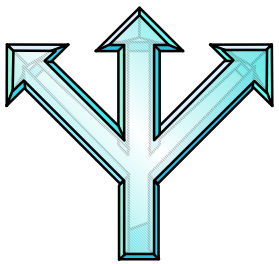
Statistical Techniques II

EXST7015

Extra SS



Multiple regression

- **Multiple regression involves two or more independent variables (X_i), but still only a single dependent variable (Y_i).**
  - **There is an analysis for multiple dependent variables, it is called multivariate regression.**
- **The sample equation is;**
  - $Y_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + e_i$

Multiple regression *(continued)*

- **The objectives in multiple regression are generally the same as SLR.**
  - **Testing hypotheses about potential relationships (using correlations),**
  - **fitting and documenting relationships, and**
  - **estimating parameters with confidence intervals.**

Multiple regression *(continued)*

- **The good news, most of what we know about simple linear regressions applies to multiple regression.**
  - **The regressions equation is similar.**
  - **The assumptions for the regression are the same as for Simple Linear Regression**
  - **The interpretation of the parameter estimates are the same (units are Y units per X units, and measure the change in Y for a 1 unit change in X).**

Multiple regression *(continued)*

- **The diagnostics used in simple linear regression are mostly the same for multiple regression.**
  - **Residuals can still be examined for outliers, homogeneity, normality, curvature, influence, etc., as with SLR.**
  - **The only difference is that, since we have several X's, we would usually plot the residuals on Yhat instead of a single X variable.**

Multiple regression *(continued)*

- **So what is different?**
  - **Obviously, the calculations are more complicated. Algebraic equations basically do not exist. Matrix algebra must be used.**
  - **Also, we now have several independent variables, X_1, X_2, X_3, etc. We will need some mechanism to evaluate these individually. To this end, we will discuss a new type of Sum of Squares not needed for simple linear regression.**

Multiple regression *(continued)*

- **The use of several independent variables also creates some new problems. If the independent variable are highly correlated we have a problem called multicollinearity. We will need some diagnostics to evaluate this problem.**
- **Outside of the new diagnostics needed to deal with several independent variables, SLR and Multiple regression are very similar.**

Multiple regression *(continued)*

- **To do multiple reg. in SAS we specify a model with the variables of interest.**
- **For example, a regression on Y with 3 variables X1, X2 and X3 would be specified as**
  - **PROC REG;**
  - **MODEL Y = X1 X2 X3;**
- **To get the SS1 and SS2 we add the options /ss1 ss2;**

Extra Sum of Squares

- **When a variable is added to a model, it usually accounts for some variation.**
- **In rare circumstances the variable will account for zero sum of squares. This is rare in practice.**

Extra Sum of Squares *(continued)*

- **Most often, when a variable is added to a model it reduces the error sum of squares (SSE) and increases the model sum of squares (SSReg). The sum of squares that a variable causes to be removed from the error SS and added to the model SS is called its "ExtraSS".**

Extra Sum of Squares *(continued)*

- **If each variable had its unique ExtraSS, the concept would be simple. However, two variables in a model are rarely wholly independent. Two variables in a model may "compete" for SS, so that if one enters the model first it get SS that a second variable could have taken had it entered first.**

Extra Sum of Squares *(continued)*

- **Or, one variable may actually enhance another, so that the second variable may actually account for more SS after the first variable is entered than if it had entered first.**
- **As a result we cannot talk simply about the ExtraSS for a variable. We have to consider the ExtraSS in the context of what variables were already in the model.**

Extra Sum of Squares *(continued)*

- **Extra SS notation.**
  - **The ExtraSS will simply be denoted SSX_i for each variable X_i.**
  - **Since we must consider which variables are already in the model we will add to this notation a designation of what is already in the model. For example, the extra SS for X_2, given that X_1 and X_3 are already in the model will be $SSX_2 \mid X_1, X_3$.**

Extra Sum of Squares *(continued)*

- **We could also add a designation for the intercept (usually X_0). So the extra SS mentioned earlier could be $SSX_2 \mid X_0, X_1, X_3$.**
- **However, since ALL models we will discuss are adjusted for the intercept first, we will usually leave off the X_0.**
- **Using this notation, a simple linear regression Extra SS for X would be just SSX. Adjustment for X_0 is assumed.**

Extra SS Example 1

- **The first example is with a two factor multiple regression. The data is given below.**
- **We will fit the simple linear regressions for each variable,**
- **and then the two factor multiple regression to calculate the extra ss.**
- **To do this in SAS we simply list the variables in the PROC REG model statement, `MODEL Y = X1 X2;`**

Extra SS Example 1 *(continued)*

Obs	Y	X1	X2
1	18	3	4
2	22	1	8
3	34	2	11
4	36	6	5
5	42	8	1
6	54	6	5
7	68	7	10
8	77	4	10
9	87	9	11
10	92	6	8

Extra SS Example 1 *(continued)*

- **Simple linear regression for Y on X1.**

Analysis of Variance

Source	DF	Sum of Squares	Mean Square
Model	1	2381.31494	2381.31494
Error	8	4054.68506	506.83563
C Total	9	6436.00000	

F Value	Prob>F
4.698	0.0621

Extra SS Example 1 *(continued)*

- **Note the slope is not significantly different from zero. Now Y on X2.**

Analysis of Variance

Source	DF	Sum of Squares	Mean Square
Model	1	1446.14793	1446.14793
Error	8	4989.85207	623.73151
C Total	9	6436.00000	

F Value	Prob>F
2.319	0.1663

Extra SS Example 1 *(continued)*

- **The slope for X2 also is not significantly different from zero either.**
- **Now we fit the two factor multiple regression of Y on both X1 and X2, and we will compare the results.**
-

Extra SS Example 1 *(continued)*

- Analysis of Variance
-

Source	DF	Sum of Squares	Mean Square
Model	2	4523.43473	2261.71736
Error	7	1912.56527	273.22361
C Total	9	6436.00000	

-

F Value	Prob>F
8.278	0.0143

- **Note that together the variables are significant.**

Extra SS Example 1 *(continued)*

- **Now calculate the Extra SS. Since the SSReg increases by the same amount that the SSError decreases for each variable, either SS can be used. I will use the Error SS.**
- **Error SS for X1 alone 4054.68506**
- **Error SS for X2 alone 4989.85207**
- **Error SS for X1 and X2 1912.56527**
- **SSTotal (all models) 6436.00000**

Extra Sum of Squares *(continued)*

- **The first variable alone (X1) had an error SS equal to 4054.68506 out of the corrected total of 6436. The difference is 2381.31494, and this is the ExtraSS for X1, the amount of the total accounted for by the variable.**
  - **SSX1 = 6436 - 4054.68506 = 2381.31494**
- **Likewise, the ExtraSS for X2 alone is**
  - **SSX2 = 6436 - 4989.85207 = 1446.14793**

Extra Sum of Squares *(continued)*

- **SSX1 = 6436 - 4054.68506 = 2381.31494**
- **SSX2 = 6436 - 4989.85207 = 1446.14793**
- **Now, how did they do together in the multiple regression?**

Extra Sum of Squares *(continued)*

- **The two variables together had an error term of 1912.56527.**
- **Since the first variable alone had an error of 4054.68506, the second must have reduced the model by an additional amount equal to the difference.**
- **This is an additional amount of SSX2|X1 = 4054.68506 - 1912.56527 = 2142.119793.**

Extra Sum of Squares *(continued)*

- **So, SSX2|X1 = 4054.68506 - 1912.56527 = 2142.119793.**
- **And likewise, the ExtraSS for X1 is SSX1|X2 = 4989.85207 - 1912.56527 = 3077.286793.**

Extra Sum of Squares *(continued)*

- **In summary,**
  - **SSX1 = 2381.31494**
  - **SSX2 = 1446.14793**
  - **SSX1|X2 =3077.286793**
  - **SSX2|X1 = 2142.119793**
- **Note that in this case the variables actually enhanced each other, performing better together than alone. Although not the rule, this can happen.**

3 factor model

- **The calculation of extra SS is exactly the same for larger models. The following example is a 3 factor multiple regression.**
- **In SAS this model would be**
  - `PROC REG; MODEL Y = X1 X2 X3;`
-
- **The raw data is given below.**
-

3 factor model *(continued)*

Obs	Y	X1	X2	X3
1	1	2	9	2
2	3	4	6	5
3	5	7	7	9
4	3	3	5	5
5	6	5	8	9
6	4	3	4	2
7	2	2	3	6
8	8	6	2	1
9	9	7	5	3
10	3	8	2	4
11	5	7	3	7
12	6	9	1	4

3 factor model *(continued)*

- The results of the regressions are;
- The SSTotal is 62.91667 for all models
- For the 1 factor models the results are;
  - Regression of Y on X1
    - SSError = 38.939, SSModel = 23.978
  - Regression of Y on X2
    - SSError = 58.801, SSModel = 4.115
  - Regression of Y on X3
    - SSError = 62.680, SSModel = 0.237

3 factor model *(continued)*

- **The extra SS are equal to the model SS for 1 factor models.**
- **SSX1 = 23.978 (or SSX1|X0)**
- **SSX2 = 4.115 (or SSX2|X0)**
- **SSX3 = 0.237 (or SSX3|X0)**
- **These SS are adjusted for the intercept (correction factor). This will always be the case for our examples, so the X0 is often omitted.**

3 factor model *(continued)*

- **Fitting X1 and X2 and X3 together TWO AT A TIME we get the following results.**
- **Regression of Y on X1 and X2**
  - **SSError = 38.842, SSModel = 24.074**
- **Regression of Y on X1 and X3**
  - **SSError = 37.546, SSModel = 25.371**
- **Regression of Y on X2 and X3**
  - **SSError = 58.779, SSModel = 4.137**

3 factor model *(continued)*

- **Now lets get the Extra SS for variables fitted together.**
  - **SSX1 alone = 23.978**
  - **SSX2 alone = 4.115**
  - **SSX1 and X2 together = 24.074**
- **Again, look for the improvement in the model due to the second variable. Calculate how much each variable adds to a model with the other variable already in the model.**

3 factor model *(continued)*

- **For the model with X1 and X2, subtract the amount accounted for by each variable alone from the amount together.**
  - **Start with X1 (SSX1=23.978), add X2 (SSX1,X2 = 24.074). The improvement with X2 is 24.074-23.978 = 0.096 = SSX2|X1**
  - **Start with X2 (SSX2=4.115), Add X1 (SSX1,X2=24.074). The improvement is 24.074-4.115 = 19.959 = SSX1|X2**
- **So, SSX1|X2=19.959, and SSX2|X1=0.096**

3 factor model *(continued)*

- **Likewise for the model with X1 and X3.**
  - **Start with X1 (SSX1=23.978), Add X3 (SSX1,X3 = 25.371). The improvement is 25.371-23.978 = 1.393 = SSX3|X1**
  - **Start with X3 (SSX3=0.237), Add X1 (SSX1,X3 = 25.371). The improvement is 25.371-0.237 = 25.134 = SSX1|X3**
- **So, SSX1|X3=25.134, and SSX3|X1=1.393**

3 factor model *(continued)*

- **Likewise for the model with X2 and X3.**
  - **Start with X2 (SSX2=4.115), Add X3 (SSX2,X3 = 4.137). The improvement is 4.137-4.115 = 0.022 = SSX3|X2**
  - **Start with X3 (SSX3=0.237), Add X2 (SSX2,X3 = 4.137). The improvement is 4.137-0.237 = 3.900 = SSX2|X3**
- **So, SSX2|X3=3.900, and SSX3|X2=0.022**

3 factor model *(continued)*

- **Finally, and most important (?), for all 3 variables in the model. How much does each variable improve the model over a model with the other two variables present in the model?**
  - **Start with the full model, SSX1,X2,X3=26.190**
  - **SSX1,X2 = 24.074, So SSX3|X1,X2=2.116**
  - **SSX1,X3 = 25.371, So SSX2|X1,X3=0.819**
  - **SSX2,X3 = 4.137, So SSX1|X2,X3=22.053**

3 factor model *(continued)*

- **Summarizing the Extra SS calculations.**

Extra SS	SS	d.f. Error	Error SS
SSX1	23.978	10	38.939
SSX2	4.115	10	58.802
SSX3	0.237	10	62.680
SSX1|X2	19.959	9	38.843
SSX2|X1	0.096	9	38.843
SSX1|X3	25.134	9	37.546
SSX3|X1	1.393	9	37.546
SSX2|X3	3.900	9	58.780
SSX3|X2	0.022	9	58.780

3 factor model *(continued)*

- **Summarizing the Extra SS calculations. Note that all Extra SS are also corrected for the intercept.**

Extra SS	SS	d.f. Error	Error SS
SSX1|X2,X3	22.053	8	36.727
SSX2|X1,X3	0.819	8	36.727
SSX3|X1,X2	2.116	8	36.727

More extra SS

- **A final note on extra SS. It is also useful to be able to express SS for two or more variables with two or more degrees of freedom as extra SS. For example, the SS due to X1 and X2 together (adjusted only for the intercept) is SSX1, X2. These extraSS can be obtained directly from the two factor models fitted in SAS.**
-

More extra SS *(continued)*

- **Another possibility is the two variable SS fitted after one or more other variables. For example, the SS for X1 and X2 adjusted for X3 (and of course the intercept) is SSX1, X2 | X3.**
- **To calculate this we start with the full model (SSX1,X2,X3=26.190). We know X3 alone fits SSX3 = 0.237.**
- **So SSX1,X2 | X3 = 26.190 - 0.237 = 25.953**

More extra SS *(continued)*

These extra SS for the example just mentioned are given in the table below.

Extra SS	SS	d.f. Error	Error SS
SSX1, X2	24.074	9	38.843
SSX1, X3	25.371	9	37.546
SSX2, X3	4.137	9	58.780
SSX1,X2|X3	25.953	8	36.727
SSX1,X3|X2	22.075	8	36.727
SSX2,X3|X1	2.212	8	36.727
SSX1,X2,X3	26.190	8	36.727

Type I SS

- So, what is important here? Or why do we need extra SS?
- SAS will provide us with two types of sum of squares. We need to understand both, and extra SS is the key to this understanding.
- The first one is the SAS type 1 SS, the second is SAS type 2 or 3 or 4 (which are the same for regression analysis).

Type I SS *(continued)*

- **The SAS Type 1 SS are called the sequentially adjusted SS. They have a number of potential problems.**
  - **These SS are adjusted in a sequential or serial fashion. Each SS is adjusted for the variables previously entered in the model, but not for variables entered later. For the model [Y = X1 X2 X3], X1 would be first and adjusted for nothing else (except X0). X2 would enter second, be adjusted for X1, but not for X3. X3 enters last and is adjusted for both X1 and X2.**

Type I SS *(continued)*

- **The result,**
  - **SSX1**
  - **SSX2|X1**
  - **SSX3|X1,X2.**
- **The SAS Type 1 SS are called the sequentially adjusted SS. They have a number of potential problems.**

Type I SS *(continued)*

- **Type I SS (continued)**
  - **Unfortunately, these SS are different depending on the order of the variables, so different researchers could get different results for the same data.**
  - **Use of this SS type is rare, it is only used where there is a mathematical reason to place the variables in a particular order.**
  - **Use is restricted mostly to polynomial regressions (which we will later) and a few other applications we will discuss.**

Type I SS *(continued)*

- **Type I SS (continued)**
  - **For the model Y = X1 X2 X3 in SAS the Type I SS are**
  - **SSX1, SSX2|X1, SSX3|X1,X2**
- **Other orders would give different SS and different results.**
- **So, we will not usually use Type I. However, they are provided by default by SAS PROC GLM.**

Type II SS

- **The SS Type II SS (or type III or IV for regression) are called PARTIAL SS, or fully adjusted SS, or uniquely attributable SS. These are the ones most often used.**
- **From the "fully adjusted" terminology you might guess that we are talking about each variable fitted after the other variables.**
- **This is correct.**

Type II SS *(continued)*

- **Note that in SAS, for regression, Type II and TYPE III and TYPE IV are the same. SAS provides TYPE II in PROC REG and it provides TYPE I and TYPE III by default in PROC GLM.**
- **Testing and evaluation of variables is usually done with the TYPE II or TYPE III SS.**

Type II SS *(continued)*

- **ANOVA table for our example, using the TYPE III SS (Partial SS). Note: tabular $F_{0.05,1,8}=5.32$.**

Source	d.f.	SS	MS	F value
SSX1|X2,X3	1	22.053	22.053	4.804
SSX2|X1,X3	1	0.819	0.819	0.178
SSX3|X1,X2	1	2.116	2.116	0.461
ERROR	8	36.727	4.591	

Type II SS *(continued)*

- **Since SSX2|X1,X3 and SSX3|X1,X2 are not significant we might want to remove them.**
- **However, since they are fully adjusted for each other we don't know how the SS might change when we remove one variable.**
- **So we remove variables ONE AT A TIME and check the remaining variables.**

Type II SS *(continued)*

- **ANOVA table for analysis of the variables X1 and X3 alone. ($F_{0.05,1,9}$=5.117).**
- **Note that X1 is now significant, but X3 is not and may be removed.**

Source	d.f.	SS	MS	F value
SSX1|X3	1	25.134	25.134	6.024
SSX3|X1	1	1.393	1.393	0.334
ERROR	9	37.546	4.172	

Evaluation of Multiple Regression *(continued)*

- The variable X1 is still significant. ($F_{0.05,1,10}$=4.965)
- This one at a time variable removal process is called "backward stepwise regression".

Source	d.f.	SS	MS	F value
SSX1	1	23.977	23.977	6.158
ERROR	10	38.939	3.894	

Summary

- **The primary new aspect of multiple regression (compared to SLR) is the need to evaluate and interpret several independent variables. A major tool for this are the SS produced for each variable.**
- **There are two types of SS for regression, Sequential and Partial. Extra SS are needed to understand the difference between these two types of SS.**

Summary (continued)

- **Extra SS are simply the SS that each variable accounts for, and causes to be removed from the SSError and placed in the SSModel or SSReg.**
- **It is necessary, however, to know which variables, if any, have been entered in the model in advance of the variable being examined.**

Summary (continued)

- **Also note a curious behavior of the variables when they occur together.**
  - **When one X_i is adjusted for another independent (X) variable, sometimes it's SS are larger, and sometimes smaller. This is unpredictable and can go either way.**
  - **For example. SSX1 was 23.978, but dropped to 19.959 when adjusted for X2 and increased to 25.134 when adjusted for X3. It dropped to 22.053 when adjusted for both (see X1 next page).**

Summary (continued)

Extra SS	SS
SSX1	23.978
SSX2	4.115
SSX3	0.237
SSX1|X2	19.959
SSX2|X1	0.096
SSX1|X3	25.134
SSX3|X1	1.393
SSX2|X3	3.900
SSX3|X2	0.022
SSX1|X2,X3	22.053
SSX2|X1,X3	0.819
SSX3|X1,X2	2.116

Summary (continued)

- **Not only will the SS of one variable increase or decrease as other variables are added to the model, but the regression coefficient values will also change. They may even change sign, and hence interpretation. So variables in combination do not necessarily have the same interpretation as they might have alone, though the interpretation does not usually change.**

Summary

- **Multiple regression shares a lot in interpretation and diagnostics with SLR.**
- **The coefficients should be adjusted for each other. This is the Type III SS in SAS. This is the big and important difference from SLR.**