“EXTRA" Sums of Squares

These are values of SS attributable to adding some variable(s) to a model which already has some variable(s) in the model. Take for example the model,

$$Y_i = b_0 + b_k X_{ki} + b_l X_{li} + b_m X_{mi} + e_i$$

where we may be interested in the contribution of a single variable to the model (eg X_m).

We wish to get $SSRX_m$ given that X_k and X_l are already in the model. This measures the IMPROVEMENT in the model from entering another variable or variables (X_m) in this case.

This is EXACTLY the SSDifference one would get with the General Linear Test if the Model with X_m is viewed as the Full model and the model without X_m is taken as the Reduced model.

The textbook refers to this SS as the “EXTRA" Sums of Squares. This is a measure of the improvement in the model (or reduction in the Error) by one or more variables given that some other variable or variables are already in the model.

for example, the Full model is

$$Y_i = b_0 + b_k X_{ki} + b_l X_{li} + b_m X_{mi} + b_n X_{ni} + e_i$$

the reduced model is (any combination of X variables can be removed)

$$Y_i = b_0 + b_k X_{ki} + b_l X_{li} + e_i$$

The difference between these models is the EXTRA SS, and is calculated as

$$SSR(X_m, X_n | X_k, X_l) = SSR(X_k, X_l, X_m, X_n) - SSR(X_k, X_l) = SSE(X_k, X_l) - SSE(X_k, X_l, X_m, X_n)$$

I will continue using problem 7.20 as an example in class. You can follow a similar discussion of the material in your textbook, where they use Table 8.1. To assist in this endeavor, I have provided SAS output for this example as well.

Think about the ways the SSRegression can be partitioned between the various variables in a model. For the mathematician salary example

First look at the SSRegression for the various combinations of variables. These values are simply the SSReg for the fitted models.

SSCorrected Total = 689.26

$SSR(X_1) = 306.73233$	$SSE = 382.52767$
$SSR(X_2) = 508.06883$	$SSE = 181.19117$
$SSR(X_3) = 214.76157$	$SSE = 474.49843$
$SSR(X_1,X_2) = 570.52677$	$SSE = 118.73323$
$SSR(X_1,X_3) = 397.19152$	$SSE = 292.06848$
$SSR(X_2,X_3) = 593.39849$	$SSE = 95.86151$
$SSR(X_1,X_2,X_3) = 627.817$	$SSE = 61.44300$

Now look at the gain from entering each X_k after each of the others has been entered in the model.

These values are calculated as $SSR(X_k|X_l) = SSR(X_k,X_l) - SSR(X_l)$

$SSR(X_2|X_1) = 263.79444$
$SSR(X_3|X_1) = 90.45919$
$SSR(X_1|X_2) = 62.45794$
$SSR(X_3|X_2) = 85.32966$
$SSR(X_1|X_3) = 182.42995$
$SSR(X_2|X_3) = 378.63692$

Now look at the gain from entering each X_k after all of the others has been entered in the model.

These values are calculated as $SSR(X_k|X_l,X_m) = SSR(X_k,X_l,X_m) - SSR(X_k,X_l)$

$SSR(X_1|X_2,X_3) = 34.41851$
$SSR(X_2|X_1,X_3) = 230.62548$
$SSR(X_3|X_1,X_2) = 57.29023$

Which of all of of these possible EXTRA SS are most likely to be of interest?

This will depend primarily on the hypotheses to be tested.

Also, how can we set up models to most easily estimate these with some computer package?

SAS provides two types of SS calculations to estimated EXTRA SS. SAS terms these TYPE I and TYPE II, TYPE III or TYPE IV (the last 3 are the same in regression, but not in Design).

TYPE I SS gives the contribution to the model of each variable entered in order. If we enter three variables X_k, X_l and X_m in this order then we would get TYPE I SS which give

$SSR(X_k)$ =

$SSR(X_l|X_k)$ =

$SSR(X_m|X_k,X_l)$ =

These SS are also referred to as the Sequentially adjusted SS, and are entirely order dependent. If order is changed, the results will be different.

These can be obtained from PROC REG with the SS1 option on the MODEL statement. These are given by default by PROC GLM along with SS3.

PROC GLM will also give SS2 and SS4 upon request, while PROC REG gives only SS1 and SS2.

Output from PROC REG

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.84693	2.00188	8.92	<.0001
X1	1	1.10313	0.32957	3.35	0.0032
X2	1	0.32152	0.03711	8.66	<.0001
X3	1	1.28894	0.29848	4.32	0.0003

Variable	DF	Type I SS	Type II SS
Intercept	1	37446	244.17168
X1	1	306.73233	34.41851
X2	1	263.79445	230.62548
X3	1	57.29022	57.29022

The Hypotheses of interest generally include one or more of the following

For the model

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

Common hypotheses to test

1) Test all β_k jointly

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
H_1: some $\beta_k \neq 0$

This test we have seen. It is a test of the whole model. The necessary SS is the SSR(X_1,X_2,X_3) tested with the SSE for the full model. This test is provided by most software.

2) Test some subset of β_k, jointly

H_0: $\beta_2 = \beta_3 = 0$

3) Test individual β_k

H_0: $\beta_2 = 0$

4) Test relationships between 2 or more β_k

H_0: $\beta_2 = \beta_3$

1) Test all β_k jointly

$$H_o:\ \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

$$H_1:\ \beta_j \neq 0 \text{ for at least one j}$$

The test is done as

$$F = \frac{MSReg(b_1, b_2, \ldots, b_k)}{MSE}$$

A different approach to this same test is given as

a) Full Model: $Y = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$

$$SSE_{Full} = Y'Y - SSReg_{(b_0, b_1, \ldots, b_k)}$$

b) Reduced Model: $Y = \beta_0$

$$SSE_{Reduced} = Y'Y - n\bar{Y}^2$$

The test statistic is now - expressions using SSErrors

$$F_o = \frac{\dfrac{SSE_{Red} - SSE_{Full}}{{}^{2}\text{ parameters in full - }{}^{2}\text{ parameters in reduced}}}{\dfrac{SSE_{Full}}{dfFull}}$$

or expressions using SSRegressions

$$F_o = \frac{\dfrac{SSReg_{(b_0, b_1, \ldots, b_k)} - n\bar{Y}^2}{k+1-1}}{\dfrac{SSE_{Full}}{n-k-1}}$$

where; n = number of observations
k = number of parameters
“1" is for the intercept (β_0)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	627.8169964	209.2723321	68.12	<.0001
Error	20	61.4430036	3.0721502		
Corrected Total	23	689.2600000			

Utility of TYPE I SS

1) Some special types of models do have an order. eg. Polynomials and some cases of analysis of covariance. We will see both of these later.

2) This is also convenient to use to set up the General Linear Test, the reduced model variables are entered first, and the SSDifference is given by the sum of the remaining variables.

For example, suppose we wish to test to see of several β's jointly = 0

H_0: $\beta_2 = \beta_3 = 0$

H_0: both β_2 and β_3 do not = 0

The SS for the reduced model will be $SSR(X_1)$

The SS for the full model will be $SSR(X_1,X_2,X_3)$

The SS for the difference will be $SSR(X_2,X_3|X_1)$

The SS for the difference can be calculated as

$SSR(X_2,X_3|X_1) = SSR(X_1,X_2,X_3) - SSR(X_1)$

or as $SSR(X_2|X_1) + SSR(X_3|X_1,X_2)$

Test of a subset of coefficients

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

versus

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

The hypotheses are

$$H_o: \ \beta_2 = \beta_3 = 0$$

versus

H_1: at least one of either β_2 or β_3 does not equal 0

The test is done as

$$F_o = \frac{\dfrac{SSE_{Red} - SSE_{Full}}{2 \text{ parameters in full} - 2 \text{ parameters in reduced}}}{MSE_{Full}}$$

The computations are

$$SSE_{Full} = Y'Y - b'X'Y \quad \text{where,} \quad X = (1, X_1, X_2, X_3)$$

$$b = (X'X)^{-1}X'Y$$

$$SSE_{Red} = Y'Y - b_R' X_R' Y \qquad \text{where,} \quad X_R = (1, X_1)$$

$$b_R = (X'X)^{-1}X_R' Y$$

$$F_o = \frac{\dfrac{b'X'Y - b_R' X_R' Y}{\text{difference in number of parameters} = \gamma}}{MSE_{Full}}$$

If $F_o \geq F_{\alpha, \gamma, \text{n-p}}$ then we reject H_o; otherwise accept

If the variables to be tested are entered last in the model, all of the necessary SS are available in SAS from a single run of PROC GLM or PROC REG with the SS1 option.

Output from PROC GLM with tests of TYPE I SS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	627.8169964	209.2723321	68.12	<.0001
Error	20	61.4430036	3.0721502		
Corrected Total	23	689.2600000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	306.7323285	306.7323285	99.84	<.0001
X2	1	263.7944455	263.7944455	85.87	<.0001
X3	1	57.2902224	57.2902224	18.65	0.0003

To test β_2 and β_3 jointly, we use

Full Model = $SSR(X_1,X_2,X_3) = 627.8169964$

Reduced model is given by "X1" in the TYPE I SS = $SSR(X_1) = 306.7323285$

The difference can be calculated as either

Full Model − Reduced model = $SSR(X_1,X_2,X_3) - SSR(X_1)$

$= 627.8169964 - 306.7323285 = 321.08467$

or as the sum of "X2" and "X3" from the TYPE I SS

$SSR(X_2|X_1) + SSR(X_3|X_1,X_2) = 263.7944455 + 57.2902224 = 321.08467$

Then, for either case, the F test for the General Linear Test is

$$F = \frac{MSDifference}{MSError} = \frac{\frac{321.08467}{2}}{3.0721502} = 52.25732305$$

The Tabular F values are

$F_{a=0.05:1,20} = 4.35$ $\mathbf{F_{a=0.05:2,20} = 3.49}$ $F_{a=0.05:3,20} = 3.10$

TYPE II SS gives the contribution to the model of each variable entered after the model has been adjusted for all other variables.

$SSR(X_l|X_m,X_n)$ =

$SSR(X_n|X_l,X_m)$ =

$SSR(X_m|X_l,X_n)$ =

These SS are also referred to as the Partial SS or the fully adjusted SS. These SS are NOT order dependent. If order is changed, the results will not change

Utility of TYPE II SS

1) These are the statistics generally examined to determine the “unique" contribution of each variable.

The test done by this type of SS is the test of individual β_k , given that all other variables are already in the model.

H_0: $\beta_2 = 0$

H_1: $\beta_2 \neq 0$

This can also be tested by the t-test of the regression coefficient. These two tests (ie. F test of TYPE II SS and t-test of β_k are identical).

NOTE that this implies that the b_k are fully adjusted. The b_k in Multiple Regression are also referred to as the Partial Regression Coefficient.

Test of individual coefficients - β_j against an hypothesized β_{jo}

$$t_o = \frac{b_j - \beta_{jo}}{\sqrt{\hat{\sigma}^2 c_{jj}}} \sim t_{(n-p)}$$

Special tests

H_0: $\beta_1 = \beta_3$
H_1: $\beta_1 \neq \beta_3$

This can be tested with the full model

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

$$Y = \beta_0 + \beta_1 X_{2i} + \beta_2 (X_{1i} + X_{3i}) + \epsilon_i$$

To test this, create a variable X1X2SUM=X1+X2; , and fit this as a reduced model.

The original estimates were β_1=**1.10313** and β_3=**1.28894**.

The hypotheses are H_0: $\beta_1 = \beta_3$ H_1: $\beta_1 \neq \beta_3$

The variable X1X2Sum was created as = X1+X3; and run as the reduced model.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	627.38353	313.69176	106.46	<.0001
Error	21	61.87647	2.94650		
Corrected Total	23	689.26000			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.89290	1.95684	9.14	<.0001
X2	1	0.31865	0.03556	8.96	<.0001
X2X3SUM	1	1.20345	0.18912	6.36	<.0001

The full model error was **3.0721502** with 20 df

The F test is

$$F = \frac{\text{MSDifference}}{\text{MSError}} = \frac{61.87647 - 61.4430036}{3.0721502} = \frac{0.4334664}{3.0721502} = 0.1410954$$

$$\mathbf{F_{a=0.05:1,20} = 4.35} \; F_{a=0.05:2,20} = 3.49 \quad F_{a=0.05:3,20} = 3.10$$

See the test results for the second SAS technique below.

This test is facilitated in PROC REG several ways;

Using the "RESTRICT x1=x3;" statement the results are:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	627.38353	313.69176	106.46	<.0001
Error	21	61.87647	2.94650		
Corrected Total	23	689.26000			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.89290	1.95684	9.14	<.0001
X1	1	1.20345	0.18912	6.36	<.0001
X2	1	0.31865	0.03556	8.96	<.0001
X3	1	1.20345	0.18912	6.36	<.0001
RESTRICT	-1	-2.33286	6.08223	-0.38	0.7111*

Using the "TEST x1=x3;" statement the results are:

The results of the test were

Test 1 Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.43347	0.14	0.7111
Denominator	20	3.07215		

NOTE that the F value is the same. The difference is not significant, implying that the reduced model is as good as the full model, so we can conclude that the data is consistent with the Null hypothesis, H_0: $\beta_1 = \beta_3$

NOTE: that one results in the full model ANOVA and testing with the full model error, the other the reduced model ANOVA. Full model error is still used in testing.