Chapter 10 : Inferential tools for Multiple Regression

A note on regression analysis in SAS –In SAS, regression can be done in PROC REG, PROC GLM, PROC MIXED and numerous other specialty procedures. PROC GLM has an advantage in dealing with dummy or indicator variables because these variables can be set up automatically when listed in a CLASS statement.

PROC REG has the advantage of having many special regression diagnostic tools and good facilities for testing hypotheses about the regression coefficients (CLB) and producing confidence intervals.

PROC MIXED is a relatively new procedure that has some aspects of both REG and GLM. This procedure has a CLASS statement and will handle dummy variable the same as GLM. This procedure also has some of the regression diagnostics available in REG and facilities for confidence intervals. This procedure does not normally produce sums of squares.

The major elements discussed in Chapter 10 (statistical inference) are discussed in the example below. Many of the usual diagnostics usually included in our regression analysis are not germane to the discussion and have been omitted from this handout, but are included in the SAS program posted on the WWW.

Bat echolocation example.

Data is given for the energy expenditure for selected species of echolocating bats, non-echolocating bats and birds. The energy expenditure for flight is a function of the mass of the animal. The question here is "Do bats that use echolocation expend more energy for flight than animals that do not use echolocation?"

Plots of the data are given below where the echolocating bats are assigned type=3, non-echolocating bats are type=1 and birds are type=2.

```
Plot of Expenditure*Mass.  Symbol is value of Type.

Expenditure |
            |
         45 +
            |                                                        2                                              1
            |
            |
         40 +
            |
            |
            |
            |
         35 +                                                                                    1
            |                                            2
            |
            |
         30 +                                                           2
            |
            |                                                               2
            |
            |                                                 2 2
         25 +
            |
            |                                   1  2
            |                                          1
            |
         20 +
            |
            |
            |
            |
         15 +                             2
            |                 2
            |
            |
            |
         10 +
            |          2  3
            |
            |
            |
          5 +
            |      2
            |
            |    2
            |  3
          0 +
            |
            ---+------------+------------+------------+------------+------------+------------+------------+------------+--
               0           100          200          300          400          500          600          700          800
                                                                Mass
NOTE: 2 obs hidden.
```

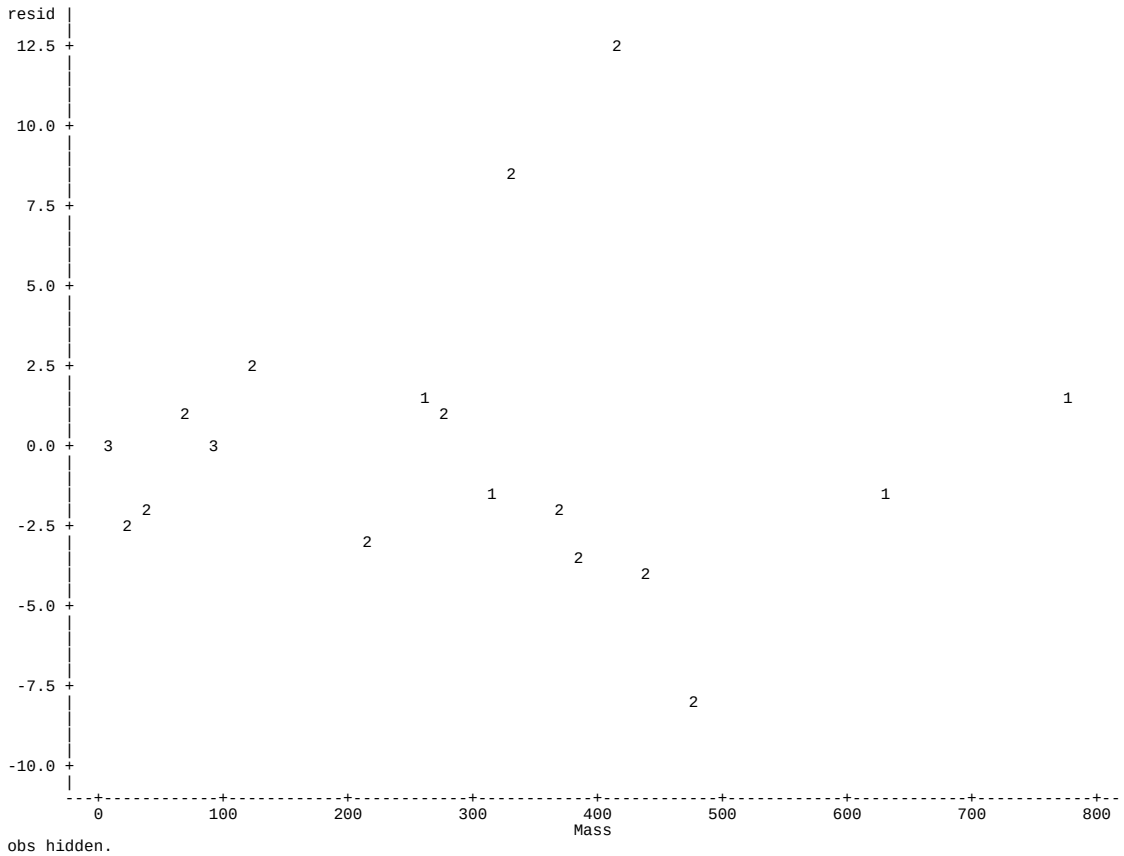
The regression analysis was run on Expenditure (the dependent variable) and mass (the quantitative independent variable). Indicator variables were included for the types (TypeNON=1 for non-echolocating bats, 0 otherwise; TypeBird=1 for birds, 0 otherwise) along with a the interaction of the indicator variables with the quantitative variable mass. Major results are given below.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3367.37636	673.47527	26.50	<.0001
Error	14	355.73236	25.40945		
Corrected Total	19	3723.10872			

Root MSE	5.04078	R-Square	0.9045
Dependent Mean	19.51800	Adj R-Sq	0.8703
Coeff Var	25.82631		

Plot of resid*Mass. Symbol is value of Type.



NOTE: 2 obs hidden.

Tests for Normality

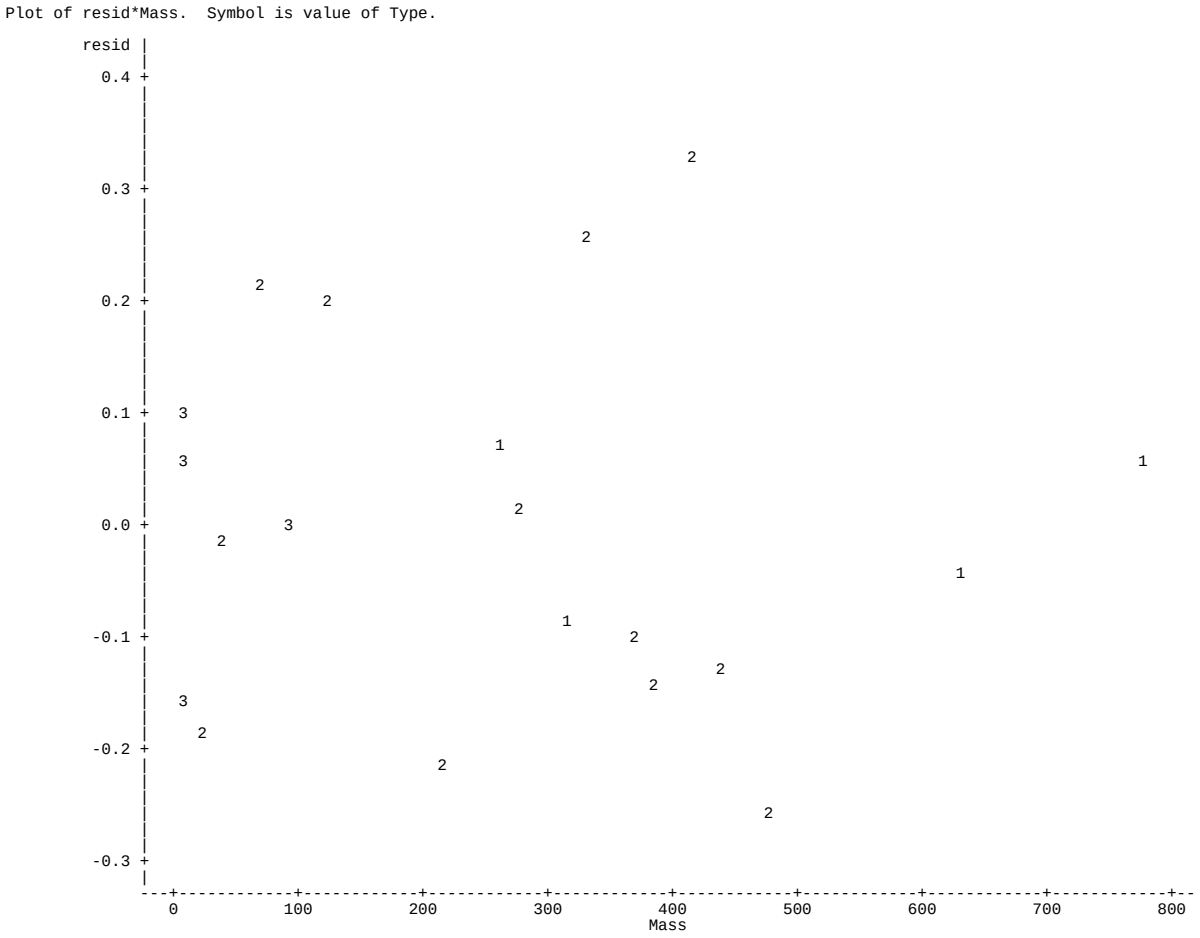
Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.870782	Pr < W	0.0121
Kolmogorov-Smirnov	D	0.193724	Pr > D	0.0474
Cramer-von Mises	W-Sq	0.170174	Pr > W-Sq	0.0118
Anderson-Darling	A-Sq	1.023665	Pr > A-Sq	0.0087

The same analysis was then rerun using the logarithm of expenditure and the log of mass with the following results.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	29.46993	5.89399	163.44	<.0001
Error	14	0.50487	0.03606		
Corrected Total	19	29.97480			

Root MSE	0.18990	R-Square	0.9832
Dependent Mean	2.48220	Adj R-Sq	0.9771
Coeff Var	7.65047		



Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.96598	Pr < W	0.6688
Kolmogorov-Smirnov	D	0.097033	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.032292	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.236314	Pr > A-Sq	>0.2500

In comparing the non-log transformed analysis to the log transformed analysis, clearly the log transformation was a superior model terms of the F test of the relationship (F = 26.50 versus 163.44), the R^2 value (0.9045 versus 0.9832) the homogeneity (as indicated by the residual plots) and in meeting the assumption of normality (P>W = 0.0121 versus 0.6688). The analysis and its interpretation will proceed using the logarithm transformed values.

Inference on regression coefficients – many research objectives investigated by regression can be addressed with hypotheses tests or confidence intervals. Most of these issues were discussed in simple linear regression and extend to multiple regression fairly easily.

Least squares estimates and standard errors – These estimates are readily available in the SAS output. These are produced by default in PROC REG and can be requested with the option "solution" added to the model statement of PROC GLM and PROC MIXED. Confidence intervals on the estimates can also be requested in PROC REG and the α values specified (α = 0.05 by default).

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-1.47052	0.24767	-5.94	<.0001	-2.00172	-0.93932
LMass	1	0.80466	0.08668	9.28	<.0001	0.61874	0.99058
TypeNON	1	1.26807	1.28542	0.99	0.3406	-1.48888	4.02502
TypeBIRD	1	-0.11032	0.38474	-0.29	0.7785	-0.93551	0.71487
LMassNON	1	-0.21487	0.22362	-0.96	0.3529	-0.69450	0.26475
LMassBird	1	0.03071	0.10283	0.30	0.7696	-0.18984	0.25127

In our example the first question to ask is if the energy expenditure indeed is a function of mass, as one would expect. The test of the regression coefficient on measuring expenditure per mass has a P-value of 0.0126, so there does appear to be a correlation between mass and energy usage.

The other regression coefficients fit intercept differences and slope differences as follows.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} X_{2i} + \beta_5 X_{1i} X_{3i} + \varepsilon_i ,$$

where X_1 is the quantitative variable, mass
X_2 and X_3 are dummy variables for non-echolocating bats and birds
X_4 and X_5 are interactions of the two dummy variables with the quantitative variable

Using simpler notation,

$$Y_i = \beta_0 + \beta_1 X_{MASSi} + \beta_2 X_{NONECHOi} + \beta_3 X_{BIRDSi} + \beta_4 X_{MASSi} X_{NONECHOi} + \beta_5 X_{MASSi} X_{BIRDSi} + \varepsilon_i ,$$

For the group getting a 0 for both dummy variables (echolocating bats) the model reduces to

$$Y_i = \beta_0 + \beta_1 X_{MASSi} + \varepsilon_i$$

When $X_2 = 1$, indicating a non-echolocating bat the model reduces to $Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_{MASSi} + \varepsilon_i$, so β_2 is the difference in intercepts between the two bat types and β_4 is the slope difference

When $X_3 = 1$, indicating a bird the model reduces to $Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_{MASSi} + \varepsilon_i$, so β_3 is the difference in intercepts between echolocating bats and birds while β_5 is the slope difference

Least squares estimates and standard errors for each of these parameters are given in the SAS output, along with the test of each estimate against an hypothesized value of zero.

So, do non-echolocating bats differ from echolocating bats (H_0: $\beta_3 = 0$)? To test the intercept adjustment only, a model should be fitted without the slope. The parameter estimates are the same as above, but β_4 and β_5 are not included. Estimates for this model are given below.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-1.49770	0.14987	-9.99	<.0001	-1.81540	-1.17999
LMass	1	0.81496	0.04454	18.30	<.0001	0.72053	0.90938
TypeNON	1	-0.07866	0.20268	-0.39	0.7030	-0.50832	0.35100
TypeBIRD	1	0.02360	0.15760	0.15	0.8828	-0.31050	0.35770

The value of β_3 is the estimated difference non-echolocating bats and echolocating bats. The difference in intercepts (without slopes in the model) is given by TypeNON = 0.07866. This value

indicates that the line for non-echolocating bats is lower (since it is negative) by –0.079 energy units than the line for echolocating bats. When tested against zero, the P-value = 0.7030. Therefore, we find no difference in the intercepts, or levels, between the two types of bats. A confidence interval on the true value of the parameter is also available in SAS, and was calculated as $P(-0.50832 \leq \beta_3 \leq 0.35100) = 0.95$.

The estimated values and their significance depend on what other values are included in the model. The same hypothesis test (H_0: $\beta_3 = 0$) will be different if done after the interactions are included in the model. With the interactions present the estimate was 1.26807 (note the change of sign). The value was also not significantly different from zero (P = 0.3406) and the confidence interval is $P(-1.48888 \leq \beta_3 \leq 4.02502) = 0.95$.

Tests and confidence intervals for linear combinations of coefficients. In SAS PROC REG the TEST statement can test linear combinations of parameter estimates. A common approach with group, or indicator variables, is to test them jointly, in groups. PROC GLM and PROC MIXED will do this automatically. However, PROC REG has no CLASS statement and tests each term in the model individually. To test H_0: $\beta_2 = \beta_3 = 0$, or the equivalent H_0: $\beta_2 = 0$ and $\beta_3 = 0$, use the test statement would be "`TEST TypeNON = TypeBird = 0;`". The results of this test are given below.

Joint test of intercepts with slope interactions in the model.

```
Test Test_of_intercepts Results for Dependent Variable LExpend
Source              DF    Mean Square    F Value    Pr > F
Numerator            2        0.02061       0.57    0.5773
Denominator         14        0.03606
```

Joint test of intercepts with slope interactions in the model.

```
Test Test_of_intercepts Results for Dependent Variable LExpend
Source              DF    Mean Square    F Value    Pr > F
Numerator            2        0.01479       0.43    0.6593
Denominator         16        0.03458
```

Joint test of slope interactions – The interactions with the dummy variables would also be tested jointly, usually with the intercepts in the modle. To test H_0: $\beta_4 = \beta_5 = 0$, the test statement would be "`TEST LMassNON = LMassBird = 0;`". The results of this test are given below.

```
Test Test_of_slopes Results for Dependent Variable LExpend
Source              DF    Mean Square    F Value    Pr > F
Numerator            2        0.02422       0.67    0.5265
Denominator         14        0.03606
```

PROC GLM can accomplish many of the same tests more easily with the CLASS statement. Results for the SAS statements "`PROC GLM DATA=BatDat; class Type; MODEL LExpend = LMass Type Type*LMass / solution;`" are given below. Note that the estimates of the joint tests and individual tests are all the same.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LMass	1	29.39190909	29.39190909	815.04	<.0001
Type	2	0.02957359	0.01478680	0.41	0.6713
LMass*Type	2	0.04844954	0.02422477	0.67	0.5265

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LMass	1	3.37875389	3.37875389	93.69	<.0001
Type	2	0.04122472	0.02061236	0.57	0.5773
LMass*Type	2	0.04844954	0.02422477	0.67	0.5265

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-1.470515265	B	0.24767033	-5.94	<.0001
LMass		0.804657049	B	0.08668453	9.28	<.0001
Type	1	1.268067693	B	1.28542004	0.99	0.3406
Type	2	-0.110322504	B	0.38474216	-0.29	0.7785
Type	3	0.000000000	B	.	.	.
LMass*Type	1	-0.214874992	B	0.22362264	-0.96	0.3529
LMass*Type	2	0.030713281	B	0.10283304	0.30	0.7696
LMass*Type	3	0.000000000	B	.	.	.

Redefining the reference level – The reference level is that level of the categorical treatment that is not coded with a "1". As a result, it is the level against which all other levels are compared. In the fits with PROC REG above the references level was "echolocating bats". The dummy variables were coded for non-echolocating bats and birds, so β_2 and β_3 were differences between these groups and the reference level, echolocating bats. Likewise, the parameter estimates for β_4 and β_5, the interactions of the two dummy variables with the quantitative variable were comparisons of non-echolocating bats and birds to the references level, echolocating.

The GLM procedure automatically codes the group variable with dummy variables. The default reference level chosen as the reference level is the one in the last alphanumeric position. In the example above the first dummy variable is for type = 1 (non- echolocating bats), and the second for type = 2 (birds). These both use echolocating bats as a reverence level, since type=3 is the last alphanumeric position. Since the same reference level was chosen for the PROC REG coding, the results are the same.

It should be noted that the choice of a reference level is arbitrary. Obviously, if there is one level against which all others are to be compared then this level should be the reference level. The choice does not affect the joint tests,

Notes on the R^2 statistic – The text suggest that the value of R^2 can always be made 100% by adding enough independent variables. However, their choice of example is a little unfortunate, since it is a polynomial. What is true about polynomials is that if there are k different values of the independent variable (X), each with a single observation (n_i–1), a polynomial of order k will provide a perfect fit to all points. If, however, there are multiple values at any of the values of the X variable (n_i>1) the fit will not be perfect (100%). If the number of replicates at different values of X increase the value of R^2 will decrease.

It is also true that if there are n different observations in a dataset, and the investigator attempts to fit this data with a model having n–1 variables, the fit will be "perfect" in that there will be no

error term. That means no tests and not confidence intervals, so it is not "perfect" in a statistical sense.

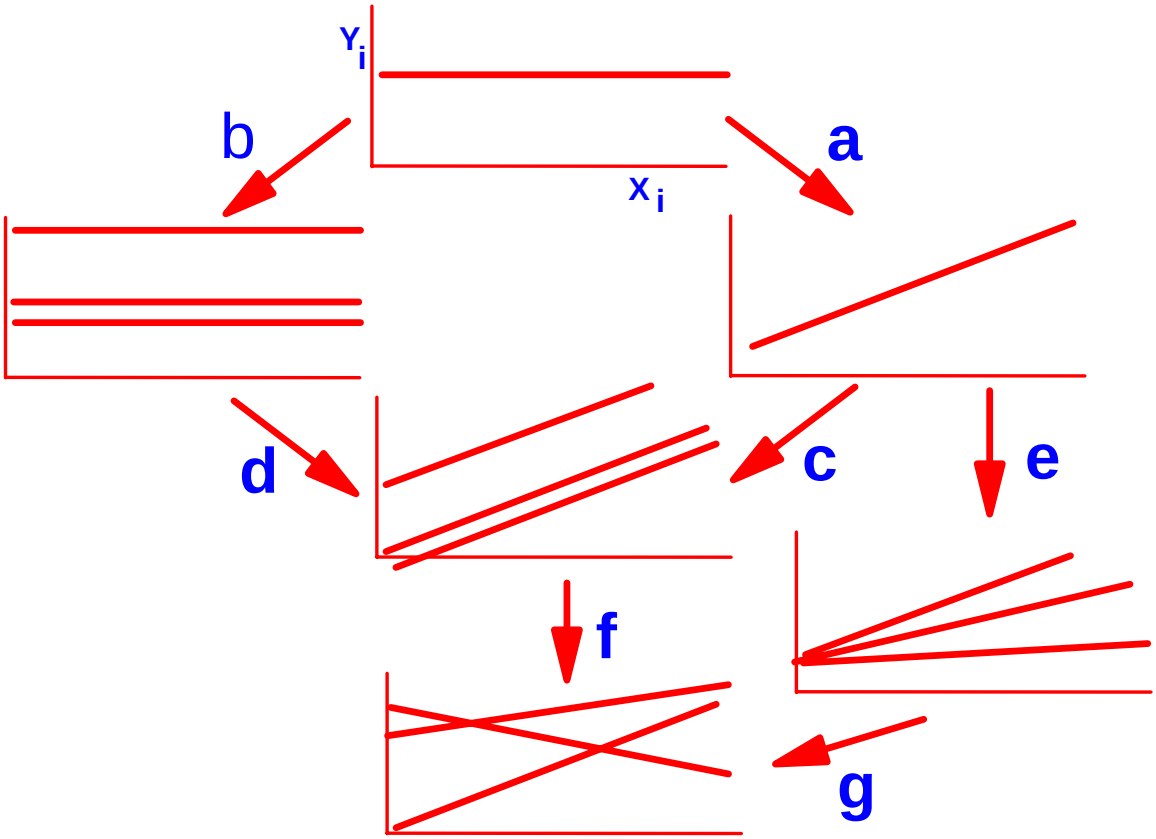
Transformed data – When using transformed data as in the analysis above, all analyses are done on the data in the transformed form. This means all tests of hypotheses and the calculation of all estimates and confidence intervals. Once calculated, estimates and the upper and lower limit sof the confidence intervals can be back-transformed or detransformed. The reason that confidence values must be calculated on the transformed data is that the estimates of standard errors cannot be detransformed.

Scope of inference – At the end of the "bat echolocation" analysis the text discusses scope of inference. The text points out that statistical inference should be restricted to the species used, and that any inference to a larger population would be speculative. This is somewhat related to a concept that will be very important when we finish regression and return to Analysis of Variance. If the material included in a study includes all levels of a treatment or all categories of interest in the treatment, then the treatment is called a "FIXED" effect, and conclusions are limited to those levels that were included. If, however, the treatment levels are randomly selected from a large number of levels they are called "RANDOM" effects. These represent the variability among all levels, and we can draw some inferences to the whole population. More on this later when we continue our discussion of Analysis of Variance.

Extra SS

Should tests of the intercepts be done before or after a slope is present in the model? Should the tests of slopes be done before or after any tests of intercepts? Actually, this depends on the objectives of the study and the hypotheses the investigator wishes to test.

In our previous discussion of analysis of covariance we discussed the graphic below.



In this graph each arrow and letter represents the differences between a full and a reduced model. Assuming we start with a correction factor, the decision to then add indicator variables (arrow b) or a quantitative variable (arrow a) depends on what the investigator believes would be the most meaningful full model. This full model then becomes the reduced model when additional variables are entered, following arrow c or d.

At each step along this graphic we have a full and reduced model and a corresponding test an extra sum of squares. Note that the graphic does not depict adding the dummy variables one at a time, these are usually treated in groups. These tests are best represented by the tests in the GLM or MIXED model analysis.

The extra SS in the PROC GLM are as follows.

Source	DF	Type I SS	Mean Square	F Value	Pr > F	TEST
LMass	1	29.39190909	29.39190909	815.04	<.0001	a
Type	2	0.02957359	0.01478680	0.41	0.6713	c
LMass*Type	2	0.04844954	0.02422477	0.67	0.5265	f

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
LMass	1	3.37875389	3.37875389	93.69	<.0001	
Type	2	0.04122472	0.02061236	0.57	0.5773	g
LMass*Type	2	0.04844954	0.02422477	0.67	0.5265	

Graphic	Meaning	Hypothesis	Extra SS
a	Test of a slope	H_0: $\beta_1 = 0$	SSX1|X0
b	Test of separate means or levels	H_0: $\beta_2 = \beta_3 = 0$	SSX2, X3|X0
c	Test of separate intercepts or levels	H_0: $\beta_2 = \beta_3 = 0$	SSX2, X3|X0, X1
d	Test of a slope	H_0: $\beta_1 = 0$	SSX1|X0, X2, X3
e	Test of separate slopes	H_0: $\beta_4 = \beta_5 = 0$	SSX4, X5|X0, X1
f	Test of separate slopes	H_0: $\beta_4 = \beta_5 = 0$	SSX4, X5|X0, X1, X2, X3
g	Test of separate intercepts	H_0: $\beta_2 = \beta_3 = 0$	SSX2, X3|X0, X1, X4, X5

In the expression of the previous tests, note that various “hypotheses” are indistinguishable. Again, the results of a test depend on the other variables included in the model. For example, the hypotheses for a and c are the same, but from the Extra SS we can see that in a the analysis is a simple linear regression with only the intercept and slope. In c the slope is also fitted, but the model is adjusted for two intercepts instead of one. The results can be very different.

The extra SS and tests b, d and e are not estimated in the analysis above. To get these tests the GLM would be rerun with the order of the first two independent variables reversed (e.g. (Type LMass LMass*Type). This would give tests for b and d instead of a and c. To get the test labeled “e”, the model would have to include the interaction before the group variable term (e.g. LMass LMass*Type Type).