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:

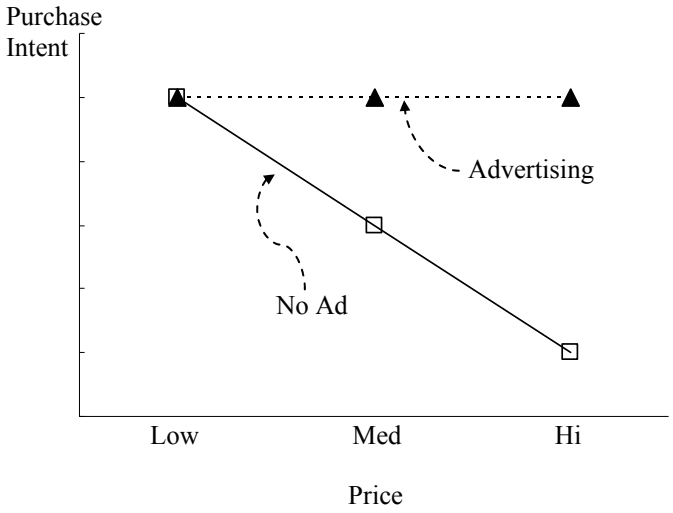
$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

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

$x_{\cdot 4} = x_{\cdot 1} x_{\cdot 2}$

$x_{\cdot 5} = x_{\cdot 1} x_{\cdot 3}$

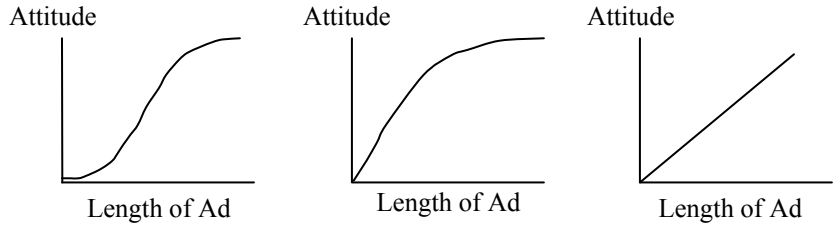
The subscripts on the dependent variable values run from L to M to H (low, medium and high) to index the level of the price variable and from A to N to indicate advertising vs. no-advertising. Column 0 of the **X** matrix codes for the usual intercept term. Column 1 uses orthogonal coding to register 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.