**ON NESTED MODELS**

CONTEXT

Consider a modeling situation where you are trying to predict Math Achievement from a standardized test (Y) based on a number of predictor variables, X's.   If you think about the nature of the predictor variables in a dataset, they often group into categories.  For this example, the X's could group like:  a)  academic ability; and, b) psychological variables like self-concept, motivation, locus of control. For example, suppose we have a variable Y(math) that is math scores that we wish to predict. We have academic variables like, READING (RDG), WRITING(WRTG), SCIENCE(SCI), CIVICS(CIV); and psychological variables like locus of control (LOCUS), motivation (MOT), and self-concept(CONCEPT) that we can use to predict Y(math).

MODEL SET UP

When you fit a multiple regression models, the results give you tests on the individual slopes for each of the variables separately, even though multicollinearity is taken into account in the estimation process.    As a researcher, you may have a broader hypothesis then the individual variables, such that in general the psychological variables do not play a role in predicting math achievement.   The individual tests don't answer this question, and really, what is being asked is a comparison of two models.

For example, suppose we have a variable Y(math) that is math scores that we wish to predict. We have academic variables like, READING (RDG), WRITING(WRTG), SCIENCE(SCI), CIVICS(CIV); and psychological variables like locus of control (LOCUS), motivation (MOT), and self-concept(CONCEPT) that we can use to predict Y(math). You might think that as a group of variables, the psychological variables LOCUS, MOT, and CONCEPT really are not important in predicting math scores. Rather, it is about academic skills that is most predictive of MATH scores. Here are the models:

FULL MODEL:

Y(math) = b0 + b1\*X(rdg) + b2\*X(wrtg) + b3\*X(sci) + b4\*X(civ) + b5\*X(locus) + b6\*X(mot) + b7\*X(concept)

REDUCED MODEL:

Y(math) = b0 + b1\*X(rdg) + b2\*X(wrtg) + b3\*X(sci) + b4\*X(civ)

Notice, the reduced model is the full model with the psychological variables removed.   The reduced model is "nested" within the full model since every variable of the reduced model is also included in the full model.

ANOVA SUMS OF SQUARES

Using nested models rely on the basic accounting mechanism that you learned in 401 for ANOVA. In studying ANOVA, you learned that the total variability in a variable can be partitioned into variability due to systematic mean shift (or mean differences between groups) and unsystematic variability (or error). In measuring this you used Sums of Squares, which are found in ANOVA tables. The accounting mechanism was that:

SS(total) = SS(treatment) + SS(error)

When we fit regression models, the same concept is in play. The total variability in Y can be split into error (the deviation of the observed values from the regression line – i.e. the residuals) plus the variability due to regression (the deviation of the regression line from the mean of the Y’s). Well, if you fit the full model, the Regression Sum of Squares is the amount of variability in Y attributable to the full model.   If you fit the reduced model, the Regression Sum of Squares is the amount of variability in Y attributable to the reduced model.  But, you always have:

SS(total) = SS(regression) + SS(error)

USING NESTED MODELS

We can use the concept of partitioning sums of squares to our advantage. The SS(total) is a constant, once you select the Y variable . It never changes. The SS(error) and SS(model) change depending on the variables included in the model. Here is the ANOVA table from a model where reading (X) is used to predict math(Y) test scores. The first model is:

**Model 1: Y(math) = b0 + b1\*X(rdg)**

If you use the ANOVA table for that regression model you obtain the following:

SS(total) = 53093.72

SS(regression\_1) = 24498.26 NOTE: SS(total) = SS(regression\_1) + SS(error\_1)

SS(error\_1) = 28595.46

If we add another variable, say writing skill, to the model. Our model becomes:

**Model 2: Y(math) = b0 + b1\*X(rdg) + b2\*X(wrtg)**

Note, Model 1 is nested within Model 2. Here is the ANOVA table for this regression model:

If you use the ANOVA table for that regression model you obtain see the following:

Note, SS(total) = 53093.72 has not changed. But, SS(regression) and SS(error) have changed. In what way?

SS(regression\_2) = 28211.58

SS(error\_2) = 24882.14

Notice that the SS for error has decreased from Model 1 to Model 2. This should make sense to you. We added in another variable to “Account” for more variability in Y, so error should be reduced. By how much?

SS(error\_1) – SS(error\_2) = 28595.46 - 24882.14 = 3713.32

Also, notice that the SS for the model has increased from Model 1 to Model 2. This also should make sense to you. We added in another variable to “Account” for more variability in Y, so the variability accounted for should go up. By how much?

SS(regression\_2) – SS(regression\_1) = 24882.14 - 24498.26 =3 713.32

Notice that these two differences are IDENTICAL! Not numerically, but measurement wise, or contextually, what is the difference between SS(regression\_2) and SS(regression\_1), or SS(error\_1) and SS(error\_2). You should think about that for a while before you read the rest of this. The difference is easily seen by looking at the linear statistical models – Model 1 and Model 2. They are nested models and the difference between the models is the one variable, X(wrtg), the writing variable. The numerical difference observed, 3713.32, is the unique contribution of X(wrtg) in explaining the total Sums of Squares for Y. You can get at this difference, either by looking a differences in errors from model to model, or from differences in regression sums of squares.

WHAT’s THE BIG AHAAA POINT?

The point is that SS(total) can be partitioned and the partitioning utilizes basic accounting principles.

1. You start with SS(total) for a response variable Y. This is a fixed amount. SS(total) = 53093.72
2. If you fit MODEL 1 (above). SS(total) gets partitioned into SS(regression\_1) + SS(error\_1). But, SS(regression\_1) = 24498.26 is the unique contribution to explaining SS(total) that is due to the first variable added to the model, X(rdg), reading. We have:

SS(total) = SS(regression\_1) + SS(error\_1) = 24498.26 + 28595.46 = 53093.72

= unique contribution of X(rdg) + error

1. If you fit MODEL 2 (above). Again, SS(total) gets partitioned into SS(regression\_2) + SS(error\_2). But, SS(regression\_2) is really made up of 2 parts. That which is uniquely attributable to X(rdg) and that which is attributable to the newly added variable, X(wrtg).

SS(total) = SS(regression\_2) + SS(error\_2) = 28211.58 + 24882.14 = 53093.72

= SS[X(rdg)] + SS[X(wrtg)] + SS(error\_2) = 24498.26 + 3713.32 + 24882.14 = 53093.72

1. It should be obvious, that if you keep adding variables to the model, one at a time, you can isolate the unique contribution to SS(total) that can be attributable to each variable, separately. But, you have to use the notion of nested models for this to work. It is a continual building up of a model, one variable at a time.
2. When you isolate the unique contribution to the total sums of squares, SS(total), for each individual variable, note that each variable accounts for 1 degree of freedom. Also note, that the model that has the largest number of variables in it, has accounted for the most SS(total) that are possible. As such, the SS(error) for this model is the best measure of unsystematic variability that is available.
3. This then is perfectly set up for F-tests. Sums of squares can be converted to Mean Squares by dividing by degrees of freedom. F statistics are always ratios of Mean Squares.

F = MS(model)/MS(error)

HOW DOES THIS WORK IN GENERAL?

For example, suppose we have a variable Y(math) that is math scores that we wish to predict. We have variables like, READING (RDG), WRITING(WRTG), SCIENCE(SCI), CIVICS(CIV), locus of control (LOCUS), motivation (MOT), and self concept(CONCEPT) that we can use to predict Y(math). You might think that as a group of variables, LOCUS, MOT, and CONCEPT really are not important in predicting math scores. Rather, it is about academic skills that is most predictive of MATH scores. Here are the models:

FULL MODEL:

Y(math) = b0 + b1\*X(rdg) + b2\*X(wrtg) + b3\*X(sci) + b4\*X(civ) + b5\*X(locus) + b6\*X(mot) + b7\*X(concept)

REDUCED MODEL:

Y(math) = b0 + b1\*X(rdg) + b2\*X(wrtg) + b3\*X(sci) + b4\*X(civ)

Note the reduced model is fully nested within the full model. All variables of reduced model are in full model. We could go through a process like was done above and isolate the unique contributions to each of the variables. {If you need practice I suggest you do that, the database is on the SAS OnDemand Cloud and is called: hsb\_high\_school\_and\_beyond\_data}. But, we really don’t need to do that.

What is the difference in the two models? Conceptually, it is the unique contributions of LOCUS, MOT, and CONCEPT. Computationally, it is:

SS(unique to LOCUS, MOT, CONCEPT) = SS(regression full model) – SS(regression reduced model)

Notice there were 3 variables isolated, so there are 3 degrees of freedom.

If you examine the ANOVA table from fitting the full model, you will see:

SS(regression full model) = 30413.53 and that SS(error full model) = 22680.19

Now, if you examine the ANOVA table from fitting the reduced model, you will see:

SS(regression reduced model) = 30338.32

From all of this, we have that

SS(unique to LOCUS, MOT, CONCEPT) = SS(regression full model) – SS(regression reduced model)

= 30413.53 - 30338.32 = 75.21

Since, there are 3 variables isolated, this Sum of Squares has 3 degrees of freedom associated with it.

THE END GAME

Being able to isolate the Sum of Squares for a sub-set of variables allows you to isolate and test hypotheses. For example,

Ho:  b5 = b6 = b7 = 0   (from the full model)  
  
The test statistic is an F test, where the numerator is the Mean Square for the unique contribution associated with the variables for b5, b6 and b7 (LOCUS, MOT and CONCEPT). This numerator is:

SS(unique to LOCUS, MOT, CONCEPT) / Degrees of freedom = 75.21 / 3 = 25.07

The denominator has to be the Mean Square for Error. But which error? Well, this is easy! The full model has the most variables included into to account for variability in Y. As such, it has the best estimate of error. Use MS(error) from the full model:

MS(Error from full model) = 22680.19 / 592 = 38.31

Finally, the F statistic is:

F = MS(unique variables) / MS(error full model) = 25.07 / 38.31 = 0.654

Here, you would fail to reject the null hypothesis, so you would retain the notion that b5 = b6 = b7 = 0 .

I do hope this helps. Different texts use different notations to convey these same ideas. Just keep in mind that all that is happening is that Sums of Squares are being partitioned and uniquely attributable to individual variables.