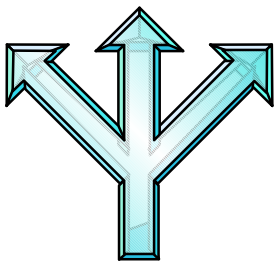
Statistical Techniques II

EXST7015

Multiple Regression (Part 1)



Multiple regression

- **The population equation is;**
  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$
- **The sample equation is;**
  - $Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + e_i$
- **Always remember that our estimates of the b_i are sample estimates of the true population values.**

Multiple regression *(continued)*

- **The objectives in multiple regression are generally the same as SLR.**
  - **Testing hypotheses (about β_i values, predicted values, correlations),**
  - **quantifying relationships (but NOT <u>proving</u> that there is a relationship)**
  - **estimating parameters with confidence intervals.**

Multiple regression *(continued)*

- **The assumptions for the regression are the same as for Simple Linear Regression**
  - **Normality**
  - **Independence**
  - **Homogeneity of variance**
  - **X_i measured without error**
  - **in short: ε_i ~ NIDr.v.(0,σ^2). Do not use this expression in an exam unless you can explain how it relates to the assumptions.**

Multiple regression *(continued)*

- **The interpretation of the parameter estimates are the same as simple linear regression**
  - **For the slope, the units are Y units per X units, and measure the change in Y for a 1 unit change in X).**
  - **For the intercept the units are Y units.**

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.**

Interpretation

- From our discussion of Extra SS you may recall that SAS will provide several types of SS.
- The first is called SS Type I, or the Sequential SS.
  - SSX1
  - SSX2|X1
  - SSX3|X1,X2

Interpretation *(continued)*

- **There will be some specific instances where these are desirable.**
- **However, we will usually want the variables adjusted for each other. All variables adjusted for all other variables in the model.**

Interpretation *(continued)*

- **This is desirable because when we adjust for other variables we**
  - **account for the effect of the other variables, or we**
  - **remove the effect of the other variables, or we**
  - **hold the other variables constant.**

Interpretation *(continued)*

- **So we know that in multiple regression each variable may have an effect on the dependent variable, and we want to isolate the effect of each variable while adjusting for the effect of other variables on the dependent variable (Y_i).**
- **The Type III SS do this, while the Type I SS adjust in a particular order (WHICH IS NOT UNIQUE!!)**

Interpretation *(continued)*

- **Note that the Type II or Type III SS (these are the same for regression) are also called the PARTIAL SS. The may also be referred to as the fully adjusted SS or the uniquely attributable SS.**
  - **SSX1| X2, X3**
  - **SSX2| X1, X3**
  - **SSX3| X1, X2**

Interpretation *(continued)*

- **So we will generally use the Partial SS. Remember the word PARTIAL.**
- **What about other things in regression, are they sequentially adjusted or fully adjusted? The regressions coefficients for example. Or correlations that may be calculated between the Y_i and the various X_i.**

Interpretation *(continued)*

- **The regression coefficients in a multiple regression are called the partial regression coefficients, and we will see partial correlation coefficients. The word partial suggests that these are FULLY ADJUSTED,**
- **and this is true.**

Numerical examples

- **We will look at two examples, a three factor (plus intercept) regression and a nine factor multiple regression.**
- **We will see most diagnostics with both examples.**
-

Example 1

- **Snedecor and Cochran (1967)**
- **Three types of soil phosphorus levels were determined, and the amount of phosphorus available to plants was determined. We want to do a regression that determines which of the soil measurement relate (correlate) to the plant available phosphorus.**

Example 1 *(continued)*

- **The 4 variables in the data set are;**
  - **Plant available phosphorus, the dependent variable.**
  - **Inorganic phosphorus, the first independent variable (order is not important if Type II SS are used).**
  - **Organic phosphorus hydrolyzed in hypobromite, another independent variable.**
  - **Organic phosphorus NOT hydrolyzed in hypobromite, also a independent variable.**

Example 1 *(continued)*

- **The SAS program.**
- **PROC REG is used for this problem. There were a number of new options used.**

```
31   PROC REG DATA=ONE ALL LINEPRINTER;
     TITLE2 'PROC REG OUTPUT WITH ALL OPTIONS';
32      MODEL Y = X1 X2 X3 / INFLUENCE;
33        TEST X1=2; TEST X1=X2=X3;
34   RUN;    OPTIONS PS=60 LS=120;
35      MODEL Y = X1 X2 X3 / PARTIAL;
36      PLOT RESIDUAL.*PREDICTED. / VREF=0;
37   RUN;
38   OPTIONS LS=80;
```

Example 1 *(continued)*

- **The "ALL" option produces a host of output, but not everything. The "INFLUENCE" and "PARTIAL" are also needed for some the output we will look at.**
- **The lineprinter option causes graphics output to be done with text characters in the output (and not high resolution graphics).**
- **Also note that there are two tests statements requested.**

Example 1 *(continued)*

- **OUTPUT**
- **The first output is the "Descriptive Statistics". For each variable (including the intercept, represented by a column of ones in the X matrix) this output gives some basic summary statistics including the , Sum, Mean, Uncorrected SS, Variance, and Std Deviation. There is no information here not available elsewhere (proc means or proc univariate).**

Example 1 *(continued)*

- **The second section is the "Uncorrected Sums of squares and Crossproducts". This section is completely redundant with the SS and CP matrix we already plan to discuss. We will ignore this section.**

Example 1 *(continued)*

- **The third section gives the simple correlations between the various independent and dependent variables. Is has some utility in examining for multicollinearity and will be discussed later.**

Example 1 *(continued)*

- **The fourth section contains the "Model Crossproducts X'X X'Y Y'Y" information we talked about in our discussion of matrix algebra. You are responsible for knowing what is in this section (but not how to derive this section with matrix algebra).**

Example 1 *(continued)*

- **The three matrices are contained in a small array with X'X a 4 by 4 matrix in the upper left, X'Y a 4 by 1 matrix on the upper right, and Y'Y is the scalar value (a 1 by 1 matrix) in the lower right corner. The remaining 1 by 3 matrix in the lower left is the (X'Y)'.**

X'X (4x4)	X'Y (4x1)
(X'Y)' (1x4)	Y'Y (1x1)

Example 1 *(continued)*

- **Calculated values in the resulting 5 by 5 matrix are given below.**

	Intercept	X1	X2	X3	Y
Intercept	n	ΣX_1	ΣX_2	ΣX_3	ΣY
X1	ΣX_1	ΣX^2_1	$\Sigma X_1 X_2$	$\Sigma X_1 X_3$	$\Sigma X_1 Y$
X2	ΣX_2	$\Sigma X_1 X_2$	ΣX^2_2	$\Sigma X_2 X_3$	$\Sigma X_2 Y$
X3	ΣX_3	$\Sigma X_1 X_3$	$\Sigma X_2 X_3$	ΣX^2_3	$\Sigma X_3 Y$
Y	ΣY	$\Sigma X_1 Y$	$\Sigma X_2 Y$	$\Sigma X_3 Y$	ΣY^2

Example 1 *(continued)*

- **We will occasionally refer to the X matrix and the X'X matrix as the semester progresses. The X matrix is the matrix of the various independent values (X_i). The X'X contains all of the SS and cross products of those variables.**
- **The X matrix has p columns and the X'X is a pxp square matrix.**
- **Note the symmetry of the off diagonal elements.**

Example 1 *(continued)*

- **The position of the elements of the "X'X Inverse, Parameter Estimates, and SSE" are as follows.**

$(X'X)^{-1}$ (4x4)	B (4x1)
B (1x4)	SSE (1x1)

Example 1 *(continued)*

- **When the X'X matrix is inverted we get the elements of the $(X'X)^{-1}$, the regression coefficients and the SSError. We will not go into detail on this. We will see the SSError and regression coefficients elsewhere. We are interested in the $(X'X)^{-1}$ because when multiplied by the MSE it gives the Variance- Covariance matrix. We will see this matrix later.**

Example 1 *(continued)*

- **Next in the output is the "Analysis of Variance" table.**

Source	DF	Sum of Squares	Mean Square
Model	3	6806.11145	2268.70382
Error	14	5583.49966	398.82140
C Total	17	12389.61111	

F Value	Prob>F
5.689	0.0092

Example 1 *(continued)*

- **Note that in multiple regression we have a 3 d.f. test in the ANOVA table.**
- **This is a test of H_o: $\beta_1 = \beta_2 = \beta_3 = 0$, or a joint test of H_o: $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$, so it is a 3 d.f. test.**
- **This test is usually not very interesting, because we included 3 variables and are interested in examining them individually. We may consider this 3 *a priori* test of interest.**

Example 1 *(continued)*

- **In this case we note that the test is highly significant, which suggests that there is some correlation between the dependent and independent variables.**
- **We also note that the MSE = 398.82140, and that it has 14 d.f.. These values will be used in many of the calculations that follow.**

Example 1 *(continued)*

- **Degrees of freedom (d.f.) in multiple regression.**
  - **The model will have p-1 d. f., where p is the number of parameters including the intercept.**
  - **The corrected total has n-1 d.f., where n is the number of observations.**
  - **The error has n-p d.f.**

Example 1 *(continued)*

- **Another section of some minor interest are the summary statistics below the ANOVA table.**
  - **Root MSE 19.97051 R-square 0.5493**
  - **Dep Mean 81.27778 Adj R-sq 0.4528**
  - **C.V. 24.57069**
- **Note from the R^2 that we are accounting for 55% of the variability. Is this good or bad?**

Example 1 *(continued)*

- **The R^2 value is now called the coefficient of multiple determination (instead of the coefficient of determination).**
- **Its square root (r) is a correlation coefficient. It is the correlation between the observed and predicted values of Y_i.**

Example 1 *(continued)*

- **Note that the next two sections of output are the "Parameter estimate" sections. This section is greatly expanded from the SLR discussed previously. It includes a number of new diagnostics that are needed to interpret and compare the various independent variables.**

Example 1 *(continued)*

- **The first columns in the "Parameter estimate" section are somewhat familiar. They include the actual estimates, standard errors (used for t-test and confidence intervals) and a t-test against the value zero (H_0: $\beta_i = 0$).**

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob>|T|
INTERCEP	1	43.652198	18.01021075	2.424	0.0295
X1	1	1.784780	0.53769551	3.319	0.0051
X2	1	-0.083397	0.41770557	-0.200	0.8446
X3	1	0.161133	0.11166524	1.443	0.1710

Example 1 *(continued)*

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob>|T|
INTERCEP	1	43.652198	18.01021075	2.424	0.0295
X1	1	1.784780	0.53769551	3.319	0.0051
X2	1	-0.083397	0.41770557	-0.200	0.8446
X3	1	0.161133	0.11166524	1.443	0.1710

- From this section we can get our model,
  - $\hat{Y} = 43.7 + 1.78X_1 - 0.0834X_2 + 0.161X_3$
- We also have our first tests of the variables (against an hypothesized value of zero). We see that only X_1 and the intercept are different from zero.

Example 1 *(continued)*

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob>|T|
INTERCEP	1	43.652198	18.01021075	2.424	0.0295
X1	1	1.784780	0.53769551	3.319	0.0051
X2	1	-0.083397	0.41770557	-0.200	0.8446
X3	1	0.161133	0.11166524	1.443	0.1710

- **This is a common calculation an often a prime objective of the regression.**
- **By default SAS does the tests of each regression coefficient against zero. However, is a test of a parameter against a value other than zero. This can be done with the SAS test statement (later) or by hand.**

Example 1 *(continued)*

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob>|T|
INTERCEP	1	43.652198	18.01021075	2.424	0.0295
X1	1	1.784780	0.53769551	3.319	0.0051
X2	1	-0.083397	0.41770557	-0.200	0.8446
X3	1	0.161133	0.11166524	1.443	0.1710

- **The hypothesis tested is;**
- **H_0: Parm est = Hypothesized value and is calculated as**
  - **t = (Parm est - Hypothesized value) / stderror. For example, test H_0: $\beta_1 = 0$**
  - **t = (1.78 - 0)/ 0.538 = 3.319 > 2.145**
  - **So reject H_0:, conclude results are NOT consistent with the null hypothesis.**

Example 1 *(continued)*

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob>|T|
INTERCEP	1	43.652198	18.01021075	2.424	0.0295
X1	1	1.784780	0.53769551	3.319	0.0051
X2	1	-0.083397	0.41770557	-0.200	0.8446
X3	1	0.161133	0.11166524	1.443	0.1710

- **A confidence interval, say on β_1, would be calculated as (using α value of 0.05)**
  - **parm est(b_1) ± $t_{(\alpha/2,\ 14\ d.f.)}$*std error**
  - **= 1.785 ± 2.145*0.5377**
  - **P(0.63154 ≤ β_1 ≤ 2.93802) = 0.95**

Example 1 *(continued)*

- **Note that confidence intervals are available from SAS PROC REG**
- **and that the alpha value can be specified**
- **I did not specify "CLB" here, but it is covered in the "ALL" option.**

Parameter Estimates

Variable	DF	Squared Semi-partial Corr Type II	Squared Partial Corr Type II	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	1	.	.	.	0	5.02414	82.28026
X1	1	0.35466	0.44040	0.78692	1.27077	0.63154	2.93802
X2	1	0.00128	0.00284	0.72432	1.38060	-0.97929	0.81249
X3	1	0.06703	0.12947	0.89915	1.11216	-0.07837	0.40063

Ex 1: New Statistics

- There are also some new statistics in the parameter estimate section.

Parameter Estimates			Standardized
Variable	Type I SS	Type II SS	Estimate
INTERCEP	118909	2342.896465	0.00000000
X1	5957.022495	4394.149832	0.67133970
X2	18.646037	15.897886	-0.04208963
X3	830.442921	830.442921	0.27302912

- The Type I SS and Type II SS we know about from our discussion of extra SS. These are not tested in PROC REG, but we can get tests from PROC GLM. For the moment we will use the t-tests just discussed, since F tests of these SS are identical to the t-test of H_o: $\beta_i = 0$.

Ex 1: New Statistics *(continued)*

- As a reminder, Type I SS are
  - SSX1
  - SSX2 | X1
  - SSX3 | X1, X2
- and type II or III SS (same for reg) are
  - SSX1 | X2, X3
  - SSX2 | X1, X3
  - SSX3 | X1, X2
- Note that the last variable is the same for both types. This is always true.

Ex 1: New Statistics *(continued)*

- **Do an F test of these, first calculate the Mean Square (all have one d.f.), and then divide the MS by the MSError, which in this example has 14 d.f.**
- **The result would then be compared to tabular values from the F table with 1, 14 d.f.**
- **This can also be used to test each parameter estimate against zero.**

Ex 1: New Statistics *(continued)*

- **Another note on the types of SS. The Sequential SS (Type I) will always sum to the SSRegression (not counting the intercept).**
- **The Partial SS (Type II or III) may sum to less than the SSRegression or to more than the SSRegression (not counting the intercept).**

Ex 1: New Statistics *(continued)*

- **In this case the SSRegression is 6806.11145.**
  - **The Type I SS sum exactly to this value**
  - **5957.022+18.646+830.443=6806.111**
  - **The Type II SS sum to less than the SSRegression**
  - **4394.150+15.898+830.443=5240.491**

Ex 1: New Statistics *(continued)*

- **Another new statistic here is the standardized regression coefficient.**

Parameter Estimates			**Standardized**
Variable	Type I SS	Type II SS	**Estimate**
INTERCEP	118909	2342.896465	**0.00000000**
X1	5957.022495	4394.149832	**0.67133970**
X2	18.646037	15.897886	**-0.04208963**
X3	830.442921	830.442921	**0.27302912**

  - **When we see the word "standardized" we are usually talking about some transformation of a variable to a mean of zero and variance of 1 (one).**
  - **This is like a Z or t score.**

Ex 1: New Statistics *(continued)*

- **The standardization is applied to the raw data, the original Y_j and X_{ij} values.**
  - **Y_j = standardized Y_j value = $(Y_j - \bar{Y}) / S_Y$**
  - **Note that S_Y denotes the standard deviation and not the standard error.**
- **The X'X, X'Y and Y'Y matrices are calculated with these values, and the $(X'X)^{-1}$ matrix is a correlation matrix.**
- **For simpler calculations see last pages of handout.**

Ex 1: New Statistics *(continued)*

- **Standardization is sometimes used to put variables on the same scale.**
- **For example, if our slope (Y units per X unit) is meaningful in terms of the original scale (e.g. mg phosphorus available per mg in the soil) we may want to keep the original scale for interpretative purposes.**

Ex 1: New Statistics *(continued)*

- **However, in other cases our scales may be arbitrary. For example, if we are trying to predict a Freshman's first semester college performance (scale 0 to 4) from SAT (scale 0 to 36), ACT verbal (scale 200 to 800) and High School GPA (scale 0 to 4), then the arbitrary scales may confuse and complicate the study. We could "standardize" the 4 variables so that all have a mean = 0 and variance = 1.**

Ex 1: New Statistics *(continued)*

- **Since the original scales are arbitrary we loose little by doing this. The resulting regression would have regression coefficients that are without scale, and whose "relative size" would give an indication of the "relative importance" or "relative impact" of the variable in determining the value of the predicted value.**

Ex 1: New Statistics *(continued)*

- **When standardized, the bigger the value of the regression coefficient, the more important the variable.**
- **These are the "standardized regression coefficients, and they are used extensively in some disciplines.**
- **In your discipline, take note of the statistics presented in the literature.**

Ex 1: New Statistics

Parameter Estimates			Standardized
Variable	Type I SS	Type II SS	Estimate
INTERCEP	118909	2342.896465	0.00000000
X1	5957.022495	4394.149832	0.67133970
X2	18.646037	15.897886	-0.04208963
X3	830.442921	830.442921	0.27302912

- The standardized regression coefficients indicate that the X_1 variable is the most "important", while the X_2 variable is the least "important".
- Note that tests of the Type II SS (or t-tests of the slopes) would give similar results in this case.

Ex 1: New Statistics *(continued)*

- **More new statistics. The second section of the "Parameter estimates" provides some squared "correlations".**

Parameter Estimates

Variable	DF	Squared Semi-partial Corr Type I	Squared Partial Corr Type I	Squared Semi-partial Corr Type II	Squared Partial Corr Type II
INTERCEP	1	.	.	.	.
X1	1	0.48080787	0.48080787	0.35466406	0.44039930
X2	1	0.00150497	0.00289868	0.00128316	0.00283921
X3	1	0.06702736	0.12947464	0.06702736	0.12947464

- **This is another measure of "importance" for each variable.**

Ex 1: New Statistics *(continued)*

- **Recall the R^2 is the SSModel / SSTotal (both corrected). However, since we are not very interested in the overall model could we get an R^2 type statistics for each variable? Of course we could, in fact there are four.**
- **What would you imagine the individual R^2 values to be?**
- **You might guess the Type I or Type II SS divided by the total.**

Ex 1: New Statistics *(continued)*

- **Congratulations, you just invented the "Squared semi-partial correlation Type I" and the "Squared semi-partial correlation Type II".**
  - **Squared semi-partial correlation TYPE I = SCORR1 = $SeqSSX_j$ / SSTotal**
  - **Squared semi-partial correlation TYPE II = SCORR2 = $PartialSSX_j$ / SSTotal**

Ex 1: New Statistics *(continued)*

- **But think about the Extra SS we talked about earlier. When variables go into a model they account for some of the SSTotal, and that fraction of the SS is not available to later variables, or it may even enhance the SS accounted for by later variables.**
- **Doesn't it seem that we should look at the SS available to a variable when we consider how much SS it accounts for?**

Ex 1: New Statistics *(continued)*

- **This is more in keeping with the concept of "Partial" SS we talked about earlier.**
- **So, when a variable (say X_1) enters the model after the other variables, what SS are available to it?**
- **Obviously, the SS it accounted for was available ($SSX_1|X_2,X_3$). And the SSError was also available, though not accounted for.**

Ex 1: New Statistics *(continued)*

- **So, the SS available to each variable is the part it accounts for (SSX_i|all other variables), plus the part no variable accounts for (SSError).**
- **If we use this as the available SS to be accounted for, instead of the SSTotal, we have the "Squared Partial correlations".**

Ex 1: New Statistics *(continued)*

- **These are calculated as**
  - **Squared partial correlation TYPE I = PCORR1 = $SeqSSX_j$ / ($SeqSSX_j$ + SSError*)**
    - *** Note that for sequential SS the error changes as each variable enters. This must be taken into account.**
  - **Squared partial correlation TYPE II = PCORR2 = $PartialSSX_j$ / ($PartialSSX_j$ + SSError)**

Ex 1: New Statistics *(continued)*

- **So how are these used?**
- **The interpretation is similar to that of the R^2, except that it is a fraction for each variable.**
- **The ones that make the most sense to me are**
  - **For models using the Type I SS the Squared semi-partial correlation TYPE I**
  - **For models using the Type II SS, the Squared partial correlation TYPE II**

Ex 1: New Statistics *(continued)*

- **Since the Type I SS sum to the SSReg, the Semi-partial R^2 Type I will sum to the overall R^2.**
- **Since the Partial SS may sum to more or less than the SSReg, and the denominator is not the SSTotal, the sum of these partial R^2 values is unpredictable.**

Multicollinearity

- **One last statistic to evaluate variables. There is a problem that exists in multiple regression when two independent variables are very highly correlated. The problem is called multicollinearity.**
  - **At one extreme of this phenomenon is the case where two independent variables are perfectly correlated. This results in "singularity", and the X'X matrix that cannot be inverted.**

Multicollinearity *(continued)*

- **To illustrate the problem, take the following data set.**

Y	X1	X2
1	1	2
2	2	3
3	3	4

Multicollinearity *(continued)*

- **If entered in PROC REG, SAS will report problems and will fit only the first variable, since the second one is perfectly correlated. Suppose we did want to fit both parameters for X_1 and X_2, what b_i values could we get. The table below shows some possible values for b_1 and b_2.**

Multicollinearity *(continued)*

- Acceptable values of b_0, b_1 and b_2.

b0	b1	b2
0	1	0
-1	0	1
99	100	-99
999	1000	-999
-101	-100	101
-1001	-1000	1001

Multicollinearity *(continued)*

- **There are an infinite number of solutions when singularity exists, and that is why no program can, or should, fit the parameter estimates.**
- **But suppose that I took and added to one of the X_i observations the value 0.0000000001.**
- **Now the two independent variables are not perfectly correlated!!! SAS will report no error and will give a solution.**

Multicollinearity *(continued)*

- **How good is that solution. Remember how the b_i values could go way up or way down as long as they were balanced by the other?**

b0	b1	b2
0	1	0
-1	0	1
99	100	-99
999	1000	-999
-101	-100	101
-1001	-1000	1001

Multicollinearity *(continued)*

- **Typically when very high correlations exist (but NOT perfect correlations) small changes in the data result in large fluctuations of the regression coefficients.**
- **Basically, under these conditions, the regression coefficient estimates are useless.**
- **Also, the variance estimates are inflated.**

Multicollinearity *(continued)*

- **So how do we detect these problems?**
  - **First, look at the correlations, the simple correlations among the X_i variables produced by the PROC REG in the summary statistics section.**
  -
  - **Correlation**
  - **CORR X1 X2 X3 Y**
  - **X1 1.0000 0.4616 0.1520 0.6934**
  - **X2 0.4616 1.0000 0.3175 0.3545**
  - **X3 0.1520 0.3175 1.0000 0.3617**

Multicollinearity *(continued)*

- **Large correlations (usually > 0.9) can indicate potential multicollinearity problems.**
- **However, to detect Multicollinearity these statistics alone are not enough. It is possible that there is no pairwise correlation, but that some combination of X_i variables correlates with some other combination. So we need another statistic to address this.**

Multicollinearity *(continued)*

- **The Variance Inflation Factor (VIF) is the statistic most commonly used to detect this problem.**

Variable	Tolerance	Variance Inflation
INTERCEP	.	0.00000000
X1	0.78692352	1.27077152
X2	0.72432171	1.38060199
X3	0.89915421	1.11215627

Multicollinearity *(continued)*

- **VIF values over 5 or 10, or a mean of the VIF values much over 2 indicate potential problems with multicollinearity.**
- **Tolerance is just the inverse of the VIF, so as VIF go up, Tolerance goes down. Both can be used to detect multicollinearity. We will ignore Tolerance.**

Multiple Regression

- **So, the multiple regression differs from the SLR in that it has several variables.**
- **We need new statistics to examine parameter estimates from these variables, and to determine if there are problems among the variables.**
- **I will collectively refer to these as the "variable diagnostics".**

Multiple Regression *(continued)*

- **Variable diagnostics include,**
  - **The partial regression coefficients and their tests.**
  - **The standardized partial regression coefficient (note also partial).**
  - **The squared partial and semi-partial correlation values**
  - **The VIF values**
-

Additional output

- **The variances and covariances of the regression coefficients are given by $(X'X)^{-1}$. This produces the Variance-Covariance matrix.**

Covariance of Estimates

COVB	INTERCEP	X1	X2	X3
INTERCEP	324.36769134	0.7651495821	-4.545863489	-0.974950169
X1	0.7651495821	0.2891164632	-0.09904623	-0.00038649
X2	-4.545863489	-0.09904623	0.1744779464	-0.013158898
X3	-0.974950169	-0.00038649	-0.013158898	0.0124691254

Additional output *(continued)*

- **Another section of interest is the section called the "Sequential Parameter Estimates".**
- **This section gives the estimates of the b_i sequentially as each variable is entered. We are rarely interested in the sequential SS or the sequential parameter estimates themselves.**

Additional output *(continued)*

- **However, recall that multicollinearity can cause the regression coefficients to fluctuate greatly. Examining the Sequential Parameter Estimates for large fluctuations as variables enter is another indicator of multicollinearity.**

Sequential Parameter Estimates

INTERCEP	X1	X2	X3
81.277777778	0	0	0
59.258958792	1.8434360081	0	0
56.251024085	1.7897741162	0.08664925	0
43.652197791	1.7847796802	-0.083397057	0.161132691

Additional output *(continued)*

- **The last section of interest related to the variables is the output from the requested TEST statements.**
- **TEST X1=2; TEST X1=X2=X3;**
- **SAS will conduct these tests and provide F tests. Results are the same as t-tests for the same hypotheses for one degree of freedom tests. However, F tests can also conduct joint tests.**

Additional output *(continued)*

- **There are several tests provided automatically by SAS.**
- **Of course, the tests of the individual regression coefficients against zero.**
- **And the test of the model. This tests the hypothesis H_o: $\beta_1 = \beta_2 = \beta_3 = 0$, which is the same as the joint test H_o: $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$. Note that 3 parameters are tested and this is a 3 d.f. test.**

Additional output *(continued)*

- **This last test (of the full model) is not the same as the test H_o: $\beta_1 = \beta_2 = \beta_3$. This is actually equivalent to the test H_o: $\beta_1 = \beta_2$, $\beta_2 = \beta_3$ which is a 2 d.f. test. This test was requested in SAS with the following result.**
- **TEST X1=X2=X3;**

Test 2 Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	1760.86020	4.42	0.0326
Denominator	14	398.82140		

Additional output *(continued)*

- **The usual test of interest are tests of the individual regression coefficients against an hypothesized value of zero. These tests are provided automatically by SAS.**
  - **H_o: $\beta_1 = 0$**
  - **H_o: $\beta_2 = 0$**
  - **H_o: $\beta_3 = 0$**

Variable	DF	Estimate	Std Error	Parameter=0	Prob>|T|
INTERCEP	1	43.652198	18.01021075	2.424	0.0295
X1	1	1.784780	0.53769551	3.319	0.0051
X2	1	-0.083397	0.41770557	-0.200	0.8446
X3	1	0.161133	0.11166524	1.443	0.1710

Additional output *(continued)*

- It is not unusual to want a test of some hypothesized value other than zero. This can be requested with a test statement.
- for example, TEST X1=2;
- $H_o: \beta_1 = 2$

Test 1 Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	63.89578	0.16	0.6950
Denominator	14	398.82140		

Multiple Regression

- **This concludes the "variable diagnostic" section of notes. The observation diagnostics are in Part 2 of this series of slides.**
- **An Example of GLM output is also included in the second section.**