Review of Multiple Regression

Richard Williams, University of Notre Dame, https://academicweb.nd.edu/~rwilliam/
Last revised January 3, 2022

Assumptions about prior knowledge. This handout attempts to summarize and synthesize the basics of Multiple Regression that should have been learned in an earlier statistics course. It is therefore assumed that most of this material is indeed "review" for the reader. (Don't worry too much if some items aren't review; I know that different instructors cover different things, and many of these topics will be covered again as we go through the semester.) Those wanting more detail and worked examples should look at my course notes for Grad Stats I. Basic concepts such as means, standard deviations, correlations, expectations, probability, and probability distributions are not reviewed.

In general, I present formulas either because I think they are useful to know, or because I think they help illustrate key substantive points. For many people, formulas can help to make the underlying concepts clearer; if you aren't one of them you will probably still be ok.

Linear regression model

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + ... + \beta_k X_{kj} + \varepsilon_j = \alpha + \sum_{i=1}^{k} \beta_i X_{ij} + \varepsilon_j = E(Y_j \mid X) + \varepsilon_j$$

β_i = partial slope coefficient (also called partial regression coefficient, metric coefficient). It represents the change in E(Y) associated with a one-unit increase in X_i when all other IVs are held constant.

α = the intercept. Geometrically, it represents the value of E(Y) where the regression surface (or plane) crosses the Y axis. Substantively, it is the expected value of Y when all the IVs equal 0.

ε = the deviation of the value Y_j from the mean value of the distribution given X. This error term may be conceived as representing (1) the effects on Y of variables not explicitly included in the equation, and (2) a residual random element in the dependent variable.

Parameter estimation (Metric Coefficients): In most situations, we are not in a position to determine the population parameters directly. Instead, we must estimate their values from a finite sample from the population. The sample regression model is written as

$$Y_j = a + b_1 X_{1j} + b_2 X_{2j} + ... + b_k X_{kj} + e_j = a + \sum_{i=1}^{k} b_i X_{ij} + e_j = \hat{Y}_j + e_j$$

where a is the sample estimate of α and b_k is the sample estimate of β_k.

Computation of b_k		
Case	*Formula(s)*	*Comments*
All Cases	$\widehat{\mathrm{B}} = \left(X'X\right)^{-1} X'Y$	This is the general formula but it requires knowledge of matrix algebra to understand that I won't assume you have.
1 IV case	$b = \frac{s_{xy}}{s_x^2}$	Sample covariance of X and Y divided by the variance of X
Computation of a (all cases)	$a = \bar{y} - \sum_{k=1}^{K} b_k \bar{x}_k$	Compute the betas first. Then multiply each beta times the mean of the corresponding X variable and sum the results. Subtract from the mean of y.

Question. Suppose $b_k = 0$ for all variables, i.e. none of the IVs have a linear effect on Y. What is the predicted value of Y? What is the predicted value of Y if all Xs have a value of 0?

Standardized coefficients. The IVs and DV can be in an infinite number of metrics. Income can be measured in dollars, education in years, intelligence in IQ points. This can make it difficult to compare effects. Hence, some like to "standardize" variables. In effect, a Z-score transformation is done on each IV and DV. The transformed variables then have a mean of zero and a variance of 1. Rescaling the variables also rescales the regression coefficients. Formulas for the standardized coefficients include

1 IV case	$b' = r_{yx}$	In the one IV case, the standardized coefficient simply equals the correlation between Y and X
General Case	$b'_k = b_k * \frac{s_{x_k}}{s_y}$	As this formula shows, it is very easy to go from the metric to the standardized coefficients. There is no need to actually compute the standardized variables and run a new regression.

We interpret the standardized coefficients as follows: a one standard deviation increase in X_k results in a b'_k standard deviation increase in Y.

Standardized coefficients are somewhat popular because variables are in a common (albeit weird) metric. Hence, it is possible to compare magnitudes of effects, causing them to sometimes be used as a measure of the "importance" of a variable. They are easier to work with mathematically. The metric of many variables is arbitrary and unintuitive anyway. Hence, you might as well make the scaling standard across variables.

Nevertheless, standardized effects tend to be looked down upon because they are not very intuitive. Worse, they can be very misleading; for example, when making comparisons across groups. As we will see, Duncan argues this point quite forcefully.

The ANOVA Table: Sums of squares, degrees of freedom, mean squares, and F.
Before doing other calculations, it is often useful or necessary to construct the ANOVA (Analysis of Variance) table. There are four parts to the ANOVA table: sums of squares, degrees of freedom, mean squares, and the F statistic.

Sums of squares. Sums of squares are actually sums of squared deviations about a mean. For the ANOVA table, we are interested in the Total sum of squares (SST), the regression sum of squares (SSR), and the error sum of squares (SSE; also known as the residual sum of squares).

Computation of sums of squares	
Case	*Formula(s)*
General case:	$SST = \sum_{j=1}^{N} (y_j - \bar{y})^2 = SSR + SSE$ $SSR = \sum_{j=1}^{N} (\hat{y}_j - \bar{y})^2 = SST - SSE$ $SSE = \sum_{j=1}^{N} (y_j - \hat{y}_j)^2 = \sum_{j=1}^{N} e_j^{\,2} = SST - SSR$

Question: What do SSE and SSR equal if it is always the case that $y_j = \hat{y}_j$, i.e. you make "perfect" predictions every time? Conversely, what do SSR and SSE equal if it is always the case that $\hat{y}_j = \bar{y}$, i.e. for every case the predicted value is the mean of Y?

Other calculations. The rest of the ANOVA table easily follows (K = # of IVs, not counting the constant):

Source	SS	DF	MS	F
Regression (or explained)	SSR	K	MSR = SSR/K	F = MSR / MSE
Error (or residual)	SSE	N - K - 1	MSE = SSE/(N - K - 1)	
Total	SST	N - 1	MST = SST/(N - 1)	

An alternative formula for F, which is sometimes useful when the original data are not available (e.g. when reading someone else's article) is

$$F = \frac{R^2 * (N - K - 1)}{(1 - R^2) * K}$$

The above formula has several interesting implications, which we will discuss shortly.

Uses of the ANOVA table. As you know (or will see) the information in the ANOVA table has several uses:

- The F statistic (with df = K, N-K-1) can be used to test the hypothesis that $\rho^2 = 0$ (or equivalently, that all betas equal 0). In a bivariate regression with a two-tailed alternative hypothesis, F can test whether $\beta = 0$. F (along with N and K) can be used to compute R^2.
- MST = the variance of y, i.e. s_y^2.
- SSR/SST = R^2. Also, SSE/SST = 1 - R^2.
- MSE is used to compute the standard error of the estimate (s_e).
- SSE can be used when testing hypotheses concerning nested models (e.g. are a subset of the betas equal to 0?)

Multiple R and R^2. Multiple R is the correlation between Y and $\hat{Y}$. It ranges between 0 and 1 (it won't be negative.) Multiple R^2 is the amount of variability in Y that is accounted for (explained) by the X variables. If there is a perfect *linear* relationship between Y and the IVs, R^2 will equal 1. If there is no linear relationship between Y and the IVs, R^2 will equal 0. Note that R and R^2 are the sample estimates of ρ and ρ^2.

Some formulas for R^2.

R^2 = SSR/SST	Explained sum of squares over total sum of squares, i.e. the ratio of the explained variability to the total variability.
$R^2 = \dfrac{F * K}{(N - K - 1) + (F * K)}$	This can be useful if F, N, and K are known
One IV case only: $R^2 = b'^2$	Remember that, in standardized form, correlations and covariances are the same.

Incidentally, R^2 is biased upward, particularly in small samples. Therefore, *adjusted* R^2 is sometimes used. The formula is

$$\text{Adjusted } R^2 = 1 - \left(\frac{(N-1)(1-R^2)}{(N-K-1)} \right) = 1 - (1-R^2) * \frac{N-1}{N-K-1}$$

Note that, unlike regular R^2, Adjusted R^2 can actually get smaller as additional variables are added to the model. One of the claimed benefits for Adjusted R^2 is that it "punishes" you for including extraneous and irrelevant variables in the model. Also note that, as N gets bigger, the difference between R^2 and Adjusted R^2 gets smaller and smaller.

Sidelight. Why is R^2 biased upward? McClendon discusses this in "Multiple Regression and Causal Analysis", 1994, pp. 81-82.

Basically he says that sampling error will always cause R^2 to be greater than zero, i.e. even if no variable has an effect R^2 will be positive in a sample. When there are no effects, across multiple samples you will see estimated coefficients sometimes positive, sometimes negative, but either way you are going to get a non-zero positive R^2. Further, when there are many Xs for a given sample size, there is more opportunity for R^2 to increase by chance.

So, adjusted R^2 wasn't primarily designed to "punish" you for mindlessly including extraneous variables (although it has that effect), it was just meant to correct for the inherent upward bias in regular R^2.

Standard error of the estimate. The standard error of the estimate (s_e) indicates how close the actual observations fall to the predicted values on the regression line. If $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, then about 68.3% of the observations should fall within $\pm 1s_e$ units of the regression line, 95.4% should fall within $\pm 2s_e$ units, and 99.7% should fall within $\pm 3s_e$ units. The formula is

$$s_e = \sqrt{\frac{SSE}{N-K-1}} = \sqrt{MSE}$$

Standard errors. b_k is a *point* estimate of β_k. Because of sampling variability, this estimate may be too high or too low. s_{bk}, the standard error of b_k, gives us an indication of how much the point estimate is likely to vary from the corresponding population parameter. We will discuss standard errors more when we talk about multicollinearity. For now I will simply present this formula and explain it later.

Let H = the set of all the X (independent) variables.

Let G_k = the set of all the X variables *except* X_k.

The following formulas then hold:

General case:	$s_{b_k} = \sqrt{\frac{1-R^2_{YH}}{(1-R^2_{X_kG_k})*(N-K-1)}} * \frac{s_y}{s_{X_k}}$	This formula makes it clear how standard errors are related to N, K, R^2, and to the inter-correlations of the IVs.

Hypothesis Testing. With the above information from the sample data, we can test hypotheses concerning the population parameters. Remember that hypotheses can be one-tailed or two-tailed, e.g.

H_0: $\beta_1 = 0$ or H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$ H_A: $\beta_1 > 0$

The first is an example of a two-tailed alternative. Sufficiently large positive *or* negative values of b_1 will lead to rejection of the null hypothesis. The second is an example of a 1-tailed alternative. In this case, *we will only reject the null hypothesis if b_1 is sufficiently large and positive.* If b_1 is negative, we automatically know that the null hypothesis should not be rejected, and there is no need to even bother computing the values for the test statistics. *You only reject the null hypothesis if the alternative is better.*

EXAMPLE:

H_0: $\beta_1 = 0$
H_A: $\beta_1 > 0$

N = 1000, $b_1 = -10$, $t_1 = -50$. Should you reject the Null? (HINT: Most people say reject. Most people are wrong. Explain why.)

With regression, we are commonly interested in the following sorts of hypotheses:

Tests about a single coefficient. To test hypotheses such as

H_0: $\beta_1 = 0$ or H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$ H_A: $\beta_1 > 0$

we typically use a T-test. The T statistic is computed as

$$T_{N-K-1} = \frac{b_k - \beta_{k0}}{s_{b_k}} = \frac{b_k}{s_{b_k}}$$

The latter equality holds if we hypothesize that $\beta_k = 0$. The degrees of freedom for T are N-K-1. If the T value is large enough in magnitude, we reject the null hypothesis.

If the alternative hypothesis is two-tailed, we also have the option of using *confidence intervals* for hypothesis testing. The confidence interval is

$$b_k - (t_{\alpha/2} * s_{b_k}) \leq \beta_k \leq b_k + (t_{\alpha/2} * s_{b_k})$$

If the null hypothesis specifies a value that lies within the confidence interval, we will not reject the null. If the hypothesized value is outside the confidence interval, we will reject. Suppose, for example, $b_k = 2$, $s_{bk} = 1$, and the "critical" value for T is 1.96. In this case, the confidence interval will range from .04 to 3.96. If the null hypothesis is $\beta_k = 0$, we will reject the null, because 0 does not fall within the confidence interval. Conversely, if the null hypothesis is $\beta_k = 1$, we will not reject the null hypothesis, because 1 falls within the confidence interval.

If the alternative value is two-tailed *and* there is only one IV, we can also use the F-statistic. In the one IV case, $F = T^2$.

Global F Test: Tests about all the beta coefficients. We may want to test whether any of the betas differ from zero, i.e.

H_0: $\beta_1 = \beta_2 = \beta_3 = ... = \beta_K = 0$

H_A: At least one $\beta \neq 0$.

This is equivalent to a test of whether $\rho^2 = 0$ (since if all the betas equal 0, ρ^2 must equal 0). The F statistic, with d.f. = K, N-K-1, is appropriate. As noted above, F = MSR/MSE; or equivalently,

$$F = \frac{R^2 * (N - K - 1)}{(1 - R^2) * K}$$

Question: Looking at the last formula, what happens to F as R^2 increases? N increases? K increases? What are the implications of this?

Note that

- Larger R^2 produce bigger values of F. That is, the stronger the relationship is between the DV and the IVs, the bigger F will be.
- Larger sample sizes also tend to produce bigger values of F. The larger the sample, the less uncertainty there is whether population parameters actually differ from 0. Particularly with a large sample, it is necessary to determine whether statistically significant results are also substantively meaningful. Conversely, in a small sample, even large effects may not be statistically significant.
- If additional variables do not produce large enough increases in R^2, then putting them in the model can actually decrease F. (Why?) Hence, if there is too much "junk" in the model, it may be difficult to detect important effects.
- The F statistic does not tell you which effects are significant, only that at least one of them is.
- In a bivariate regression (and only in a bivariate regression) Global $F = T^2$.

Tests about a subset of coefficients. We sometimes wish to test hypotheses concerning a subset of the variables in a model. For example, suppose a model includes 3 demographic variables (X1, X2, and X3) and 2 personality measures (X4 and X5). We may want to determine whether the personality measures actually add anything to the model, i.e. we want to test

H_0: $\beta_4 = \beta_5 = 0$
H_A: β_4 and/or $\beta_5 \neq 0$

Another common example is when we have multi-category qualitative variables (e.g. Religion, where 1= Catholic, 2 = Protestant, 3 = Other). Here, we compute a set of **Dummy Variables** (see Appendix A) and then test whether the entire set of dummies is statistically significant. (e.g. X4 = 1 if Catholic, 0 otherwise; X5 = 1 if Protestant, 0 otherwise. Note that there are always at least one fewer dummy variables than there are categories, and that one category (the "excluded category") is coded 0 on all the dummies.)

We will talk about such situations much more later in the semester. For now we will simply note that there are various ways to conduct such tests:

- An incremental F test can be used. This procedure is described in the optional Appendix C. It requires estimating an *unconstrained* model (in which all variables of interest are included in the model) and a *constrained* model (in which the subset of variables to be tested are excluded from the model. In Stata the user-written `ftest` command makes this easy. Note that the exact same cases must be used for estimating both the constrained and unconstrained models.
- A Wald test can be used. A Wald test requires only the estimation of the unconstrained model in which all variables of interest are included. The Stata `test` and `testparm` commands make this easy.

Appendix A: Dummy Variables [Important but not too hard. Read this on your own and get help from the professor or TA if needed]

(1) We frequently want to examine the effects of both quantitative and qualitative independent variables on quantitative dependent variables. Dummy variables provide a means by which qualitative variables can be included in regression analysis. The procedure for computing dummy variables is as follows:

(a) Suppose there are L groups. You will compute L-1 dummy variables. For example, suppose you had two categories for race, White and Black. In this example, L = 2, since you have two groups. Hence, one dummy variable would be computed. If we had 3 groups (for example, White, Black, and Other) we would construct 2 dummy variables.

(b) The first group is coded 1 on the first dummy variable. The other L-1 groups are coded 0. On the second dummy variable (if there is one), the second group is coded 1, and the other L-1 groups are coded zero. Repeat this procedure for each of the L-1 dummy variables.

(c) Note that, under this procedure, the Lth group is coded 0 on every dummy variable. We refer to this as the "excluded category." Another way of thinking of it is as the "reference group" against which others will be compared.

For example, suppose our categories were White, Black, and Other, and we wanted White to be the excluded category. Then,

Dummy1 = 1 if Black, 0 if Other, 0 if White
Dummy2 = 0 if Black, 1 if Other, 0 if White

Incidentally, note that if we wanted to compute it, Dummy3 = 1 - Dummy1 - Dummy2. We do not include Dummy3 in our regression models, because if we did, we would run into a situation of perfect collinearity. (This should make intuitive sense: If we know someone is not Black and not Other, then we know that the person is White.)

Also note that, before computing dummies, you may want to combine some categories if the Ns for one or more categories are very small. For example, you would have near-perfect multicollinearity if you had a 1000 cases and one of your categories only had a dozen people in it. In the present case, if there were very few others, you might just want to compute a single dummy variable that stood for White/NonWhite.

(2) When a single dummy variable has been constructed and included in the equation, a T test can be used. When there are multiple dummy variables, an incremental F test or Wald Test is appropriate.

(3) EXAMPLE: The dependent variable is income, coded in thousands of dollars. BLACK = 1 if Black, 0 otherwise; OTHER = 1 if other, 0 otherwise. White is the "excluded" category, and Whites are coded 0 on both BLACK and OTHER.

```
Variable        B

BLACK        -10.83
OTHER        - 5.12
(Constant)    29.83

For whites, predicted income = 29.83 -(10.83 * 0) - (5.12 * 0) = 29.83.
For blacks, predicted income = 29.83 -(10.83 * 1) - (5.12 * 0) = 19.00.
For others, predicted income = 29.83 -(10.83 * 0) - (5.12 * 1) = 24.71.
```

In this simple example, the constant is the mean for members of the excluded category, Whites. The coefficients for BLACK and OTHER show how the means of Blacks and Others differ from the mean of Whites. That is why Whites can be thought of as the "reference group": the dummy variables show how other groups differ from them.

In a more complicated example, in which there were other independent variables, you can think of the dummy variable coefficients as representing the average difference between a White, a Black and an Other who were otherwise identical, i.e. had the same values on the other IVs.

Example:

```
Variable        B

BLACK        - 4.00
OTHER        - 2.10
EDUCATION    + 1.50
(Constant)   +12.00
```

This model says that, if a White, a Black, and an Other all had the same level of education, on average the Black would make $4,000 less than the White, while the Other would make on average $2,100 less than the White. So if, for example, each had 10 years of education,

```
white: predicted income = 12.00 -(4.00* 0) - (2.10 * 0) + (1.50 * 10) = 27.0
black: predicted income = 12.00 -(4.00* 1) - (2.10 * 0) + (1.50 * 10) = 23.0
other: predicted income = 12.00 -(4.00* 0) - (2.10 * 1) + (1.50 * 10) = 24.9.
```

As a substantive aside, note that the dummy variable coefficients in this hypothetical example became much smaller once education was added to the model. This often happens in real world examples. In this case, a possible explanation might be that much, but not all, of the differences in mean income between the races reflects the fact that Whites tend to have more years of education than do other racial groups.

Appendix B: Semipartial and Partial correlations/ Stepwise regression

[Review on your own]

Semipartial correlations (also called part correlations) indicate the "unique" contribution of a variable. They are sometimes used as a way of assessing the "importance" of a variable.

Specifically, the squared semipartial correlation for a variable tells us how much R^2 will decrease if that variable is removed from the regression equation. Some relevant formulas are

$$sr_k^2 = R_{YH}^2 - R_{YG_k}^2 = b_k'^2 * (1 - R_{X_kG_k}^2) = b_k'^2 * Tol_k,$$
$$sr_k == b_k' * \sqrt{1 - R_{X_kG_k}^2} = b_k' * \sqrt{Tol_k}$$

As before, H = the set of all the X (independent) variables, G_k = the set of all the X variables *except* X_k. Thus, to get X_k's unique contribution to R^2, first regress Y on all the X's to get R^2_{YH}. Then regress Y on all the X's except X_k to get R^2_{YGk}. The difference between the R^2 values is the squared semipartial correlation. Or alternatively, the standardized coefficients and the Tolerances can be used to compute the squared semipartials. Note that, the more "tolerant" a variable is (i.e. the less highly correlated it is with the other IVs), the greater its unique contribution to R^2 will be.

Optional. The partial correlation measures the strength of the association between X_i and Y when all other X's are controlled for. Formulas are below but I mostly want to focus on semipartial correlations.

$$pr_k^2 = \frac{sr_k^2}{1 - R_{YG_k}^2} = \frac{sr_k^2}{1 - R_{YH}^2 + sr_k^2},$$
$$pr_k = \frac{sr_k}{\sqrt{1 - R_{YG_k}^2}} = \frac{sr_k}{\sqrt{1 - R_{YH}^2 + sr_k^2}}$$

Optional. Some alternative formulas that may occasionally come in handy are

$$sr_k = \frac{T_k * \sqrt{1 - R_{YH}^2}}{\sqrt{N - K - 1}}$$

$$pr_k = \frac{T_k}{\sqrt{T_k^2 + (N - K - 1)}}$$

With these formulas you only need to estimate the model that has all the variables in it. Note that the only part of the calculations that will change across X variables is the T value; therefore the X variable with the largest partial and semipartial correlations will also have the largest T value (in magnitude).

Note that

- Once one variable is added or removed from an equation, all the other semipartial correlations can change. The semipartial correlations only tell you about changes to R^2 for one variable at a time.
- Semipartial correlations are used in Stepwise Regression Procedures, where the computer (rather than the analyst) decides which variables should go into the final equation. In a forward stepwise regression, the variable which would add the largest increment to R^2 (i.e. the variable which would have the largest semipartial correlation) is added next (provided it is statistically significant). In a backwards stepwise regression, the variable which would produce the smallest decrease in R^2 (i.e. the variable with the smallest semipartial correlation) is dropped next (provided it is not statistically significant.)
- The squared semipartials provide another possible formula for R^2

Let H = the set of all the X (independent) variables.

Let G_k = the set of all the X variables *except* X_k.

$R^2_{YH} = R^2_{YG_k} + sr^2_k$ $R^2_{YG_k} = R^2_{YH} - sr^2_k$	The squared semipartial for X_k tells you how much R^2 will go up if X_k is added to the model or how much R^2 will decline if X_k is dropped from the model.

Appendix C: Incremental F Tests about a subset of coefficients. [Optional]

[NOTE: We will talk about incremental F tests in great detail toward the middle of the course. For now, I will just briefly review what they are.] We sometimes wish to test hypotheses concerning a subset of the variables in a model. For example, suppose a model includes 3 demographic variables (X1, X2, and X3) and 2 personality measures (X4 and X5). We may want to determine whether the personality measures actually add anything to the model, i.e. we want to test

H_0: $\beta_4 = \beta_5 = 0$
H_A: β_4 and/or $\beta_5 \neq 0$

One way to proceed is as follows.

1. Estimate the model with all 5 IVs included. This is known as the *unconstrained model*. Retrieve the values for SSE and/or R^2 (hereafter referred to as SSE_u and R^2_u.) [NOTE: If using the R^2 values, copy them to several decimal places so your calculations will be accurate.]
2. Estimate the model using only the 3 demographic variables. We refer to this as the *constrained model*, because the coefficients for the excluded variables are, in effect,

constrained to be 0. Retrieve the values for SSE and/or R^2 (hereafter referred to as SSE_c and R^2_c).

3. Compute the following:

$$F_{J,N-K-1} = \frac{(R_u^2 - R_c^2)*(N-K-1)}{(1-R_u^2)*J} = \frac{(SSE_c - SSE_u)/J}{SSE_u/(N-K-1)} = \frac{(SSE_c - SSE_u)*(N-K-1)}{SSE_u * J}$$

where J = the number of constraints imposed (in this case, 2) and K = the number of variables in the *unconstrained* model (in this case, 5). Put another way, J = the error d.f. for the constrained model minus the error d.f. for the unconstrained model.

If J = 1, this procedure will lead you to the same conclusions a two-tailed T test would (the above F will equal the T^2 from the unconstrained model.)

If J = K, i.e. all the IVs are excluded in the constrained model, the incremental F and the Global F become one and the same; that is, the global F is a special case of the incremental F, where in the constrained model all variables are constrained to have zero effect. You can see this by noting that, if there are no variables in the model, $R^2 = 0$.

When you can use incremental F. In order to use the incremental F test, it must be the case that

- The sample is the same for each model estimated. This assumption might be violated if, say, missing data in variables used in the unconstrained model caused the unconstrained sample to be smaller than the constrained sample. You should be careful how missing data is getting handled in your statistical routines
- One model must be "nested" within the other; that is, one model must be a constrained, or special case, of the other. For example, if one model contains IVs X1-X5, and another model contains X1-X3, the latter is a special case of the former, where the constraints imposed are $\beta 4 = \beta 5 = 0$. If, however, the second model included X1-X3 and X6, it would not be nested within the first model and an incremental F test would not be appropriate.
- Other types of constraints can also be tested with an incremental F test. For example, we might want to test the hypothesis that $\beta 1 = \beta 2$, i.e. two variables have equal effects. We'll discuss such possibilities later.

Other comments

- Constrained and unconstrained are relative terms. An unconstrained model in one analysis can be the constrained model in another. In reality, every model is "constrained' in the sense that more variables could always be added to it.
- Wald tests, which are easily done in Stata and which we will discuss this semester, are an alternative to incremental F tests.
- We are also often interested in doing such tests when estimating sequences of nested models. So, for example, Model 1 may include X1, X2 and X3; Model 2 may add X4 and X5; Model 3 adds X6 and X7; and so on. With each model we may want to test whether the variables added in that model have effects that significantly differ from 0.