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.