] the derivative is

$$\frac{d}{dp} p \cdot f(p) - g[f(p)] = f'(p) - g'[f(p)] \cdot f'(p) .$$

References

Blattberg, Robert C. and Kenneth J. Wisniewski (1989) "Price-Induced Patterns of Competition," *Marketing Science,* 8 (fall), 291-309.

Gerard J. Tellis (1988) "The Price Elasticity of Selective Demand: A Meta-Analysis of Econometric Models of Sales," *Journal of Marketing Research,* 25 (November), 331-41.

Russell, Gary J. and Ruth N. Bolton (1988) "Implications of Market Structure for Elasticity Structure," *Journal of Marketing Research,* 25 (August), 229241.

Hauser, John R. (1988) "Competitive Price and Positioning Strategies,´ *Marketing Science* 7 (Winter), 76-91.

Bolton, Ruth N. (1989) "The Relationship between Market Characteristics and Promotional Price Elasticities," *Marketing Science,* 8 (Spring), 153-69.

Chapter 17: Econometrics

Prerequisites: Chapter 6, Sections 3.5 - 3.8

17.1 *The Problems with Nonrecursive Systems*

This chapter contains a mixture of ideas extended from the Chapters on Regression, in particular Chapter 5 and 6, and the chapters on covariance structure, in particular 9 and 10. To anticipate a theme of this chapter, econometricians have come up with a variety of ways to use the basic least squares philosophy to look at models with latent variables and complex causal structures. In this section we are concerned with nonrecursive systems, with equations of the form $\mathbf{y} = \mathbf{By} + \mathbf{\Gamma x} + \boldsymbol{\zeta}$, where $V(\boldsymbol{\zeta})$ is not diagonal, or it is impossible to arrange the sequence of y variables such that $\mathbf{B}$ is lower triangular. To illustrate the problems caused by nonrecursion, we start with a deceptively simple two equation system:

$$\begin{aligned} y_1 &= \beta_{12} y_2 + \zeta_1 \\ y_2 &= x_1 + y_1, \end{aligned} \tag{17.1}$$

where y_1 represents the expenditures on our product category, y_2 is income, and x_1 is all other expenditures, including savings. As we did in Chapter 10, we are dropping any subscript that references the individual observation in this section. However, the reader should keep in mind that ζ_1 is a random input to the model, and varies from one observation to the next. The second equation is known as an identity, since there is no error term. If we were to assume that $V(\zeta_1) = \sigma^2\mathbf{I}$, can we use the OLS approach of Chapter 5? Unfortunately not, since problems arise due to the covariance between y_2 and ζ_1. This becomes clear when we substitute the y_1 equation into the y_2 identity:

$$y_2 = (\beta_{12} y_2 + \zeta_1) + x_1$$

$$y_2 - \beta_{12} y_2 = \zeta_1 + x_1$$

$$y_2 = \frac{\zeta_1}{1-\beta_{12}} + \frac{x_1}{1-\beta_{12}}. \tag{17.2}$$

If we assume that $E(\zeta_1) = 0$ then we can say

$$E(y_2) = \frac{x_1}{1-\beta_{12}} \tag{17.3}$$

which means, by the definition of variance [Equation (4.7)], we get:

$$\begin{aligned} \mathrm{Cov}(\zeta_1, y_2) &= E\{[\zeta - E(\zeta)][y_2 - E(y_2)]\} \\ &= E\{\zeta_1[y_2 - E(y_2)]\} \end{aligned}$$

where we get to the second line above since $E(\zeta_1) = 0$. Now, substituting the results of Equation (17.2) and Equation (17.3) into the line above, we get

$$\mathrm{Cov}(\zeta_1, y_2) = E\left[\zeta_1 \left(\frac{\zeta_1}{1-\beta_{12}} \right) \right]$$

$$= \frac{1}{1-\beta_{12}} E(\zeta_1 \zeta_1) = \frac{\sigma^2}{1-\beta_{12}}.$$

Thus, y_2, which functions as an independent variable in the equation for y_1, is correlated with the error for that equation, ζ_1. This is a no-no. In this situation the usual least squares estimator $\hat{\boldsymbol{\beta}}$ is not consistent [consistency is defined in Equation (5.11), but see Johnson p 281-2 for a proof].

There are three solutions to this problem. First, there is what econometricians call *Full Information Maximum Likelihood* which is basically the covariance structure model covered in Chapter 10. Estimating a nonrecursive system using coviarance structural models can be tricky however. Second, there is what is known as *Indirect Least Squares* which takes advantage of reduced form, covered elsewhere [Equation (10.6)]:

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$$

$$\mathbf{y} - \mathbf{B}\mathbf{y} = \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$$

$$(\mathbf{I} - \mathbf{B})\,\mathbf{y} = \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$$

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{x} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta}$$

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{e}$$

We can use OLS to estimate the elements in $\hat{\mathbf{G}}$. The major problem here is that unless the model is just identified, with exactly the right number of unknowns, you cannot recover the structural parameters of theoretical importance in **B** and **Γ**.

Third, there is a technique called Two Stage Least Squares and we will now cover that.

17.2 *Two Stage Least Squares*

The basic strategy of Two Stage Least Squares, sometimes called *2SLS*, is to replace y_2 with $\hat{y}_2$ in Equation (17.1) above. To discuss the technique further, we need to revert to the notational convention of Chapters 5, 6 and 8 which explicitly makes reference to individual observations. Rather than refer to a particular endogenous variable as y_2, lets say, it is now a particular column of the **Y** matrix which has n rows, one row for each observation. To get the discussion started, we introduce some key vectors and matrices:

Array	Order	Description
$\mathbf{y}_{\cdot 1}$	$n \cdot 1$	Endogeneous variable of interest
$\mathbf{Y}_2$	$n \cdot (p-1)$	Other endogenous variables in the equation for $\mathbf{y}_{\cdot 1}$

β_2	(p-1) · 1	Structural parameters for $\mathbf{Y}_2$
$\mathbf{X}_1$	n · k_1	Exogenous variables in equation for $\mathbf{y}_{\cdot 1}$
γ	K_1 · 1	Structural parameters for $\mathbf{X}_1$
$\zeta_{\cdot 1}$	n · 1	Error in the equation for $\mathbf{y}_{\cdot 1}$

The model looks like

$$\mathbf{y}_{\cdot 1} = \mathbf{Y}_2 \beta_2 + \mathbf{X}_1 \beta_1 + \zeta_{\cdot 1}$$

Now we define the full set of exogenous variables as $\mathbf{X} = [\mathbf{X}_1 \mid \mathbf{X}_2]$. . In stage 1 we regress $\mathbf{Y}_2$ on $\mathbf{X}$ to produce:

$$\hat{\mathbf{Y}}_2 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_2 .$$

In stage 2 we regress $\mathbf{y}_{\cdot 1}$ on $\hat{\mathbf{Y}}_2$ and $\mathbf{X}_1$. This produces a formula for the unknowns as below:

$$\begin{bmatrix} \hat{\beta}_2 \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{Y}}_2'\hat{\mathbf{Y}}_2 & \hat{\mathbf{Y}}_2'\mathbf{X}_1 \\ \mathbf{X}_1'\hat{\mathbf{Y}}_2 & \mathbf{X}_1'\mathbf{X}_1 \end{bmatrix}^{-1} \begin{bmatrix} \hat{\mathbf{Y}}_2'\mathbf{y}_{\cdot 1} \\ \mathbf{X}_1'\mathbf{y}_{\cdot 1} \end{bmatrix}.$$

While $\mathbf{Y}_2$ may be correlated with $\zeta_{\cdot 1}$ we expect that $\hat{\mathbf{Y}}_2$ is not. It is not literally necessary to execute two stage least squares in two stages. Instead you can use

$$\begin{bmatrix} \hat{\beta}_{\cdot 2} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_2'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_2 & \mathbf{Y}_2'\mathbf{X}_1 \\ \mathbf{X}_1'\mathbf{Y}_2 & \mathbf{X}_1'\mathbf{X}_1 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Y}_2'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}_{\cdot 1} \\ \mathbf{X}_1'\mathbf{y}_{\cdot 1} \end{bmatrix}$$

or define $\mathbf{Y}_2 = \hat{\mathbf{Y}}_2 + \mathbf{E}_2$ so that

$$\hat{\mathbf{Y}}_2'\hat{\mathbf{Y}}_2 = \hat{\mathbf{Y}}_2'(\mathbf{Y}_2 - \mathbf{E}_2) = \hat{\mathbf{Y}}_2'\mathbf{Y}_2 = \mathbf{Y}_2'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1$$

Now rewriting,

$$\begin{bmatrix} \hat{\beta}_{\cdot 2} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_2'\mathbf{Y}_2 - k\mathbf{E}_2'\mathbf{E}_2 & \mathbf{Y}_2'\mathbf{X}_1 \\ \mathbf{X}_1'\mathbf{Y}_2 & \mathbf{X}_1'\mathbf{X}_1 \end{bmatrix}^{-1} \begin{bmatrix} (\mathbf{Y}_2' - k\mathbf{E}_2)\mathbf{y}_{\cdot 1} \\ \mathbf{X}_1'\mathbf{y}_{\cdot 1} \end{bmatrix}.$$

For k = 0 we have OLS and for k = 1 we have 2SLS. There is a technique called Limited Information Maximum Likelihood in which k is itself estimated.

17.3 *Econometric Approaches to Measurement Error*

We begin by noting that measurement error in the **y** vector is not a problem for regression. Assume the real model is

$$\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where $\tilde{\mathbf{y}}$ is the true value of the dependent variable vector. Instead, unfortunately, we observe

$$\mathbf{y} = \tilde{\mathbf{y}} + \boldsymbol{\delta}$$

where $\boldsymbol{\delta}$, in general, is not a null vector. We can write the true model

$$\mathbf{y} - \boldsymbol{\delta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} + \boldsymbol{\delta},$$

so that we just get a slightly different error term. Unless Cov($\boldsymbol{\delta}$, $\mathbf{X}$) $\neq$ $\mathbf{0}$ we will be OK. Now, however, lets contemplate what happens when there is measurement error on the x side. Imagine that we have the true model

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{e}$$

but we observe

$$\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{F} \tag{17.4}$$

instead. Rewriting the true model, we get

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{e}$$

$$= (\mathbf{X} - \mathbf{F})\boldsymbol{\beta} + \mathbf{e}$$

$$= \mathbf{X}\boldsymbol{\beta} - \mathbf{F}\boldsymbol{\beta} + \mathbf{e}$$

$$= \mathbf{X}\boldsymbol{\beta} + (\mathbf{e} - \mathbf{F}\boldsymbol{\beta}).$$

In this case we find out that the Cov($\mathbf{X}$, $\mathbf{F}\boldsymbol{\beta}$) is not going to vanish since $\mathbf{F}$ is a component of $\mathbf{X}$. Thus the error and the independent variables are correlated and the OLS estimator is not consistent. We can get around this problem using a technique called *Instrumental Variables*. We need to find a set of instruments, $\mathbf{X}_{(i)}$, that are independent of both the error vector $\mathbf{e}$ and the errors in the $\mathbf{X}$-variables, $\mathbf{F}$. We then estimate $\boldsymbol{\beta}$ below such that

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X})^{-1}\mathbf{X}'_{(i)}\mathbf{y}$$

and $\hat{\boldsymbol{\beta}}_{(i)}$ will then consistently estimate $\boldsymbol{\beta}$. From time to time we might use $\mathbf{Z}$ with 1's and -1's from a median split of the x variables.

17.4 *Generalized Least Squares*

GLS estimation has been discussed in Sections 6.8, 12.4 and 13.3. Here we review and further develop the concept of GLS with an eye to applying it to data that are collected across time and so cannot be considered independent. In the basic linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

in this section we will assume that $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$ where in general, $\mathbf{V} \neq \mathbf{I}$. Regardless as to the distribution of e, if we estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

we find that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, but this estimator no longer produces the best, or smallest, variance, $V(\hat{\boldsymbol{\beta}})$. Assuming that V is of full rank (see Section 3.7), $\mathbf{V}^{-1}$ exists and we can decompose it in the manner of Equation 3.38) such that

$$\mathbf{V} = \mathbf{P}'\mathbf{P}.$$

Using P to premultiply the linear model, we get

$$\mathbf{Py} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{Pe} \text{ or}$$

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}^*.$$

What are the properties of the new error term, $\mathbf{e}^*$? According to Theorem (4.9) we have

$$V(\mathbf{e}^*) = \mathbf{P}[\sigma^2\mathbf{V}]\,\mathbf{P}'$$

$$= \sigma^2\mathbf{P}[\mathbf{P}'\mathbf{P}]^{-1}\mathbf{P}'$$

and since V is of full rank, P is square and also of full rank so we can say that

$$V(e^*) = \sigma^2\mathbf{P}\,\mathbf{P}^{-1}\,(\mathbf{P}')^{-1}\mathbf{P}' = \sigma^2\mathbf{I}.$$

While we cannot believe in the Gauss-Markov assumption with **e**, we can with $\mathbf{e}^*$! Rather than minimizing $\mathbf{e}'\mathbf{e}$ as in OLS, we should minimize

$$\mathbf{e}^{*\prime}\mathbf{e}^* = \mathbf{e}\mathbf{P}'\mathbf{P}\mathbf{e} = \mathbf{e}'\mathbf{V}^{-1}\mathbf{e}$$

instead. Doing so, we pick our objective function as

$$f = \mathbf{e}^{*\prime}\,\mathbf{e}^* = (\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta})'(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta})$$

$$= \mathbf{y}^{*\prime}\,\mathbf{y}^* - 2\mathbf{y}^*\,\mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\beta}\mathbf{X}^{*\prime}\,\mathbf{X}^*\boldsymbol{\beta}.$$

In order to minimize f, we should set $\partial f/\partial\boldsymbol{\beta} = \mathbf{0}$ and solve for $\boldsymbol{\beta}$, as we will now do:

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^{*\prime}\boldsymbol{\beta} - 2\mathbf{X}^{*\prime}\mathbf{y}^{*} = \mathbf{0}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\prime}\mathbf{y}^{*}$$

and of course we end up with the usual formula, but using the transformed data matrices $\mathbf{X}^*$ and $\mathbf{y}^*$. Substituting back $\mathbf{PX} = \mathbf{X}^*$ and $\mathbf{Py} = \mathbf{y}^*$, we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\prime}\mathbf{y}^{*}$$

$$= (\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{y}$$

$$= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

The variance of this estimator is

$$V(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\sigma^2\mathbf{V})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

$$= \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.$$

This is all fine and dandy, but since $\mathbf{V}$ contains $\frac{n(n+1)}{2}$ unique elements, it is necessary that most of them be known a priori. But there is another identification issue. Since $V(\mathbf{e}) = \sigma^2\mathbf{V}$, we cannot uniquely identify both σ^2 and the elements of $\mathbf{V}$. That this is so can be seen by simply multiplying σ^2 by some value a and then dividing all of the elements of $\mathbf{V}$ by a and the model is unchanged. What we do is to set $Tr(\mathbf{V}) = Tr(\mathbf{I}) = n$.

We can estimate σ^2 using

$$s^2 = \frac{SS_{Error}}{n-k}$$

where

$$SS_{Error} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

We can construct *t*-statistics that allow us to test hypotheses of the form

$$H_0: \beta_i = 0$$

using the ith diagonal element of $s^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ in the denominator to create a *t*. One can also test one degree of freedom hypotheses such as

$$\mathbf{a}'\boldsymbol{\beta} = c$$

using

$$\hat{t} = \frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - c}{s^2\mathbf{a}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{a}}$$

and for more complex hypotheses of the form

$$H_0: \mathbf{A\beta} - \mathbf{c} = \mathbf{0}$$

we use

$$SS_H = (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{A}]^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})$$

to construct an F ratio numerator (with degrees of freedom equal to the number of rows in **A**), with s^2 in the denominator (with n - k degrees of freedom).

One area that we can apply GLS to occurs when the error in a regression model is not independent because the data are collected over time, leading to autocorrelated error. This may happen if we are analyzing the behavior of a particular firm, a particular store, category sales, purchases in a particular geographic region, and in many other cases in marketing where we look at data not collected across independent subjects. The next section speaks to that application of GLS.

17.5 *Autocorrelated Error*

When we collect data over time, rather than across a set of independent individuals, we run the risk that the error from observations that are closer together in time will be more closely related than a pair of errors that are farther apart from time. For example, looking at industry-wide sales of motor homes, we may fail to include every possible exogenous factor that there could be in a model for such sales. In fact, unless our model fits without error, it must be the case that we have omitted some important independent variables. Now, if any of those independent variables that did not find their way into our regression equation vary in a systematic way over time, for example, the weather, or consumer confidence, then the errors in our regression equation will also vary systematically over time. Of course, that would violate the Gauss-Markov assumption and necessitate some counter measure. Such as GLS. To begin to sketch this out, consider observation t on the dependent variable and the model for it,

$$\mathbf{y}_t = \mathbf{x}'_{t\cdot}\boldsymbol{\beta} + \mathbf{e}_t \tag{17.5}$$

where, needless to say, $\mathbf{x}'_{\cdot t}$ represents the t-th row of the matrix of independent variables, **X**. Given the argument in the preceding paragraph, we note that values of e_t are not independently distributed, but rather, adjacent observations follow the model

$$e_t = \rho e_{t-1} + \varepsilon_t. \tag{17.6}$$

In this context, the values ε_t represent an error for the error, if you will. We would also be remiss if we did not point out that a requirement of the model is that $|\rho| < 1$. The distribution of the ε_t is characterized as

$$\varepsilon_t \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{17.7}$$

which is to say that the the ε_t, unlike the e_t, are independently distributed. They behave like a white noise process, in summary. Repeating our model for the error,

$$e_t = \rho e_{t-1} + \varepsilon_t \tag{17.8}$$

we see that, since e_{t-1} appears in the right hand side, the model for e_{t-1} would contain e_{t-2} in it. Making that obvious substitution, we get

$$\begin{aligned} e_t &= \rho(\rho e_{t-2} + \varepsilon_{t-2}) + \varepsilon_t \\ &= \rho[\rho(\rho e_{t-3} + \varepsilon_{t-2}) + \varepsilon_{t-1}] + \varepsilon_t \\ &= \cdots. \end{aligned}$$

At this point the pattern should be obvious. Continuing the process of substitution, we end up with

$$\begin{aligned} e_t &= \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3} + \cdots \\ &= \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}. \end{aligned} \tag{17.9}$$

This last equation will look quite familiar if you have looked at Equation (15.17) or (18.15), being an infinite series. Now we wish to find out the expectation of the error. To determine the expectation of e_t from Equation (17.9), we keep in mind that the expectation of a sum is equal to the sum of the expectations [Equation (4.4)], and that therefore

$$\begin{aligned} E(e_t) &= E\left[\sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}\right] \\ &= \sum_{i=0}^{\infty} \rho^i E(\varepsilon_{t-i}) = 0, \end{aligned} \tag{17.10}$$

since ρ is a constant parameter that describes the population and by assumption $E(e_i) = 0$ for all i. Now we wish to figure out the variance of e_t, that is $V(e_t) = E[e_t - E(e_t)]^2$ according to Equation (4.7). Given that $E(e_t) = 0$, which we have just shown in Equation (17.10), we will only need to figure out $E(e_t^2)$. That will be made easier by recalling that all cross terms of the form $E(\varepsilon_t, \varepsilon_{t-j})$ will vanish as the ε_t are presumed independent, and that $a^0 = 1$ for any value a. So, squaring the second line of Equation (17.10) we have

$$E(e_t^2) = E(\varepsilon_t^2) + \rho^2 E(\varepsilon_{t-1}^2) + \rho^4 E(\varepsilon_{t-2}^2) + \cdots$$

$$= \sigma^2 + \rho^2\sigma^2 + \rho^4\sigma^2 + \cdots$$

$$= (1 + \rho^2 + \rho^4 + \cdots)\sigma^2.$$

So in the above equation we have an infinite series of the form $1 + \rho^2 + \rho^4 + \cdots$, call it s such that

$$s = 1 + \rho^2 + \rho^4 + \rho^8 + \cdots$$

$$\rho^2 s = \rho^2 + \rho^4 + \rho^8 + \cdots$$

$$s - \rho^2 s = 1$$

so that

$$s = \frac{1}{1-\rho^2}. \tag{17.11}$$

Putting all of this together, we conclude that

$$E(e_t^2) = \sigma_e^2 = \frac{\sigma^2}{1-\rho^2}. \tag{17.12}$$

To explore the covariances between e_t and e_{t-j} , we begin with j = 1. By definition, the covariance between e_t and e_{t-1} is given by

$$E(e_t e_{t-1}) = E[(\varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \cdots)(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \rho^2\varepsilon_{t-3} + \cdots)].$$

Looking at the right hand side of that equation, we will factor the ρ that appears in the left parentheses to give us

$$E(e_t e_{t-1}) = E\left\{[\varepsilon_t + \rho(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \cdots)](\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \rho^2\varepsilon_{t-3} + \cdots)\right\}.$$

Now, the two terms in the two parentheses on the right hand side are identical. We can write them as a single term squared. What's more, you will notice an ε_t all alone on the left of the right hand side. Its expectation is zero, and since there are no other values ε_t on the right hand side, the covariance of it and every other term will be zero. It thus vanishes without a trace. Rewriting, that gives us

$$E(e_t e_{t-1}) = \rho E[(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \rho^2\varepsilon_{t-3} + \cdots)^2]. \tag{17.13}$$

You will note that since ρ is a constant it can pass through the expectation operator [for a review, take a peek at Equation (4.5)]. Again, we remind you that $E(\varepsilon_t, \varepsilon_{t-1})$, that is the covariance between two different values of the ε_t are zero by the assumption of Equation (17.7). However, just

because the ε_t are independent does not mean that the e_t are. In fact, looking at Equation (17.13), we are almost ready to make a conclusion about the autocovariance of the e_t. The part in parentheses is just the model for the e_t, i.e. Equation (17.8). Its expectation squared must then be the variance of e_t, so that

$$E(e_t e_{t-1}) = \rho\sigma_e^2. \tag{17.14}$$

Following the same reasoning we find that the

$$Cov(e_t, e_{t-j}) = \rho^j\sigma_e^2. \tag{17.15}$$

Summarizing, we can say that the variance matrix of the e_t is $\sigma_e^2\mathbf{V} = \frac{\sigma^2}{1-\rho^2}\mathbf{V}$ with

$$\mathbf{V} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}.$$

Thus the GLS approach only needs to estimate two error related parameters, ρ and σ_ε^2. In the *Cochrane-Orcutt Iterative Procedure* we pick a starting value for ρ, calculate $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, then pick ρ in such a way as to minimize $\mathbf{e}'\mathbf{e}$ while holding $\hat{\boldsymbol{\beta}}$ fixed, and then re-estimate $\hat{\boldsymbol{\beta}}$ holding ρ fixed. One alternates between those two least squares steps until there is convergence. More general specifications of the nature of the error are possible. While in this section we have discussed a single autoregressive parameter, in much the same way that we talk about an AR(1) model in Section 18.4, just like with ARIMA models, you can have AR(2) or other processes.

17.6 *Testing for Autocorrelated Error*

Durbin and Watson (1950) proposed using

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{17.16}$$

as a test statistic for autocorrelated residuals. Here, the hypothesis being tested is H_0: $\rho = 0$. For positive autocorrelation the numerator will be small, while for negative autocorrelation the numerator will tend to be large. There is an upper limit (d_u) and a lower limit (d_l) for this statistic such that

$$\text{if } d < d_l, \text{ reject } H_0,$$

$$\text{if } d > d_u, \text{ fail to reject } H_0, \text{ and if}$$

$$\text{if } d_l < d < d_u$$

the test is inconclusive.

17.7 *Lagged Variables*

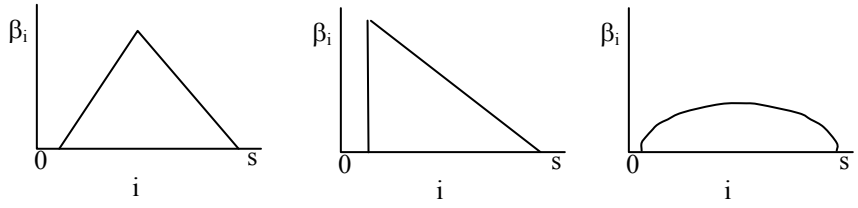
Suppose it is the case that consumers do not immediately react to a change in a marketing variable. In that case we would expect to see a relationship like the one below,

$$y_t = \beta_0 + x_{t-1}\beta_1 + e_t$$

or perhaps their reaction begins immediately but is distributed across several time periods, as in

$$y_t = \beta_0 + x_{t-1}\beta_1 + x_{t-2}\beta_2 + \cdots x_{t-s}\beta_s + e_t.$$

This is reasonable under many real life marketing situations. For example, the consumer may not immediately learn about a change in the market. Or perhaps, they are encumbered in their actions by inventory already on hand. However realistic this may be, there are unfortunately some problems with this approach. For one thing, what should "s" be? For another, we will be losing a degree of freedom for each lag, which is to say that the model is not very parsimonious. Finally, successive values of x might well be highly correlated, so that multicollinearity rears its head. What we can do is impose some sort of a priori structure on the values of the β_i. A graph of some possible structural assumptions is below:



Of course, any function can be represented by a polynomial of sufficiently high degree, fact exploited in ANOVA in Section 7.6. We can approximate, for example, a system with s = 7 lags with a polynomial of the third degree:

$$\beta_0 = a_0$$

$$\beta_1 = a_0 + a_1 + a_2 + a_3$$

$$\beta_2 = a_0 + 2a_1 + 4a_2 + 8a_3$$

$$\beta_3 = a_0 + 3a_1 + 9a_2 + 27a_3$$

$$\cdots = \quad \cdots$$

$$\beta_7 = a_0 + 7a_1 + 49a_2 + 343a_3$$

The reader will perhaps recognize that the coefficients for the a values are constant in the first column, linear in the second, quadratic in the third and cubic in the fourth. If we substitute these equations back into the model for s = 7, i. e.

$$y_t = \beta_0 + x_{t-1}\beta_1 + x_{t-2}\beta_2 + \cdots x_{t-7}\beta_7 + e_t,$$

we get after collecting the a_i terms

$$\begin{aligned} y_t = \beta_0 + &(x_t + x_{t-1} + x_{t-2} + \cdots x_{t-7})a_0 + \\ &(x_t + 2x_{t-1} + 3x_{t-2} + \cdots + 7x_{t-7})a_1 + \\ &(x_t + 4x_{t-1} + 9x_{t-2} + \cdots + 49x_{t-7})a_2 + \\ &(x_t + 8x_{t-1} + 27x_{t-2} + \cdots + 343x_{t-7})a_3 + e_t \end{aligned} \tag{17.17}$$

which is equivalent to an model with

$$y_t = \beta_0 + w_0a_0 + w_1a_1 + w_2a_2 + w_3a_3 + e_t, \tag{17.18}$$

where $w_0 = x_t + x_{t-1} + x_{t-2} + \cdots x_{t-7}$, and the other w values are defined as above in Equation (17.17). This is known as *Almon's Scheme*. If we define

$$\mathbf{K} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 7 & 27 & 343 \end{bmatrix}$$

using the coefficients for the x's, then

$$V(\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{K}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{K}' \tag{17.19}$$

lets you test the value s of the maximum lag, while

$$V(\hat{\mathbf{a}}) = \sigma^2(\mathbf{W}'\mathbf{W})^{-1} \tag{17.20}$$

lets you test the degree of the polynomial required to represent the lag structure.

While Almon's Scheme is quite compelling, another approach was proposed by Koyck, who used a geometric sequence. Koyck started with the infinite sequence

$$y_t = x_t\beta_0 + x_{t-1}\beta_1 + x_{t-2}\beta_2 + \cdots + e_t. \tag{17.21}$$

Now, assume that the β values are all of the same sign, and that

$$\sum_{i=0}^{\infty}\beta_i = c < \infty. \tag{17.22}$$

We now introduce the *backshift operator*, B, which is also prominently featured in Chapter 18. We define

$$Bx_t = x_{t-1}. \tag{17.23}$$

Of course, one can also say

$$BBx_t = B^2x_t = x_{t-2} \tag{17.24}$$

and so forth with $B^j x_t = x_{t-j}$. Given our two assumptions of Equations (17.21) and (17.22), we can rewrite the model as

$$y_t = x_t\beta_0 + x_{t-1}\beta_1 + x_{t-2}\beta_2 + \cdots + e_t$$

$$= \beta[w_0x_t + w_1x_{t-1} + w_2x_{t-2} + \cdots] + e_t$$

where $w_i \geq 0$ for i = 0, 1, 2, ···, ∞ and $\sum_{i=0}^{\infty} w_i = 1$. Given that, we can now rewrite the above equation as

$$= \beta[w_0 + w_1B + w_2B^2 + \cdots]x_t + e_t.$$

Now we introduce the major assumption of the Koyck scheme. The w's have a geometric relationship to each other as in

$$w_i = (1 - \lambda)\lambda^i \tag{17.25}$$

where $0 < \lambda < 1$. In that case

$$w_0 + w_1B + w_2B^2 + \cdots = (1-\lambda)(1 + \lambda B + \lambda^2B^2 + \cdots)$$

$$= (1-\lambda)\frac{1}{1-\lambda B}.$$

The fraction on the right hand side of the line immediately above is a consequence of the logic worked out in Equation (17.11) where we previously worked out the solution to an infinite series just like the one above. The upshot is that we can now write the model

$$\begin{aligned} y_t &= \frac{\beta(1-\lambda)}{1-\lambda B}x_t + e_t \\ (1-\lambda B)y_t &= \beta(1-\lambda)x_t + (1-\lambda B)e_t \\ y_t - \lambda By_t &= \beta(1-\lambda)x_t + e_t - \lambda Be_t \\ y_t &= \beta(1-\lambda)x_t + \lambda y_{t-1} + (e_t - \lambda e_{t-1}) \end{aligned} \tag{17.26}$$

As you can see, Koyck's scheme is characterized by autocorrelated error and lagged endogenous variables on the right hand side. Why would that be? Is there any marketing theory in which that would make sense? We will be finding out shortly.

17.8 *Partial Adjustment by Consumers*

The partial adjustment model posits that the optimal value of the y variable, y*, might depend on x. For example, y could be an amount spent on our brand and x is income. As the consumer wants to make an optimal choice, and if the relationship is linear, we would have

$$\tilde{y}_t = \beta_0 + x_t\beta_1, \tag{17.27}$$

but due to less than perfect information about the market, inventory considerations, inertia, or the cognitive costs of change, the consumer can only adjust a certain proportion of the way from his or her previous value, y_{t-1}, to the optimal value at $\tilde{y}_t$. In mathematical terms,

$$y_t - y_{t-1} = \gamma(\tilde{y}_t - y_{t-1}) + e_t \tag{17.28}$$

with $0 < \gamma < 1$. Substituting Equation (17.27) into Equation (17.28), we see that

$$y_t = \beta_0\gamma + \beta_1\gamma x_t + (1-\gamma)y_{t-1} + e_t$$

which bears a resemblance to Koyck's scheme, only here we have an intercept, and the error is not autocorrelated.

17.9 *Adaptive Adjustment by Consumers*

Another way that a similar equation may come about is through consumers adapting their expectations. Define $\tilde{x}_t$ as the expected level of x, and assume that some key consumer behavior depends on $\tilde{x}_t$. The value $\tilde{x}_t$ could be the best guess of the price of a good, something to do with its availability in the market, and so forth. The consumer's behavior should then appear as below

$$y_t = \beta_0 + \tilde{x}_t\beta_1 + e_t. \tag{17.29}$$

Now if we assume that the expectations are updated by a fraction of the discrepancy between the current observation and the previous expectation, we get

$$\tilde{x}_t - \tilde{x}_{t-1} = \delta(x_t - \tilde{x}_{t-1})$$

$$\tilde{x}_t = \delta x_t - \delta\tilde{x}_{t-1} + \tilde{x}_{t-1}$$

$$= \delta x_t + (1-\delta)\tilde{x}_{t-1}$$

$$\tilde{x}_t - (1-\delta)\tilde{x}_{t-1} = \delta x_t.$$

Define $\lambda = 1 - \delta$. Then starting with the last line of the above equation,

$$\tilde{x}_t - \lambda \tilde{x}_{t-1} = \delta x_t$$

$$(1 - \lambda B)\tilde{x}_t = \delta x_t$$

$$\tilde{x}_t = \frac{\delta}{1 - \lambda B} x_t .$$

Finally, substituting this result into the model of Equation (17.29), we find out that

$$y_t = \beta_0 + \tilde{x}_t \beta_1 + e_t$$

$$= \beta_0 + \frac{\delta \beta_1}{1 - \lambda B} x_t + e_t$$

which is the same as the Koyck Scheme of Equation (17.26). As far as estimating these models, OLS is not consistent [see Equation (5.11) for a definition of consistency] if there is autocorrelated error. You can use a two-stage estimator substituting $\hat{y}_{t-1}$ for y_{t-1}. You can also use x_{t-1} as an instrument for y_{t-1}.

17.10 *Pooling Time Series and Cross Section Data*

Suppose we had for a particular sales region, call it region 1, the model

$$\mathbf{y}^{(1)} = \mathbf{X}^{(1)} \boldsymbol{\beta}^{(1)} + \mathbf{e}^{(1)}$$

where $\mathbf{y}^{(1)}$ is the $T \cdot 1$ vector of observations on the response of the market in that region and $\mathbf{X}^{(1)}$ is a matrix of marketing instruments including such factors as advertising effort, and so forth. In region 1 we might have k_1 such instruments. Data have been collected from time period 1 through time period T and analogously, in region 2, we have done the same thing but with k_2 different independent variables:

$$\mathbf{y}^{(2)} = \mathbf{X}^{(2)} \boldsymbol{\beta}^{(2)} + \mathbf{e}^{(2)}.$$

These two regression are only *seemingly unrelated* because we would expect $\text{Cov}(e_t^{(1)}, e_t^{(2)}) \neq 0$ as long as the two regions are interconnected economically. In point of fact, there are hardly two regions left on Earth that are not interconnected economically. The covariance identified above is called *contemporaneous* for obvious reasons. Alternating the regions so as to keep contemporaneous observations next to each other as we move from row to row, we could combine the two models as below:

$$\begin{bmatrix} y_1^{(1)} \\ y_1^{(2)} \\ y_2^{(1)} \\ y_2^{(2)} \\ \cdots \\ y_T^{(1)} \\ y_T^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & x_{11}^{(1)} & \cdots & x_{1k_1}^{(1)} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & x_{11}^{(2)} & \cdots & x_{1k_2}^{(2)} \\ 1 & x_{21}^{(1)} & \cdots & x_{2k_1}^{(1)} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & x_{21}^{(2)} & \cdots & x_{2k_2}^{(2)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{T1}^{(1)} & \cdots & x_{Tk_1}^{(1)} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & x_{T1}^{(2)} & \cdots & x_{Tk_2}^{(2)} \end{bmatrix} \begin{bmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \\ \cdots \\ \beta_{k_1}^{(1)} \\ \beta_0^{(2)} \\ \beta_1^{(2)} \\ \cdots \\ \beta_{k_2}^{(2)} \end{bmatrix} + \begin{bmatrix} e_1^{(1)} \\ e_1^{(2)} \\ e_2^{(1)} \\ e_2^{(2)} \\ \cdots \\ e_T^{(1)} \\ e_T^{(2)} \end{bmatrix}$$

so that our model is now simply $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$, with $\mathbf{y}$ having n times T rows, assuming n regions and T observations across time. For now we will continue to assume that n = 2. With contemporaneous covariance we can model the error covariance matrix as

$$V(\mathbf{e}) = \left[\begin{array}{c|c|c|c} \mathbf{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{\Sigma} & \cdots & \mathbf{0} \\ \hline \cdots & \cdots & \cdots & \cdots \\ \hline 0 & \mathbf{0} & \cdots & \mathbf{\Sigma} \end{array}\right] = {}_T\mathbf{I}_T \otimes {}_n\mathbf{\Sigma}_n$$

where each null matrix and each copy of $\mathbf{\Sigma}$ is $2 \cdot 2$, and the Kronecker product operator $\otimes$ is defined in Section 1.10 (and elaborated on in Section 8.6). In fact, we might note the similarity between this and the error structure for the Multivariate General Linear model in Equation (8.36) and Equation (8.37). The difference is that in the General Linear Model, all dependent variables have the same set of independent variables. In this case, the seemingly unrelated model, we will use GLS, in which the error matrix, usually notated $\mathbf{V}$, will be $\mathbf{I} \otimes \mathbf{\Sigma}$ as you can see now

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'(\mathbf{I} \otimes \mathbf{\Sigma})^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} \otimes \mathbf{\Sigma})^{-1}\mathbf{y} \tag{17.30}$$

and

$$V(\hat{\boldsymbol{\beta}}) = [\mathbf{X}'(\mathbf{I} \otimes \mathbf{\Sigma})^{-1}\mathbf{X}]^{-1}.$$

We can apply a two step procedure in which we use OLS and estimate $\mathbf{S} = \hat{\mathbf{\Sigma}}$, then we apply GLS using $\mathbf{S}$ as a substitute for $\mathbf{\Sigma}$ in Equation (17.30). One could also use Maximum Likelihood.

Another way to deal with the pooling problem is to use dummy variables, as we did in Section 7.3 in Equation (7.5) in the context of the analysis of variance. Of course, one could also use effect or orthogonal coding. If we set up dummies for each time period and each region, this purges the dependent variable of all variance associated with regions and time. This is a somewhat drastic approach, since some of the variance of interest will get thrown out with the bath water. A less drastic approach is to treat time and cross-sections like a random effect in ANOVA. Suppose that our error term is composed of

$$e_{ij} = \alpha_i + \varphi_t + \varepsilon_{it}$$

with $V(\alpha_i) = \sigma_\alpha^2$, $V(\varphi_t) = \sigma_\varphi^2$ and $V(\varepsilon_{it}) = \sigma_\varepsilon^2$. Given that the data from each region are next to each other in the $\mathbf{y}$ and $\mathbf{e}$ vectors,

$$V(\mathbf{e}) = \sigma_\alpha^2({}_T\mathbf{1}\mathbf{1}'_T \otimes {}_n\mathbf{I}_n) + \sigma_\varphi^2({}_T\mathbf{I}_T \otimes {}_n\mathbf{1}\mathbf{1}'_n) + \sigma_e^2\mathbf{I}_{nT}$$

and we can apply a first step ANOVA to estimate the random variances σ_α^2 and σ_φ^2 while in the second step we use those estimates in GLS. Parks has proposed a model like Zellner's seemingly unrelated regressions but with autoregressive error, and De Silva has suggested a model with variance components for the cross sections but autoregression for the time component.

References

Johnston, J. (1972) *Econometric Methods, Second Edition*. New York: McGraw-Hill.

Theil, Henri (1971) *Principles of Econometrics*. New York: Wiley.

Maddala, G. S. (1977) *Econometrics*. New York: McGraw-Hill.

Chapter 18: Time Series

18.1 *Stationary Data Series*

In this chapter we consider a series of observation taken from a single entity over time much as we assumed in Section 17.5. The entity generating the data might be a particular company, Web site, household, market, geographic region or anything else that maintains a fixed identity over time. Our observations look like $y_1, y_2, \cdots, y_n$ with a joint density $Pr(y_1, y_2, \cdots, y_n)$. When data are collected over time, there is a very important concept that is called *stationarity* and in fact the concept shows up in other places in this book, notably Equation (15.1). For our purposes, we define the stationarity of a time series as

$$Pr(y_t, y_{t+1}, \cdots, y_{t+k}) = Pr(y_{t+m}, y_{t+m+1}, \cdots, y_{t+m+k}), \tag{18.1}$$

for all t, j and k. Given that, it must be the case also that for $m = \pm 1, \pm 2, \cdots$

$$Pr(y_t) = Pr(y_{t+m})$$

which then further implies that

$$E(y_t) = E(y_{t+m})$$

and

$$V(y_t) = V(y_{t+m}).$$

Presumably under stationarity it is the case as well that

$$Pr(y_t, y_{t+1}) = Pr(y_{t+m}, y_{t+m+1}) \tag{18.2}$$

which would then make obvious the notion that

$$Cov(y_t, y_{t+1}) = Cov(y_{t+m}, y_{t+m+1}) = \gamma_1.$$

In general, since

$$Pr(y_t, y_{t+j}) = Pr(y_{t+m}, y_{t+m+j}) \tag{18.3}$$

the following is implied

$$Cov(y_t, y_{t+j}) = Cov(y_{t+m}, y_{t+m+j}) = \gamma_j.$$

The parameter γ_j is known as the *autocovariance* at lag j. Putting all of these results together, we can say that

$$E\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = E(\mathbf{y}) = \begin{bmatrix} \mu \\ \mu \\ \cdots \\ \mu \end{bmatrix}$$

and

$$V(\mathbf{y}) = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{bmatrix}$$

Like all covariance matrices, V(**y**) is symmetric. If $E(y_t)$ does not depend on t, which it should not with a stationary series, then we would ordinarily expect to find the series in the neighborhood of μ. History tends to repeat itself, probabilistically. By the definition of covariance [Equation (4.7)]:

$$\gamma_j = E[(y_t - \mu)(y_{t+j} - \mu)].$$

If $\gamma_j > 0$ we would expect that a higher than usual observation would be followed by another higher than usual observation. We can standardize the covariances by defining the autocorrelation,

$$\rho_j = \frac{\gamma_j}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \frac{\gamma_j}{\gamma_0}.$$

As usual, $\rho_0 = 1$. The structure of the autocorrelations will greatly help us in understating the behavior of the series, **y**.

18.2 *A Linear Model for Time Series*

The time series models that we will be covering are called *discrete linear stochastic processes* and are of the form

$$y_t = \mu + e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots . \tag{18.4}$$

In effect, an observation within the series is conceptualized as being the result of a possibly linear combination of random inputs. The e_t values are assumed identically distributed with

$$E(e_t) = 0 \text{ and}$$

$$V(e_t) = \sigma_e^2.$$

Further, we will assume that

$$Cov(e_t, e_{t+j}) = 0 \tag{18.5}$$

for all $j \neq 0$. These e_t values are independent inputs and are often called *white noise*. We also assume that

$$\sum_{i=0}^{\infty} \psi_i = c \text{ and that}$$

$$\psi_0 = 1.$$

Given the preceding long list of notation and assumptions, what is the expectation and variance of our data? As was pointed out before, it is still the case the $E(y_t) = \mu$ since we can combine Equation (18.4) and the assumption that $E(e_t) = 0$. As for the variance of $V(y_t)$,

$$V(y_t) = E(y_t - \mu)^2$$

$$= E(\mu + e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots - \mu)^2 \tag{18.6}$$

where the two μ's will just cancel. Squaring the remaining terms, we can collect them into two sets:

$$V(y_t) = E(e_t^2 + \psi_1^2 e_{t-1}^2 + \psi_2^2 e_{t-2}^e + \cdots) + E(\text{all cross terms}).$$

We can quickly dispense of all the cross terms from Equation (18.6) because, by assumption [Equation (18.5)] the e_t are independent. Worrying just about the first part of the above equation, and noting that the expectation of a sum is equal to the sum of the expectation[Equation (4.4)], we can then say that

$$V(y_t) = \sigma_e^2 \sum_{i=0}^{\infty} \psi_i^2. \tag{18.7}$$

Are you game for figuring out the covariance at lag j of two data points from the series? Here goes. We note that the covariance between y_t and y_{t-j} is $E[(y_t - \mu)(y_{t-j} - \mu)]$. Once again, all values of μ will cancel leaving us with

$$\gamma_j = E[(e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots)(e_{t-j} + \psi_1 e_{t-j-1} + \psi_2 e_{t-j-2} + \cdots)]$$

$$= E[(\psi_j e_{t-j}^2) + (\psi_{j+1}\psi_1 e_{t-j-1}^2) + (\psi_{j+2}\psi_2 e_{t-j-2}^2) + \cdots] + E(\text{all cross terms}).$$

In this case, E(all cross terms) refers to any term involving $E(e_t , e_{t-m})$ for $m \neq 0$ and once again, with independent e_t all such covariances vanish. That leaves us with the very manageable Equation (18.8)

$$\gamma_j = \sigma_e^2(\psi_j + \psi_{j+1}\psi_1 + \psi_{j+1}\psi_2 + \cdots)$$

$$= \sigma_e^2 \sum_i^{\infty} \psi_i \psi_{i+j} \tag{18.8}$$

Neither the variance in Equation (18.7) nor the covariances in Equation (18.8) can exist unless the infinite sum in those two equations is equal to a finite value. That an infinite series can be finite is seen in the reasoning that runs between Equation (15.17) and (15.17). We will return to this concept momentarily, but first we will assume that $\psi_i = \phi^i$, with $|\phi| < 1$. Then

$$y_t = \mu + e_t + \phi e_{t-1} \ \phi^2 e_{t-2} + \cdots .$$

It can be shown that

$$\sum_{i=0}^{\infty} \psi_i = \sum_{i=0}^{\infty} \phi^i = \frac{1}{1-\phi}.$$

That this is so can be seen by defining $s = \sum_{i=0}^{\infty} \phi^i = 1 + \phi + \phi^2 + \phi^3 + \cdots$, and then multiplying by ϕ so that $\phi s - s = 1$. Solving for s leads to the result, $s = 1/1-\phi$. Combining this result with Equation (18.7), y_t then has a variance of

$$\gamma_0 = \frac{\sigma_e^2}{1-\phi^2}$$

and from Equation (18.8), autocovariances of

$$\gamma_j = \frac{\sigma_e^2 \phi^j}{1-\phi^2}.$$

Needless to say, this will only work for with $|\phi| < 1$, as otherwise, the variance will blow up. If $\phi = 1$ our model becomes

$$\begin{aligned} y_t &= \mu + e_t + e_{t-1} + e_{t-2} + \cdots \\ &= \mu + e_{t-1} + e_{t-2} + \cdots + e_t \\ &= y_{t-1} + e_t \end{aligned}$$

and so forth, as we could now substitute for y_{t-1} above. Obviously, the variance of a series with $\phi = 1$ blows up.

18.3 *Moving Average Processes*

A moving average model is characterized by a finite number of non-zero values ψ_i with $\psi_i = 0$ for $i > q$. The model will then look like the following,

$$y_t = \mu + e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots + \psi_q e_{t-q}.$$

The tradition in this area calls for us to modify the notation somewhat and utilize $\theta_i = -\psi_i$ which then modifies the look of the model slightly to

$$y_t = \mu + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}.$$

Such as model is often called a *Moving Average (q) process*, or MA(q) for short. As an example, consider the MA(1):

$$y_t = \mu + e_t - \theta_1 e_{t-1}$$

which can also be written with the *Backshift operator*, symbolized with the letter B and presented also in Equation (17.23):

$$y_t = \mu + (1 - \theta_1 B)e_t,$$

i. e.

$$Be_t = e_{t-1}, \tag{18.9}$$

$$B \cdot (Be_t) = B^2 e_t = e_{t-1} \text{ and} \tag{18.10}$$

$$B^0 e_t = e_t. \tag{18.11}$$

We will have much cause to use the backshift operator in this chapter. For now, it will be interesting to look at the autocovariances of the MA(1) model. These will be

$$\gamma_1 = E[(e_t - \theta_1 e_{t-1})(e_{t-1} - \theta_1 e_{t-2})]$$

$$= \sigma_e^2(-\theta_1).$$

OK, that's a nice result. What about the autocovariance at lag 2?

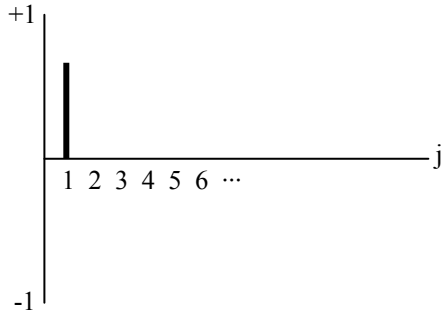
$$\gamma_2 = E[(e_t - \theta_1 e_{t-1})(e_{t-2} - \theta_1 e_{t-3})]$$

$$= 0.$$

Since none of the errors overlap with the same subscript, everything vanishes as the errors are assumed independent. Thus we note that for the MA(1),

$$\gamma_j = \begin{cases} -\sigma_e^2\theta_1 & \text{for } j = 1 \\ 0 & \text{for } j > 1 \end{cases}.$$

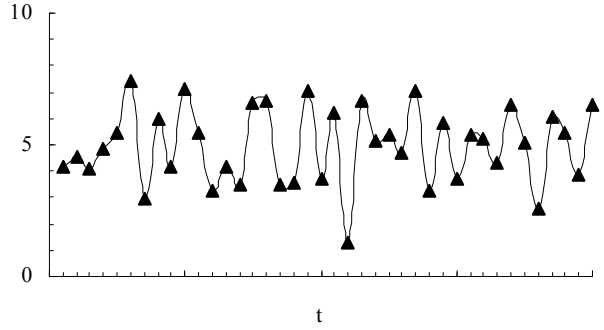
We can plot the autocorrelation function, which plots the value of the autocorrelations at various lags, j. In the case of the MA(1), the theoretical pattern is unmistakable:



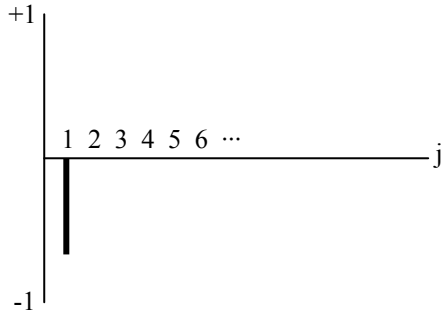
As we will see later in the chapter, the *correlogram*, as a diagram such as the one above is called, is an important mechanism to identify the underlying structure of a time series. For the sake of curiosity, it will be nice to look at a simulated MA(1) process with $\theta_1 = -.9$ and $\mu = 5$. The model would be

$$y_t = e_t + .9e_{t-1}$$

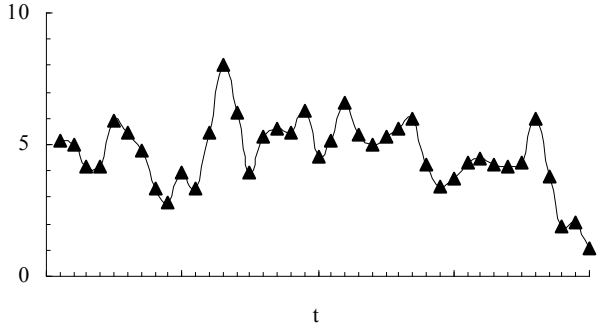
and $\sigma_e^2 = 1,\ \gamma_0 = \sigma_e^2(1+\theta_1^2) = 1.81,\ \gamma_1 = -\sigma_e^2(\theta_1) = .9,\ \rho_1 = \gamma_1/\gamma_0 = .5$ and $\rho_j = 0$ for all $j > 1$. An example of this MA(1) process, produced using a random number generator is shown below:



If $\theta_1 = +.9$ so that $\rho_1 = -.5$ the correlogram would appear as



with the spike heading off in the negative, rather than the positive direction. The plot of the time series would by more jagged, since a positive value of y_t would tend to be associated with a negative value of y_{t-1}.



For an arbitrary value of q, an MA(q) process will have autocovariances

$$\gamma_j = \begin{cases} -\sigma_e^2(-\theta_j + \theta_1\theta_{j+1} + \cdots + \theta_{q-j}\theta_q & \text{for } j = 1,\ 2,\ \cdots,\ q \\ 0 & \text{for } j > q. \end{cases}$$

For example the MA(2) process will have a correlogram that has two *spikes*:

18.4 *Autoregressive Processes*

Recall that any discrete linear stochastic process can be expressed as

$$y_t = \mu + e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots$$

as was Equation (18.4). Needless to say this implies that we can express the errors as

$$e_t = y_t - \mu - \psi_1 e_{t-1} - \psi_2 e_{t-2} - \cdots .$$

Our assumption of stationarity requires that the same basic model that holds for e_t must hold true for e_{t-1} which would then be

$$e_{t-1} = y_{t-1} - \mu - \psi_1 e_{t-2} - \psi_2 e_{t-3} - \cdots .$$

If we substitute the model for e_{t-1} into the model for y_t we get

$$y_t = \mu + e_t + \psi_1[y_{t-1} - \mu - \psi_1 e_{t-2} - \cdots] + \psi_2 e_{t-2} + \cdots$$

$$= \mu(1 - \psi_1) + e_t + \psi_1 y_{t-1} + (\psi_2 - \psi_1^2)e_{t-2} + \cdots$$

You can keep doing this - now we substitute an expression for e_{t-2} and so forth until all the e_t terms are banished and all that remains are y_t values, with various coefficients. Arbitrarily naming these coefficients with the letter π, we get something that looks like

$$y_t = \pi_1 y_{t-1} + \pi_2 y_{t-2} + \cdots + \delta + e_t. \qquad (18.12)$$

Our discrete linear stochastic process can be expressed as a possibly infinite series of past random disturbances [i. e. Equation (18.4)]. If the series is finite, we call it an MA process. Any discrete linear stochastic process can also be expressed as a possibly infinite series of its own past values disturbances [i. e. Equation (18.12)]. If the series is finite, we will call it an *autoregressive*

process, also known as an *AR* process. This is illustrated below, where we have modified Equation (18.12) by assuming that $\pi_i = 0$ for $i > p$:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \delta + e_t.$$

To the paragraph above, I would add that a finite AR is equivalent to an infinite MA and a finite MA is equivalent to an infinite AR. Below we will prove the first of these two assertions. But before we do that, it should be noted that all of this gives the data analyst a lot of flexibility in creating a parsimonious model.

The AR(1) model looks like

$$y_t = \phi_1 y_{t-1} + \delta + e_t \tag{18.13}$$

$$(1 - \phi_1 B) y_t = \delta + e_t \tag{18.14}$$

If we take Equation (18.13) and substitute the equivalent expression for y_{t-1}, we have

$$y_t = \phi_1 [\phi_1 y_{t-2} + \delta + e_{t-1}] + \delta + e_t$$

and then again

$$y_t = \phi_1 [\phi_1 ([\phi_1 y_{t-3} + \delta + e_{t-2}) + \delta + e_{t-1}] + \delta + e_t$$

and so on until we see that we end up with

$$y_t = \frac{\delta}{1-\phi_1} + e_t + \phi_1 e_{t-1} + \phi_1^2 e_{t-2} + \phi_1^3 e_{t-3} + \cdots$$

which is an infinite MA process. As claimed, an AR(1) leads to an infinite MA.

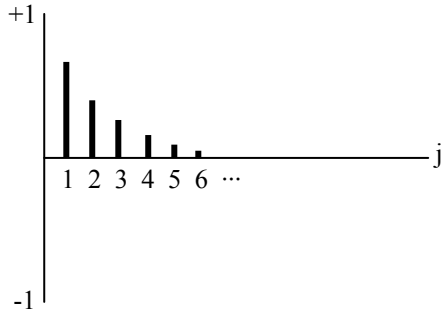
What are the moments of an AR(1) process? We have

$$E(y_t) = \frac{\delta}{1-\phi_1},$$

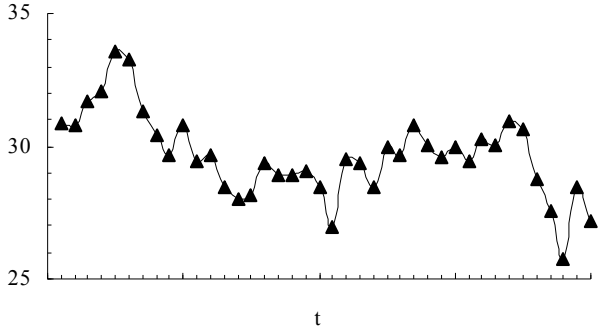
$$\gamma_j = \phi_1^j \frac{\sigma_e^2}{1-\phi_1^2} \quad \text{and}$$

$$\rho_j = \phi_1^j.$$

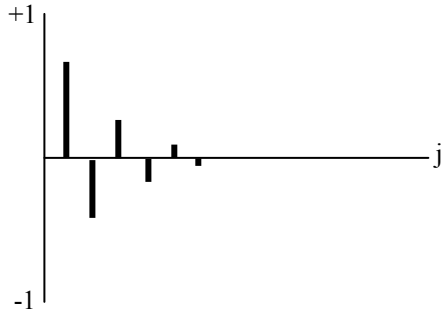
For the AR(1), the autocorrelations decline exponentially. An idealized correlogram is shown below:



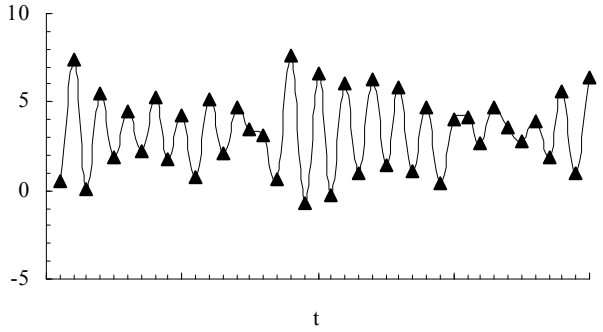
The autocorrelations damp out slowly. Next we show a random realization of the AR(1) model $y_t = .8y_{t-1} + 6 + e_t$:



Another example is identical to the first, but the sign on ϕ_2 is reversed. The correlogram appears below



and then we see a random realization of the series:



18.5 *Details of the Algebra of the Backshift Operator*

One of the most beautiful aspects of time series analysis is the use of backshift notation. Say we have an AR(1) with parameter ϕ_1. We can express the model as

$$(1 - \phi_1 B)y_t = e_t + \delta.$$

Putting the model in reduced form we have

$$y_t = (1-\phi_1 B)^{-1}\delta + (1-\phi_1 B)^{-1} e_t.$$

But what does it mean to invert a function with "B" in it? It produces an infinite series. To see that, start with the basic fact that

$$(1-\phi_1 B)^{-1} = \frac{1}{1-\phi_1 B}.$$

So far so good. However, the series

$$s = \sum_{i=0}^{\infty} \phi_1^i B^i = 1 + \phi_1 B + \phi_1^2 B^2 + \phi_1^3 B^3 + \cdots$$

and the series

$$\phi_1 B \cdot s = \phi_1 B + \phi_1^2 B^2 + \phi_1^3 B^3 + \cdots$$

differ by 1. Thus

$$s - \phi_1 B \cdot s = 1$$

and therefore

$$\begin{aligned} s = (1-\phi_1 B)^{-1} &= \frac{1}{1-\phi_1 B} \\ &= \sum_{i=0}^{\infty} \phi_1^i B^i = 1 + \phi_1 B + \phi_1^2 B^2 + \phi_1^3 B^3 + \cdots. \end{aligned} \tag{18.15}$$

Stationarity, and the need to avoid infinities in the infinite sum, require that

$$|\phi_1| < 1. \tag{18.16}$$

This is equivalent to saying that the root of that $1 - \phi_1 B = 0$ must lie outside the unit circle.

18.6 *The AR(2) Process*

The AR(2) model is

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \delta + e_t$$

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} = \delta + e_t$$

$$(1 - \phi_1 B - \phi_2 B^2) y_t = \delta + e_t$$

which is stationary if the roots of

$$1 - \phi_1 B - \phi_2 B^2 = 0$$

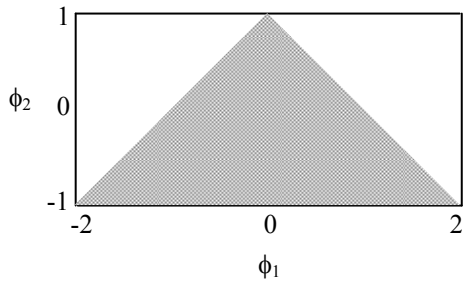
lie outside the unit circle, which is to say

$$\phi_1 + \phi_2 < 1, \tag{18.17}$$

$$\phi_1 - \phi_2 < 1 \text{ and} \tag{18.18}$$

$$|\phi_2| < 1. \tag{18.19}$$

Below, the graph shows the permissible region as a shaded triangle:



18.7 *The General AR(p) Process*

In general, an AR model of order p can be expressed as

$$(1-\phi_1 B-\phi_2 B^2-\phi_p B^p)y_t=\delta+e_t$$

$$\phi(B)y_t=\delta+e_t$$

$$y_t=\frac{\delta+e_t}{\phi(B)}.$$

Note that here we have introduced a new way of writing 1 - ϕ_1B - ϕ_2B^2 - ⋯ - ϕ_pB^p, namely to call it simply ϕ(B). The autocorrelations and the ϕ_i are related to each other via what are known as the *Yule-Walker Equations*:

$$\rho_1=\phi_1+\phi_2\rho_1+\cdots\phi_p\rho_{p-1}$$

$$\rho_2=\phi_1\rho_1+\phi_2+\cdots\phi_p\rho_{p-2}$$

$$\cdots=\cdots$$

$$\rho_p=\phi_1\rho_{p-1}+\phi_2\rho_{p-2}+\cdots\phi_p$$

which can be used to estimate $\hat{\phi}_j$ values.

18.8 *The ARMA(1,1) Mixed Process*

Consider the model

$$y_t=\delta+\phi_1 y_{t-1}+e_t-\theta_1 e_{t-1}$$

$$(1-\phi_1 B)y_t=\delta+(1-\theta_1 B)e_t.$$

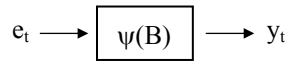
Here we have both an autoregressive and a moving average component. The AR part results in an infinite MA model with

$$y_t=\frac{\delta}{1-\phi_1 B}+\frac{1-\theta_1 B}{1-\phi_1 B}e_t.$$

In compact notation we can say that ψ(B) = ϕ^{-1}(B) · θ(B). The MA part results in an infinite AR model with

$$\frac{1-\phi_1 B}{1-\theta_1 B}y_t-\frac{\delta}{1-\theta_1 B}=e_t.$$

Again we can compactify the notation noting that $\pi(B) = \phi(B) \cdot \theta^{-1}(B)$. Mixed models let you achieve parsimony as you can represent an infinite MA with a finite AR and vice versa. The situation that we have at hand can be graphed as follows:



We conceptualize of our observed series of data as being driven by are series of random shocks, of random values or white noise inputs. These inputs are then passed through a filter with various properties and that eventually leads to an output, which consists of our data. Modeling the data requires that we come up with a parsimonious description, one with few model parameters, of the filter, i.e $\psi(B)$.

What stationarity is to the AR side, invertibility is to the MA side. Invertibility requires that the roots of

$$1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q = 0$$

lie outside of the unit circle.

18.9 *The ARIMA(1,1,1) Model*

A series may be relatively homogeneous, looking pretty much the same at all time periods, but it may end up being non-stationary simply because it shows no permanent affinity for a particular level or mean. Even though the original series of data may not be stationary, differences between successive observations may be stationary:

$$d_t = y_t - y_{t-1} = (1 - B)y_t.$$

Simply put, we can apply an ARMA model to the d_t. When we do so, this is called an ARIMA model with the middle I referring to the fact that it is integrated. If the first differences are not stationary, the second differences might be, i. e.

$$d'_t = d_t - d_{t-1} = (1 - B)(1 - B)y_t.$$

The ARIMA(1,1,1) process, with the middle number referring to the number of differences that are taken can be described as

$$d_t = \phi_1 d_{t-1} - \theta_1 e_{t-1} + e_t$$

$$y_t - y_{t-1} = \phi_1 (y_t - y_{t-1}) - \theta e_{t-1} + e_t$$

$$y_t = (1 + \phi_1) y_{t-1} - \phi_1 y_{t-2} - \theta_1 e_{t-1} + e_t.$$

Thus we see that the ARIMA(1,1,1) is an ARMA(2,1) where the first ARMA AR parameter is equal to $1 + \phi_1$ while the second ARMA(2,1) AR parameter is $-\phi_1$. These parameters violate the rules for stationarity in Equations (18.17), (18.18) and (18.19). Similarly, an ARIMA(0,1,1) process looks like

$$y_t = y_{t-1} + e_t - \theta_1 e_{t-1}$$

which violates the stationarity rule for an AR(1) [Equation (18.16)] right off the top since "ϕ_1" = 1!

Thus we see the importance of differencing the series first, if necessary, prior to fitting an ARMA model.

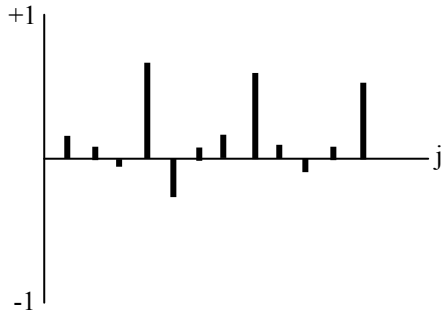
We can wrap up this section with another brief note about the backshift notation and the ARIMA(1,1,1) model. Such a model can be written quite elegantly as

$$(1 - \phi_1 B)(1 - B)y_t = \delta + (1 - \theta_1 B)e_t$$

In the model, the constant term δ implies that the average change will have the same sign as δ and the series will drift in the direction of the sign of δ.

18.10 *Seasonality*

Differencing, AR or MA parameters may be needed at various lags. For quarterly data, you may need to look at lags of 4, or for monthly data, lags of 12, which may occur whenever there are yearly patterns in data. For example, the following pattern seen in quarterly data:



may require that you difference the data at a lag of 4, i.e analyze $d_t = (1 - B^4)y^t$.

18.11 *Identifying ARIMA(p,d,q) Models*

In addition to the cues afforded in the autocorrelations, we can also look at what are known as the *partial autocorrelations*. For each lag j, you estimate ρ_j controlling for the first j - 1 values ρ_{j-1}, ρ_{j-2}, ⋯, ρ_1.

For a nonstationary process, the autocorrelations will be large at very long lags. On the other hand, over-differencing tends to produce an MA(1) with $\theta_1 = 1$.

For an AR process, the autocorrelations will decline exponentially. The partial autocorrelations will exhibit significant spikes at the first p lags.

For an MA process, the autocorrelations will exhibit significant spikes at the first q lags. The partial autocorrelations will exhibit exponential decline.

For a mixed process, the autocorrelations as well as the partial autocorrelations will decline exponentially.

It is generally a good idea to run the error of your model through the same diagnostic process to make sure that it is indeed acting like white noise. In effect, one adds a term to the ARIMA model, and then looks at the error to see if it is white noise yet. The process is repeated until the error is completely whitened.

References

Box, George E. P., Gwilym M. Jenkins (1976) *Time Series Analysis. Revised Edition*. Oakland, CA: Holden-Day.

Appendices

A. The Greek Alphabet

alpha	α	A
beta	β	B
gamma	γ	Γ
delta	δ	Δ
epsilon	ε	E
zeta	ζ	Z
eta	η	H
theta	θ	Θ
iota	ι	I
kappa	κ	K
lambda	λ	Λ
mu	μ	M
nu	ν	N
xi or ksi	ξ	Ξ
omicron	ο	O
phi	π	Π
rho	ρ	P
sigma	σ	Σ
tau	τ	T
upsilon	υ	Y
phi	φ	Φ
chi	χ	X
psi	ψ	Ψ
omega	ω	Ω

Index