Type I and Type III Sums of Squares

Supplement to Section 8.3

Brian Habing – University of South Carolina
Last Updated: February 4, 2003

PROC REG, PROC GLM, and PROC INSIGHT all calculate three types of F tests:

- The *omnibus test*: The omnibus test is the test that is found in the ANOVA table. Its F statistic is found by dividing the Sum of Squares for the Model (the SSR in the case of regression) by the SSE. It is called *the Test for the Model* by the text and is discussed on pages 355-356.
- The *Type III Tests*: The Type III tests are the ones that the text calls the *Tests for Individual Coefficients* and describes on page 357. The p-values and F statistics for these tests are found in the box labeled Type III Sum of Squares on the output.
- The *Type I Tests*: The Type I tests are also called the *Sequential Tests*. They are discussed briefly on page 372-374.

The following code uses PROC GLM to analyze the data in Table 8.2 (pg. 346-347) and to produce the output discussed on pages 353-358. Notice that even though there are 59 observations in the data set, seven of them have missing values so there are only 52 observations used for the multiple regression.

```
DATA fw08x02;
INPUT Obs    age    bed    bath     size      lot       price;
CARDS;
  1     21     3      3.0    0.951     64.904     30.000
  2     21     3      2.0    1.036    217.800     39.900
  3      7     1      1.0    0.676     54.450     46.500
  <snip rest of data as found on the book’s companion web-site>
 58      1     3      2.0    2.510       .       189.500
 59     33     3      4.0    3.627     17.760    199.000
;

PROC GLM DATA=fw08x02;
MODEL price = age bed bath size lot;
RUN;
```

With five independent variables, this procedure produces seventeen p-values for testing hypotheses.

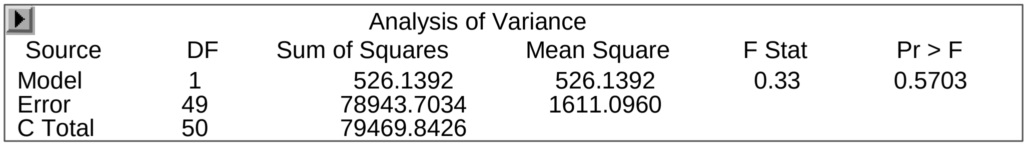
A – This is the test associated with the ANOVA table. It always (even for cases that aren’t regression) tests the null hypothesis that none of the independent variables linearly predict the dependent variable. The alternate hypothesis is that at least one of the independent variables does linearly predict the dependent variables. Because it tests all of these at once it is sometimes called the *omnibus test*. For this example it is testing the null hypothesis H_0: all of $\beta_{age}=0$, $\beta_{bed}=0$, $\beta_{bath}=0$, $\beta_{size}=0$ and $\beta_{lot}=0$.
The alternate hypothesis is that at least one of them is not zero. In this example, the p-value is less than 0.0001 and we would reject the null hypothesis. (You can check these SS, MS, and F values with what the text gives on page 356.)

```
                                       The GLM Procedure

Dependent Variable: price

                                                Sum of
       Source                      DF          Squares     Mean Square    F Value    Pr > F
       Model                        5      65695.79292     13139.15858      42.93    <.0001      ◄A
       Error                       45      13774.04972       306.08999
       Corrected Total             50      79469.84265

                       R-Square     Coeff Var      Root MSE     price Mean
                       0.826676      15.98770      17.49543       109.4305

       Source                      DF        Type I SS     Mean Square    F Value    Pr > F
       age                          1        526.13923       526.13923       1.72    0.1965   ◄B
       bed                          1      10713.09272     10713.09272      35.00    <.0001   ◄C
       bath                         1      20049.55143     20049.55143      65.50    <.0001   ◄D
       size                         1      33939.41003     33939.41003     110.88    <.0001   ◄E
       lot                          1        467.59951       467.59951       1.53    0.2229   ◄F

       Source                      DF      Type III SS     Mean Square    F Value    Pr > F
       age                          1        946.23859       946.23859       3.09    0.0855   ◄G
       bed                          1       1972.60916      1972.60916       6.44    0.0147   ◄H
       bath                         1        156.83009       156.83009       0.51    0.4778   ◄I
       size                         1      32739.75553     32739.75553     106.96    <.0001   ◄J
       lot                          1        467.59951       467.59951       1.53    0.2229   ◄K

                                                   Standard
                Parameter         Estimate            Error    t Value    Pr > |t|
                Intercept      35.28792164      14.11712107       2.50      0.0161   ◄L
                age            -0.34980453       0.19895262      -1.76      0.0855   ◄M
                bed           -11.23820158       4.42691173      -2.54      0.0147   ◄N
                bath           -4.54015206       6.34278904      -0.72      0.4778   ◄O
                size           65.94646658       6.37644152      10.34      <.0001   ◄P
                lot             0.06205081       0.05020362       1.24      0.2229   ◄Q
```

B through F – These are the Type I, or sequential, tests. The line with each of the p-values always begins with which independent variable is being tested, given that those listed above it are already included in the model. The p-value of less than 0.1965 (for $F_{df=(1,45)}$=1.72) on line B thus tests the hypotheses

H_0: β_{age}=0 given that no other independent variables are included in the model
vs. H_A: β_{age}≠0 given that no other independent variables are included in the model

This result is similar to, but not exactly the same as what you would get if you performed a simple linear regression to predict price from age (shown here).



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	1	526.1392	526.1392	0.33	0.5703
Error	49	78943.7034	1611.0960		
C Total	50	79469.8426			

Notice that the Sum of Squares for the Model in this simple linear regression is identical to the Type I Sum of Squares for the variable Age. Also note that the TSS and total degrees of freedom are the same. The reason for the difference in the F statistics and the p-values comes because the Type I test has taken all of the other variables errors out of the SSE. That is, the Type I test uses SSE and df from the ANOVA table for the entire multiple regression.

Now, the p-value of less than 0.0001 (for $F_{df=(1,45)}$=35.00) on line C takes into account that it appears on the line below *height*. It thus tests the hypotheses

H_0: β_{bed}=0 given that age is included in the model
vs. H_A: $\beta_{bed}\neq 0$ given that age is included in the model

We could continue in this way down to line F. This p-value of 0.2229 (for $F_{df=(1,45)}$=1.53) tests the hypotheses

H_0: β_{lot}=0 given that age, bed, bath, size, and lot are included in the model
vs. H_A: $\beta_{lot}\neq 0$ age, bed, bath, size, and lot are included in the model

By line B, we would thus conclude that age (by itself) is not linearly related to weight. By line C, we would conclude that bed is linearly related to weight after accounting for age, etc... If we wanted to reverse the order in which these independent variables were considered, we simply could have listed water before height in the model line.

It is important to note that when using a Type I test that it makes no sense to decide to keep something in the model if you don't also include everything listed above it. Because of this it is best to list the variables for the model in order from what you think should be most important to what you think should be least important.

Also notice the relationship between the Type I tests and the omnibus test. Notice that the sum of square on line B-F add up to the SSR on line A. It can be shown algebraically that the Type I sum of squares will always add up to the sum of squares on the model line.

G through K - These are the Type III tests. Just like the Type I tests, each line always begins with which independent variable is being tested. In this case however each of these tests is done given that all of the other variables are included in the model. The p-value of 0.0855 (for $F_{df=(1,45)}$=6.44) on line G thus tests the hypotheses

H_0: β_{age}=0 given that bed, bath, size, and lot are included in the model
vs. H_A: $\beta_{age}\neq 0$ given that bed, bath, size, and lot are included in the model

Similarly, the p-value of less than 0.0147 (for $F_{df=(1,45)}$=6.44) on line H tests the hypotheses:

H_0: β_{bed}=0 given that age, bath, size and lot are included in the model
vs. H_A: $\beta_{bed}\neq 0$ given that age, bath, size and lot are included in the model

By line G, at α=0.05, we would thus conclude that age is not a significant predictor of price if bed, bath, size, and lot are already included in the model. By line H, we would conclude bed still serves as a significant predictor even after age, bath, size and lot are already included in the model.

It is important to note that anything found significant by a Type III test is adding to the ability of the multiple regression model to predict the independent variable. Say you find that several variables are not statistically significant however. In this case it is important to realize that you found that they were not significant given that all of the others were included in the model. You can thus only safely get rid of one of those variables (maybe the others become significant then if that one is not included!)

For a Type III test the sum of squares will generally not add up to the SSR (it only happens if the independent variables are all orthogonal to each other). However, the last of the Type III tests will always equal the last of the Type I tests. (Re-read the hypotheses being tested to see that this is so.)

L through Q– Lines L through Q are simply the t-tests discussed on pages 358-359. Notice for lines G and M that $\sqrt{3.09} = 1.76$ and that the p-values are equal. Recalling that $(t_{df=n})^2 = F_{df=(1,n),}$ it is no surprise that these t-tests are exactly the same as the Type III F-tests. The reason for having SAS output this section is because it also gives the estimates of the regression coefficients and their standard errors. (Compare these lines of output to page 367 for example). Line L is the test of the hypothesis that the intercept is zero.