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

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

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

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:

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

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

Each $\hat{y}_i$ is formed as the linear combination $\mathbf{x}'_{i.}\boldsymbol{\beta}$, 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(**e**) = **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(**e**) ≠ **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: $\frac{\min f}{\boldsymbol{\beta}}$ 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 ∂f(x)/∂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?

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

Here we have relied on the identity $\hat{\mathbf{y}} \equiv \mathrm{E}(\mathbf{y})$ 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 **X** matrix as constant. Strictly speaking, the fact that **X** is fixed implies we cannot generalize beyond the values in **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 $$\mathrm{E}(\hat{\beta}_i) = \beta_i . \qquad (5.10)$$

Consistent $$\Pr(\hat{\beta}_i - \beta_i \le \varepsilon) \to 1 \text{ as n } \to \infty . \qquad (5.11)$$

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

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

In effect what is going on with consistency is that as n → ∞, $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$. 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 $$\Pr(\mathbf{y} \mid \hat{\boldsymbol{\beta}}) \text{ does not depend on } \boldsymbol{\beta} \qquad (5.12)$$

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 **e** vector will play an important role in these assumptions. Both **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 **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 **e** is normally distributed with a mean equal to the null vector, and with a variance matrix **Σ**. Since **e** is n by 1, its mean must be n by 1, and the variances and covariances among the n elements of **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, **Σ**. 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 **e** is normal, this series of assumptions is often called *NIID*, that is to say we are asserting that **e** is <u>n</u>ormally, <u>i</u>dentically and <u>i</u>ndependently <u>d</u>istributed.

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

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

where $\mathbf{x}'_{i.}$ is the ith row of **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 $E(y_i) \equiv \hat{y}_i = x'_{i.}\boldsymbol{\beta}$.

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,

$$\begin{aligned} \ell = \Pr(\mathbf{y}) &= \prod_i^n \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-(y_i - x'_{i.}\boldsymbol{\beta})^2 / 2\sigma^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\sum_i^n (y_i - x'_{i.}\boldsymbol{\beta})^2 / 2\sigma^2\right] \qquad (5.18) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) / 2\sigma^2\right]. \end{aligned}$$

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

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

with derivative

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

If we take $\partial\mathrm{L}/\partial\boldsymbol{\beta} = 0$, 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 $\mathrm{SS}_{\mathrm{Error}}$ or SSE. Now

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

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

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

so that therefore

$$\mathrm{SS}_{\mathrm{Error}} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathrm{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\mathrm{SS}_{\mathrm{Error}} = \mathrm{SS}_{\mathrm{Total}} - \mathrm{SS}_{\mathrm{Predictable}} \tag{5.22}$$

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 $\mathrm{SS}_{\mathrm{Total}}$ can be seen as the sum of the $\mathrm{SS}_{\mathrm{Error}}$ + $\mathrm{SS}_{\mathrm{Predictable}}$.

There are many ways of expressing the $\mathrm{SS}_{\mathrm{Predictable}}$, 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 $\mathbf{X'X(X'X)}^{-1}$.

Taking the last version of the $SS_{\text{Predictable}}$ 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 $SS_{\text{Predictable}}$ is the sum of the squares of the predictions of the model, the $\hat{y}_{i\cdot}$. 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

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

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

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

Note that the last line above involves two equivalent versions of $SS_{\text{Predictable}}$, 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, **S** [see Equation (2.12)]. The **S** matrix is partitioned (Section 1.4) into sections corresponding to the y variable and the x's:

$${}_k\mathbf{S}_k = \left[\begin{array}{c:c} s_{yy} & \mathbf{s}'_{xy} \\ \hdashline \mathbf{s}_{xy} & \mathbf{S}_{xx} \end{array}\right]. \tag{5.24}$$

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 **β** 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 **β 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:

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

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

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

$$
\begin{aligned}
(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} &= (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} \\
&= \mathrm{n}^{-1}\mathbf{1}'\mathbf{y} \\
&= [\mathrm{n}^{-1} \quad \mathrm{n}^{-1} \quad \cdots \quad \mathrm{n}^{-1}]\,\mathbf{y} \\
&= \frac{1}{\mathrm{n}}\sum_{\mathrm{i}}^{\mathrm{n}} \mathrm{y}_{\mathrm{i}}
\end{aligned}
$$

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

$$
{}_{\mathrm{n}}\hat{\mathrm{y}}_1 = {}_{\mathrm{n}}1_1\,\bar{\mathrm{y}}
$$

$$
\begin{bmatrix} \hat{\mathrm{y}}_1 \\ \hat{\mathrm{y}}_2 \\ \cdots \\ \hat{\mathrm{y}}_{\mathrm{n}} \end{bmatrix} = \begin{bmatrix} \bar{\mathrm{y}} \\ \bar{\mathrm{y}} \\ \cdots \\ \bar{\mathrm{y}} \end{bmatrix}.
$$

The matrix **P** is given by the expression

$$
\begin{aligned}
\mathbf{P} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= {}_{\mathrm{n}}\mathbf{1}_1(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' \\
&= \begin{bmatrix} \frac{1}{\mathrm{n}} & \frac{1}{\mathrm{n}} & \cdots & \frac{1}{\mathrm{n}} \\ \frac{1}{\mathrm{n}} & \frac{1}{\mathrm{n}} & \cdots & \frac{1}{\mathrm{n}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\mathrm{n}} & \frac{1}{\mathrm{n}} & \cdots & \frac{1}{\mathrm{n}} \end{bmatrix}
\end{aligned}
$$

so in that case the predicted values of **y** are

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

and the Sum of Squares Predictable are

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

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,

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

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

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

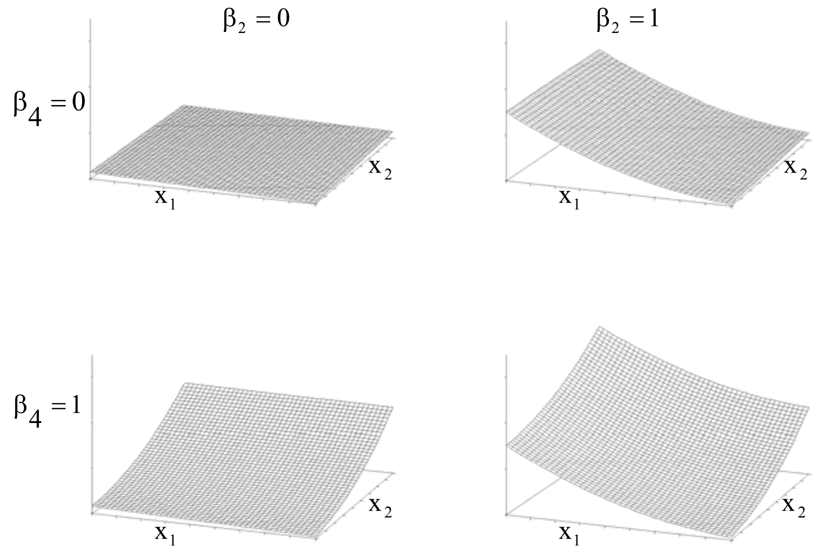
which the reader will recognize as the scalar, the corrected sum of squares from Equation (2.11).

5.10 *Response Surface Models*

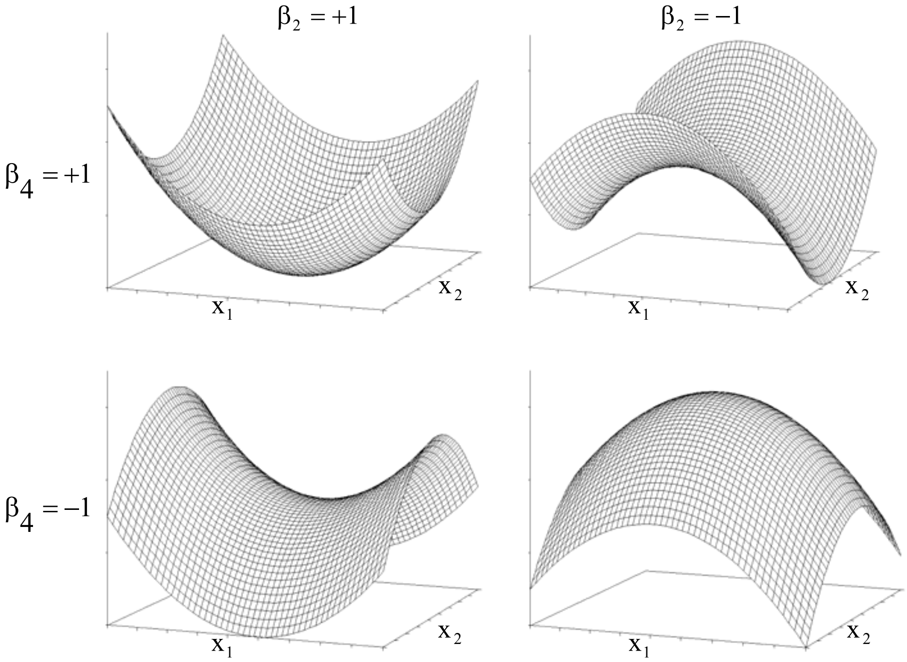
While it is known as the linear model, one can fit more complicated curves than lines or planes. It is relatively straightforward to include quadratic or higher order polynomials in a regression model, merely by squaring or cubing one of the independent variables (it is wise to mean center first). For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