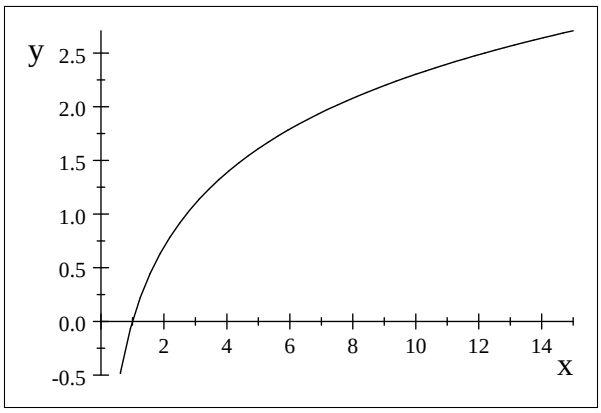
Chapter 5 outline, Econometrics

# 1 The General Linear Model

When we talk about the LINEAR regression model, we are talking about a model that is linear in the parameters (the parameters are $\beta_1, \beta_2, \beta_{3,}...$). There are models which are linear in the parameters that can test NONLINEAR relationships between X and Y. Consider the following 3 models:

1. $Y = \beta_1 + \beta_2 X_2 + \beta_3 (X_2)^2 + \varepsilon$
2. $Y = \gamma_1 + \gamma_2 \ln(X_2) + \varepsilon$
3. $\ln(Y) = \alpha_1 + \alpha_2 \ln(X_2) + \varepsilon$

Model 1 is linear in the parameters (the $\beta$'s) but suggests a nonlinear (specifically parabolic) relationship between X and Y. The equation that we might normally think of is $Y = aX^2 + bX + c$. Model 2 also suggests a nonlinear relationship between $X_2$ and $Y$. For those of you unfamiliar (or who may have forgotten) with the log function, the graph of $Y = \ln(X)$ is:




This is a nonlinear function of $X$. Model 3 can be shown to be a transformation of: $Y = \omega_1 (X_2)^{\omega_2} \varepsilon$. This should look like some of the production functions that you have seen in upper division microeconomic studies. If we take logs of both sides we get model 3 above (by the way what would the plot of $\ln(Y) = \ln(X)$ look like? You should get the 45 degree line, although the plot will not exist for values of $X \leq 0$). Some other models are:

- Exponential model: $\ln(Y) = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$
- Reciprocal model: $\frac{1}{Y} = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$
- Interaction model: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_2 X_3) + \varepsilon$

## 1.1 Interaction Model

For now we will focus on the interaction model. What is the interaction term ($\beta_4(X_2X_3)$) doing in this model? We know that in a simple linear model like $Y = \beta_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon$ the coefficients $\beta_2$ and $\beta_3$ tell us how much $Y$ will change when either $X_2$ or $X_3$ changes, holding the other constant. However, suppose we thought that an increase of $X_2$ was less important as $X_3$ increased. To make the example more concrete, consider the following model: $wage = \beta_1 + \beta_2 age + \beta_3 tenure + \varepsilon$. We hypothesize that both age and tenure are important in determining wages. However, suppose we also hypothesize that an increase in one unit (year) of tenure will have a larger effect for people who are 30 years old than for people who are 50 years old. How could we introduce this into the model? We could add an interaction term for school and tenure. The model would then be: $wage = \beta_1 + \beta_2 age + \beta_3 tenure + \beta_4(tenure)(age) + \varepsilon$. The sign on $\beta_4$ would tell us if our hypothesis is correct. If $\beta_4$ is statistically significant and the sign on $\beta_4$ is negative, then this means that a one unit increase of tenure on wages is larger for young people than for older people. I get the following results when I run this regression on the labordata sample that we are using for the computer homeworks.

| | |
|---|---|
| $\beta_2$ | 0.12783 |
| $\beta_3$ | 0.78343 |
| $\beta_4$ | −0.01091 |

What does this mean for interpreting the model? The model is: $wage = \beta_1 + \beta_2 age + \beta_3 tenure + \beta_4(tenure)(age) + \varepsilon$. What is the effect of a one unit increase in tenure? To find this effect we can take the partial derivative of wage with respect to tenure. This will give us: $\frac{\partial wage}{\partial tenure} = \beta_3 + \beta_4 age$. So what does this mean? Take a person who is 30 and compare to a person who is 50. If the 30 year-old is given one more year of tenure, then the effect on wages is given by: $\beta_3 + \beta_4 30$. If the 50 year-old is given one more year of tenure, then the effect on wages is given by: $\beta_3 + \beta_4 50$. If $\beta_4$ is positive this means the 50 year-old will gain more than the 30 year-old; if $\beta_4$ is negative then the 30 year-old will gain more than the 50 year-old. Our model says that $\beta_4$ is negative. How much will a 30 year-old gain with one additional year of tenure? $\beta_3 + \beta_4 30 = .78343 + (-.01091) * 30 = 0.456\,13$, or about 46 cents. How much will a 50 year-old gain with one additional year of tenure? $\beta_3 + \beta_4 50 = .78343 + (-.01091) * 50 = 0.237\,93$, or about 24 cents.

Direct interpretation of the coefficient on the interaction term can be difficult. In the example above, how would we interpret −.01091? We could say that a 1 cent decrease in wage occurs when either tenure is increased by one year (holding age constant) or age is increased by one year (holding tenure constant). However, the point of the model is to explain what happens when there is a one-unit increase in the variable (either age or tenure) so neglecting to incorporate the effect of $\beta_2$ for age and $\beta_3$ for tenure leads to a misinterpretation of the effect either variable has. You could make a general statement such as, "The coefficient on the interaction term suggests that a one-unit increase in either variable will be less important as the other variable increases". When

interpreting effects of independent variables on the dependent variable I suggest that when you look at interaction terms you write down the model, and then write down how a one unit increase in a particular independent variable will affect the dependent variable. Working a simple example like the one above is probably the best way to see how the interaction term affects the dependent variable.

### 1.2 Interpreting coefficients for the other models

In class I passed out a handout that explains how you calculate a one-unit effect of $X$ on $Y$ as well as a 1% effect of $X$ on $Y$ for the different model specifications. Depending on the specification of the model the coefficients will have different meanings (the numerical estimate of $\beta_2$ in $Y = \beta_1 + \beta_2 X_2 + \varepsilon$ does NOT mean the same thing as the numerical estimate of $\gamma_2$ in $\ln(Y) = \gamma_1 + \gamma_2 X_2 + \varepsilon$).

## 2 Dummy Variables

Up until now we have focused strictly on using quantitative variables (age, tenure, number of bedrooms, square feet of floor space, etc.) as independent variables. While we will continue to use ONLY quantitative variables for the dependent variables (using qualitative variables as dependent variables involves methods we may or may not cover in this class, but are covered in the book), we would like to be able to incorporate qualitative variables as independent variables. A qualitative variable is a variable that has no direct analog numerically. For example, it was suggested that gender and race be included as independent variables in the answer to question 10 of the applied portion of the first test. How do we incorporate these variables? Suppose our spreadsheet looks like:

| wage | age | tenure | school | experience | gender |
|---|---|---|---|---|---|
| 12.42 | 41 | 5 | 13 | 22 | male |
| 6.50 | 24 | 0 | 16 | 2 | male |
| 15.00 | 37 | 12 | 17 | 14 | female |
| 9.87 | 56 | 35 | 9 | 41 | female |

etc.

We want our regression model to be:

$wage = \beta_1 + \beta_2 age + \beta_3 tenure + \beta_4 school + \beta_5 experience + \beta_6 gender + \varepsilon$

There is a slight problem. When we calculate our regression coefficients we have some formulas. Consider the intercept, $\beta_1 = \bar{Y} - \beta_2\bar{X}_2 - \beta_3\bar{X}_3 - \beta_4\bar{X}_4 - \beta_5\bar{X}_5 - \beta_6\bar{X}_6$. This formula will work fine until we get to $\bar{X}_6$. What is the mean of a column that consists of the words male and female? It doesn't exist. So we need to transform our qualitative variable into a quantitative variable. To do this we create what are known as dummy variables. A dummy variable for gender (also called binary variable) is just a variable that takes on the value 1 if gender is male and 0 if gender is female (we could make our dummy variable 1 if gender is female and 0 if gender is male – it will not matter for purposes of the overall fit of the model, but the numerical sign of the dummy variable will

change in a very predictable fashion). Once this transformation is complete we can then estimate our model.

So what will dummy variables do? Suppose we have a very simple model, $Y = \beta_1 + \beta_2 X_2 + \varepsilon$, where $X_2$ is a dummy variable. What is the $E[Y]$ if $X_2 = 1$? $E[Y] = \beta_1 + \beta_2$. What is $E[Y]$ if $X_2 = 0$? $E[Y] = \beta_1$. Thus a dummy variable acts as an intercept shifter. If the dummy variable takes on the value of 1, the intercept will become $\beta_1 + \beta_2$. If the dummy variable takes on the value of 0, the intercept is just $\beta_1$. When we add more independent variables the dummy variable performs the same function – it simply shifts the intercept depending on whether the qualitative variable is classified as a 1 or a 0.

## 2.1 Dummy variable trap

The dummy variable trap occurs when you add a dummy variable for EACH of the values a qualitative variable can have. Suppose you wished to estimate a model that included a dummy variable for male (=1 if the observation is male, 0 otherwise) as well as a dummy variable for female (=1 if the observations is female, 0 otherwise). In this case you will have fallen into the dummy variable trap (assuming that the data on gender contain only male and female values) by causing PERFECT colinearity among your variables. Suppose your model is: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, where $X_2$ is a dummy variable for male and $X_3$ is a dummy variable for female. If we were to look at your spreadsheet of data it would look like:

| wage | age | tenure | school | exp | gender | male | female | constant |
|---|---|---|---|---|---|---|---|---|
| 12.42 | 41 | 5 | 13 | 22 | male | 1 | 0 | 1 |
| 6.50 | 24 | 0 | 16 | 2 | male | 1 | 0 | 1 |
| 15.00 | 37 | 12 | 17 | 14 | female | 0 | 1 | 1 |
| 9.87 | 56 | 35 | 9 | 41 | female | 0 | 1 | 1 |

etc.

Notice that I have included the column for the constant term. The constant term is just a column of ones (although it could be twos or threes). If we add together the rows for male dummy and female dummy at each observation we get a column of ones, which is exactly the same as the column of ones in the constant column. This is perfect colinearity – the three variables (male dummy, female dummy, and constant) form a PERFECT linear relationship. Even if we change the constant term to a column of twos it is STILL a perfect linear relationship because now constant=2*(male dummy + female dummy). So we have a decision to make – do we drop the constant term, the male dummy, or the female dummy? There are reasons (we will not go into the details, but mainly it has to do with calculating $R^2$) that we do not want to drop the constant term. That narrows our choice down to the male dummy and the female dummy – how will we decide? It doesn't matter. Our results will be the "same". Well, not EXACTLY the same, but very close. See the section on interpreting dummy variables that follows.

## 2.2 Interpreting dummy variables

Again, suppose you are concerned with including either the male dummy variable or the female dummy variable. Also suppose your two competing models will be: Model 1 – $wage = \beta_1 + \beta_2 tenure + \beta_3 male + \varepsilon$ and Model 2 – $wage = \gamma_1 + \beta_2 tenure + \gamma_3 female + \varepsilon$. First, if it is true that all of your "gender" observations are male and female, then your coefficient on tenure ($\beta_2$) will remain unchanged. However, the intercepts ($\beta_1$ and $\gamma_1$) and the coefficients on your dummy variables ($\beta_3$ and $\gamma_1$) will change, but in a predictable fashion. The estimate that you get for $\beta_3$ will be the exact same as your estimate for $\gamma_3$, except that it will have the OPPOSITE sign. The intercept in model 1 ($\beta_1$) will be equal to the intercept in model 2 plus the coefficient on female in model 2 ($\gamma_1 + \gamma_3$). The intercept in model 2 ($\gamma_1$) will be equal to the intercept in model 1 plus the coefficient on male in model 1 ($\beta_1 + \beta_3$). So what do these coefficients mean?

Model 1, intercept ($\beta_1$): In model 1 the intercept tells us how much a FEMALE worker with zero tenure will earn.

Model 1, coefficient on male ($\beta_3$): In model 1 this coefficient tells us how much more (or less) a MALE will make when compared to a female with the same years of tenure. When compared to a female with zero years of tenure, a male worker will earn the intercept PLUS the coefficient on male.

Model 2, intercept ($\gamma_1$): In model 2 the intercept tells us how much a MALE worker with zero tenure will earn.

Model 2, coefficient on female ($\beta_3$): In model 2 this coefficient tells us how much more (or less) a FEMALE will make when compared to a male with the same years of tenure. When compared to a male with zero years of tenure, a female worker will earn the intercept PLUS the coefficient on female.

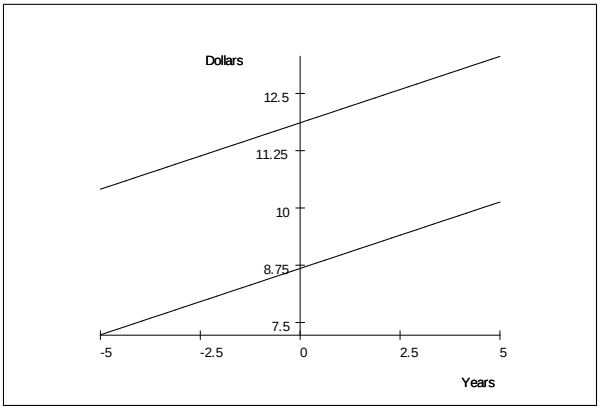
The main point is that the coefficient on the dummy variable tells us how much more (or less) the people with the characteristic captured by the dummy variable earn with respect to the group that has NOT been included as a dummy variable.

Since we know that the models yield the same results, let's look at the estimated regression equations for MALE and FEMALE (based on our data set):

MALE: $wage = 11.86 + 0.29 tenure$

FEMALE: $wage = 8.68 + 0.29 tenure$

Graphically, our regression lines look like:




where the top line is the regression line for MALE and the bottom line is the regression line for FEMALE. Note that the Y-axis is in dollars and the X-axis is in years (since tenure is measured in years).

Suppose we had three groups of people, OLD, MIDDLE-AGED, and YOUNG. We could create three dummy variables (one for each group) although we would only include TWO dummy variables in any regression model that we want to estimate (to avoid perfect colinearity). Suppose we leave out the YOUNG. Then the coefficient on OLD will tell us how much more (or less) the OLD earn when compared to the YOUNG. The coefficient on MIDDLE-AGED will tell us how much more (or less) the MIDDLE-AGED earn when compared to the YOUNG. How would we find out how much more (or less) the OLD make when compared to the MIDDLE-AGED? We could run a separate regression where we leave out the OLD (then they would be the reference group), or we could just subtract the MIDDLE-AGED coefficient from the OLD. Either way works.

## 2.3 Other uses for dummy variables

There are a few other dummy variable models that we can use.

### 2.3.1 Dummy variables as interaction terms

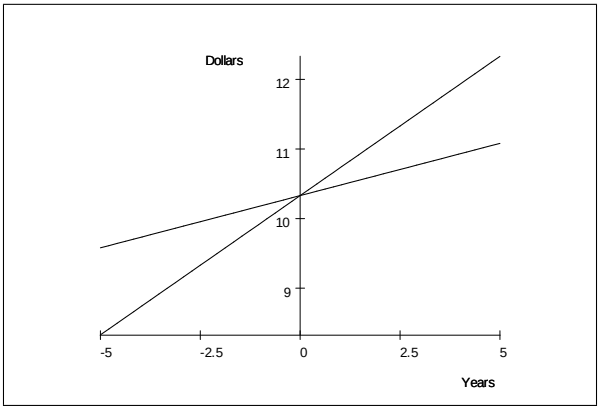
We can also make dummy variables act as interaction terms. Suppose we have the following model, $wage = \beta_1 + \beta_2 tenure + \beta_3(male * tenure) + \varepsilon$. Now, if $male = 1$, the equation becomes: $wage = \beta_1 + \beta_2 tenure + \beta_3(tenure) + \varepsilon$ which is the same as $wage = \beta_1 + (\beta_2 + \beta_3)tenure + \varepsilon$. So we are allowing the slope coefficient to change. If $male = 0$, the equation becomes: $wage = \beta_1 + \beta_2 tenure + \varepsilon$. So the slope coefficient for males is equal to $\beta_2 + \beta_3$ while the slope coefficient for females is equal to $\beta_2$.

The estimated regression equations using our data are:

MALE: $wage = 10.33 + 0.4tenure$

FEMALE: $wage = 10.33 + 0.15tenure$

These results suggest that the wages of males rise faster than the wages of females when tenure increases. Graphically, our regression lines are:




where the line with the steeper slope is the estimated regression line for males. Once again the Y-axis is in dollars and the X-axis is in years.

### 2.3.2 Using dummy variables to allow the slope and intercept to change

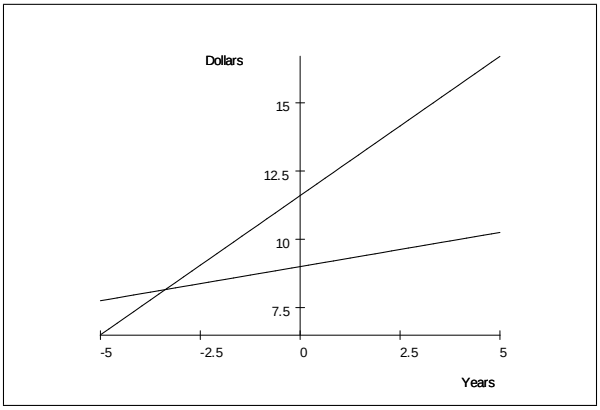
We can also use dummy variables to allow the slope and intercept to change. Suppose we think that both the slope and intercept is different for males and females. We can estimate the following equation: $wage = \beta_1 + \beta_2 tenure + \beta_3 male + \beta_4 (male)(tenure) + \varepsilon$. What will the regression model be if $male = 1$? It will be: $wage = \beta_1 + \beta_2 tenure + \beta_3 + \beta_4 (tenure) + \varepsilon$. This simplifies to: $wage = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) tenure + \varepsilon$. So the intercept for males becomes $\beta_1 + \beta_3$ and the slope becomes $\beta_2 + \beta_4$. For females, we have: $wage = \beta_1 + \beta_3 tenure$.

The estimated regression equations are:

MALE: $(9 + 2.60) + (0.25 + 0.77) tenure = 11.6 + 1.02 tenure$

FEMALE: $9 + 0.25 tenure$

All coefficients are statistically significant, suggesting that both the slope and intercept of wages (based on tenure) differ for males and females. If $\beta_3$ was not statistically different than zero we could conclude that the intercept for male and female wages is the same. If $\beta_4$ was not statistically different than zero we could conclude that the slope for male and female wages is the same. The plots of the regression lines are:




where the MALE regression line has the steeper slope and higher intercept.

### 2.3.3 Alternative version for allowing the slope and intercept to change

There is another method for allowing the slope and intercept to change. We could estimate two separate equations, one for males and one for females. We would then estimate:

MALE: $wage = \beta_1 + \beta_2 tenure + \varepsilon$ (only for the males in the sample)

FEMALE: $wage = \gamma_1 + \gamma_2 tenure + \nu$ (only for the females in the sample)

Why would we do this? When we estimated the equations for males and females we (implicitly) assumed that the error variance was constant for males and females. This may not be the case however. The error variance could be different. In that case a more correct model specification would be to run two different models, allowing the error variance to differ between the two. We will discuss tests of the error variance when we discuss heteroscedasticity.

# 3 Hypothesis testing of more than one variable

In this section we will discuss a few hypothesis tests that can be performed to test for the significance of more than one variable. So far we have discussed the t-test for the significance of one variable and the F-test for the significance of the regression. Most of what we will do in this section is modify the F-test.

## 3.1 Joint tests on several regression coefficients

Suppose we wanted to include only independent variables that were significant at the 5% level in our regression model. One possibile method of eliminating insignificant independent variables would be to remove any variables that had a t-value less than 1.96 (or a p-value greater than .05) from the model. However, there is a chance that two variables are insignificant individually but are JOINTLY significant. To make sure that we are not dropping variables that are

jointly significant from our regression equation, we must perform a test of joint significance. We have already seen a test of this type. Recall that to check the significance of the regression for the model $Y = \beta_1+\beta_2X_2+\beta_3X_3+...+\beta_kX_k+\varepsilon$, we tested:

$H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$

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

We then set up our test statistic as:

$\frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{k-1,n-k}$

We will be using a VERY similar test to test for the joint significance of regressors. In fact, the test above for the signficance of the regression is a special case of the test we will now develop.

Suppose we have the following model:

$Y = \beta_1 + \beta_2X_2 + \beta_3X_3 + ... + \beta_kX_k + \varepsilon$

We will call this the UNRESTRICTED model. The reason we will call this the unrestricted model is because we are not restricting any of our $\beta$'s to be zero. We are estimating all of them. From this model we need to other write down what the $ESS_{UR}$ (that is the error sum of squares, unrestricted) is or what the $R^2_{UR}$ (the unrestricted $R^2$) is.

Now, suppose we want to test that $q$ $\beta$'s are JOINTLY insignificant. We then need to run the RESTRICTED model:

$Y = \beta_1 + \beta_2X_2 + \beta_3X_3 + ... + \beta_{k-q}X_{k-q} + \varepsilon$

Note that in the restricted model we are only estimating $(k - q)$ coefficients, while in the unrestricted model we are estimating $k$ coefficients. From the restricted model we need to know the $ESS_R$ (the error sum of squares for the restricted model) or the $R^2_R$ (the restricted $R^2$). We also need to know $q$, which is the number of restrictions. The reason this is the number of restrictions is because we have forced $q$ coefficients to equal zero in the restricted model (by NOT including those independent variables and their coefficients we have imposed that each of those coefficients equals zero). The test statistic is as follows:

$\frac{(ESS_R-ESS_{UR})/q}{ESS_{UR}/(n-k)} \sim F_{q,n-k}$

If our test statistic is greater than the critical value we reject the null – this means that at least one of the coefficients is significantly different than zero.

What is the intuition behind this test? Focus on $ESS_R - ESS_{UR}$. First, note that $ESS_R \geq ESS_{UR}$, and so $ESS_R - ESS_{UR} \geq 0$. Why? Recall that when we add independent variables to our regression model that the $RSS$ (regression sum of squares or explained variation) will NEVER decrease. This means that the error sum of squares from the restricted model must be at least as big as the error sum of squares from the unrestricted model. Suppose that the independent variables we add to the restricted model add NOTHING to the regression sum of squares. Then, $ESS_R = ESS_{UR}$ and $ESS_R - ESS_{UR} = 0$. Intuitively, if neither of these variables adds anything to the regression sum of squares then they should be meaningless in explaining our dependent variable. This gets at the very point of the test. If $ESS_R - ESS_{UR}$ is very low then we will most likely be failng to reject the null hypothesis; if $ESS_R-ESS_{UR}$ is large,

than a larger portion of the variation in the dependent variable is explained by the additional independent variables added in the unrestricted model, and we should reject the null hypothesis.

We can also write this test as a test involving $R^2$. The appropriate test statistic is:

$\frac{\left(R^2_{UR}-R^2_R\right)/q}{(1-R^2_{UR})/(n-k)} \tilde{} F_{q,n-k}$

How does $\frac{(ESS_R-ESS_{UR})/q}{ESS_{UR}/(n-k)} = \frac{\left(R^2_{UR}-R^2_R\right)/q}{(1-R^2_{UR})/(n-k)}$? Start by multiplying $\frac{(ESS_R-ESS_{UR})/q}{ESS_{UR}/(n-k)}$ with $\frac{\frac{1}{TSS}}{\frac{1}{TSS}}$. We should get:

$\frac{\frac{(ESS_R-ESS_{UR})}{TSS(q)}}{\frac{ESS_{UR}}{TSS(n-k)}}$.

Since $\frac{ESS_R}{TSS} = 1 - R^2_R$ and $\frac{ESS_{UR}}{TSS} = 1 - R^2_{UR}$, we get:

$\frac{\left(1-R^2_R\right)-\left(1-R^2_{UR}\right)}{\frac{q}{1-R^2_{UR}}} = \frac{\frac{R^2_{UR}-R^2_R}{q}}{\frac{1-R^2_{UR}}{n-k}}$. Note that this will be positive since $R^2_{UR} - R^2_R \geq 0$.

### 3.1.1 A special case: the significance of the regression test

As was mentioned earlier the test for the significance of the regression is a special case of the test of the joint signficance of coefficients. Recall that we had our F-statistic for the significance of the regression,

$\frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tilde{} F_{k-1,n-k}$

The question is, how do we turn the F-statistic for the joint significance of coefficients test (which is $\frac{\left(R^2_{UR}-R^2_R\right)/q}{(1-R^2_{UR})/(n-k)} \tilde{} F_{q,n-k}$) into the test for the significance of the regression? First, realize that the $R^2$ in the $F_{k-1,n-k}$ test is just $R^2_{UR}$. Then, think about what $k-1$ is: it is the number of restrictions ($q$) in the second test statistic because we are forcing $k-1$ regression coefficients to equal zero. The only question is what happens to $R^2_R$ (the $R^2$ from the restricted model). What would the restricted model be in the test of the significance of the regression? It would by $Y = \beta_1$. What would the $R^2$ of such a regression be? It would be zero, because a constant cannot explain variation in the dependent variable (the constant doesn't change, so how can it explain variation?).

## 3.2 Tests of linear functions of regression coefficients

Economists are fascinated with a concept called constant returns to scale. Constant returns to scale means that if I have a production process and I double all of my inputs I will double my output. If I triple all of my inputs I will triple my output. If I cut all of my inputs in half I will get half the output I had. Hopefully you get the idea.

The typical functional form used for production functions is known as the Cobb-Douglas form. A production function is a Cobb-Douglas production function if it takes the form $Y = e^{\beta_1}(K)^{\beta_2}(L)^{\beta_3}e^{\varepsilon}$. Note that $Y$ is output, $K$ is capital, and $L$ is labor. The error term is given by $e^{\varepsilon}$, and $\beta_1$, $\beta_2$, and

$\beta_3$ are the unknown parameters that we need to estimate. The term $e^{\beta_1}$ is just a constant scaling factor. (Some of you may have seen this in a slightly different form, especially the $e^{\beta_1}$ – I have made the constant term $e^{\beta_1}$ for a reason which should become clear shortly.) One problem that we have is that we cannot estimate this model (using ordinary least squares) in the form that it is in. However, if I take the natural log of both sides I get:

$\ln(Y) = \ln(e^{\beta_1} (K)^{\beta_2} (L)^{\beta_3} e^{\varepsilon})$

Now, I can use the rules of logarithms to break this into:

$\ln(Y) = \ln\left(e^{\beta_1}\right) + \ln\left(K^{\beta_2}\right) + \ln\left(L^{\beta_3}\right) + \ln\left(e^{\varepsilon}\right)$

I can further simplify this to:

$\ln(Y) = \beta_1 + \beta_2 \ln(K) + \beta_3 \ln(L) + \varepsilon$

We now have a linear model that we can estimate using the ordinary least squares (OLS) technique. We can just put this into SAS and get coefficients.

To get back to the point, how do we know that our production function has constant returns to scale? Mathematically, this would occur if $\beta_2 + \beta_3 = 1$ in the production function (which is $Y = e^{\beta_1} (K)^{\beta_2} (L)^{\beta_3} e^{\varepsilon}$). As an example, suppose I double both my capital inputs and my labor inputs. To start with I have:

$Y_1 = e^{\beta_1} (K)^{\beta_2} (L)^{\beta_3} e^{\varepsilon}$

Note that I have added the subscript 1 to $Y$. $Y_1$ is how much output I receive BEFORE I've doubled my inputs. Now suppose that I double my inputs. My new production function will look like:

$Y_2 = e^{\beta_1} (2K)^{\beta_2} (2L)^{\beta_3} e^{\varepsilon}$, where $Y_2$ stands for the output I receive if I have doubled my inputs. We have not proved anything yet, but if we can show that $Y_2 = 2Y_1$, we will show that we have constant returns to scale. Now,

$Y_2 = e^{\beta_1} (2K)^{\beta_2} (2L)^{\beta_3} e^{\varepsilon} = e^{\beta_1} (2)^{\beta_2} (K)^{\beta_2} (2)^{\beta_3} (L)^{\beta_3} e^{\varepsilon}$

This is just using the properties of exponents. Rearranging terms we get:

$Y_2 = e^{\beta_1} (2)^{\beta_2} (K)^{\beta_2} (2)^{\beta_3} (L)^{\beta_3} e^{\varepsilon} = e^{\beta_1} (2)^{\beta_2} (2)^{\beta_3} (K)^{\beta_2} (L)^{\beta_3} e^{\varepsilon}$

Using the properties of exponents again, we get:

$Y_2 = e^{\beta_1} (2)^{\beta_2} (2)^{\beta_3} (K)^{\beta_2} (L)^{\beta_3} e^{\varepsilon} = (2)^{\beta_2+\beta_3} e^{\beta_1} (K)^{\beta_2} (L)^{\beta_3} e^{\varepsilon}$

Now, what do we know? We know that $e^{\beta_1}(K)^{\beta_2}(L)^{\beta_3} e^{\varepsilon} = Y_1$. We can substitute in to get:

$Y_2 = (2)^{\beta_2+\beta_3} Y_1$

When will $Y_2 = 2Y_1$? When $\beta_2 + \beta_3 = 1$. (This is also known as homogeneous of degree one for you mathematics people, and anyone who is interested in going to grad school for economics should be aware that you will be expected to know little things like this by the end of your first semester ( maybe first year) of grad school.)

What test will our null hypothesis be in our regression analysis? It will be $\beta_2 + \beta_3 = 1$. What will our alternative hypothesis be? $H_A : \beta_2 + \beta_3 \neq 1$. What will our test statistic be? Our initial model is:

$\ln(Y) = \beta_1 + \beta_2 \ln(K) + \beta_3 \ln(L) + \varepsilon$

If we impose the restriction that $\beta_2+\beta_3 = 1$, we would then have $\beta_3 = 1-\beta_2$. Substituting this in to our model gives us:

$\ln(Y) = \beta_1 + \beta_2 \ln(K) + (1 - \beta_2) \ln(L) + \varepsilon$

The next few steps are collecting terms and simplifying:
$\ln(Y) = \beta_1 + \beta_2 \ln(K) + \ln(L) - \beta_2 \ln(L) + \varepsilon$
$\ln(Y) - \ln(L) = \beta_1 + \beta_2 \ln(K) - \beta_2 \ln(L) + \varepsilon$
$\ln(Y) - \ln(L) = \beta_1 + \beta_2(\ln(K) - \ln(L)) + \varepsilon$
Since we are working with logarithms we can make this:
$\ln\left(\frac{Y}{L}\right) = \beta_1 + \beta_2 \left(\ln\left(\frac{K}{L}\right)\right) + \varepsilon$
($\frac{Y}{L}$ is the output to labor ratio and $\frac{K}{L}$ is the capital to labor ratio.)
Now, estimate this model.
What type of test will we use? An F-test:
$\frac{(R^2_{UR} - R^2_R)/1}{(R^2_{UR})/(n-k)} \sim F_{1,n-k}$

Although are null hypothesis involves 2 regression coefficients we only have ONE restriction. Why is this? We want $\beta_2 + \beta_3 = 1$. We can let either $\beta_2$ be anything as long as $\beta_3 = 1 - \beta_2$ (alternatively we could allow $\beta_3$ to be anything as long as $\beta_2 = 1 - \beta_3$). For our example, the $R^2_{UR}$ comes from the model $\ln(Y) = \beta_1 + \beta_2 \ln(K) + \beta_3 \ln(L) + \varepsilon$ and the $R^2_R$ comes from the model $\ln\left(\frac{Y}{L}\right) = \beta_1 + \beta_2 \left(\ln\left(\frac{K}{L}\right)\right) + \varepsilon$.

## 3.3 Tests for the equality of coefficients in different regressions

**NOTE**:This test ONLY applies when the dependent variables of the two regression models AND the independent regressors are the same. This means I need to have two regressions, perhaps one for male wages and one for female wages as was suggested above. I would then run:

Using only wages for males ($M$ observations): $wage = \beta_1 + \beta_2 tenure + \beta_3 school + \varepsilon$

Using only wages for females ($N$ observations): $wage = \alpha_1 + \alpha_2 tenure + \alpha_3 school + \varepsilon$

Suppose for the male equation I had $M$ observations and for the female equation I had $N$ observations. So in total there are $M + N$ observations on wages. My null hypothesis is:

$H_0 : \alpha_1 = \beta_1, \alpha_2 = \beta_2, \alpha_3 = \beta_3$

$H_A$ : At least one of those is not equal

**NOTE**: We are still assuming that the error variance is constant between the two models.

This is again going to be an F-test. We have run 2 unrestricted models, so we need a restricted model. The restricted model will be:

Using all wages ($N+M$ observations): $wage = \gamma_1 + \gamma_2 tenure + \gamma_3 school + \varepsilon$

What test will we use?
$\frac{(ESS_R - ESS_{UR})/q}{ESS_{UR}/(N+M-2k)} \sim F_{q,N+M-2k}$, where:

$ESS_R$ is the error sum of squares from the restricted model

$ESS_{UR}$ is the error sum of squares from the male model PLUS the error sum of squares from the female model

$q$ is the number of restrictions imposed (in our example we have imposed 3 restrictions)

$N + M - 2k$ is the number of degrees of freedom in the unrestricted model (we have $N+M$ observations and we are estimating 6 coefficients in our example – 3 $\alpha$'s and 3 $\beta$'s)

If our F-statistic is greater than the critical value of the F-distribution then we reject the null hypothesis, and we assume that we cannot "pool the data". Note that this is very similar to testing the individual significance of the gender dummy and the interaction of the gender dummy in the models above.

# 4 Piecewise linear regression

The piecewise linear regression is useful if you think there is a "structural break" in the data. This is easiest to picture with time series data, although it can happen with cross-sectional data. One example of cross-sectional data that we could use is the example of obtaining different college degrees (Bachelor's, Master's, PhD). Suppose we just want to limit ourselves to looking at people with a Bachelor's degree (say 16 years of schooling or more) versus people without a Bachelor's degree (less than 16 years of schooling). We would hypothesize that receiving a Bachelor's degree causes an increase in the rate at which your wages increase (the slope is steeper for those with a Bachelor's degree). We could just use a dummy variable approach to test this, but it may be more useful to use a continuous function (recall that with the dummy variable approach you would get two different regression lines – one for those with a Bachelor's degree and one for those without). With the piecewise linear regression what you are doing is fitting a CONTINUOUS (though non-differentiable) function through the data, although it is not necessarily a straight line. In fact, it will be pieces of two different lines, connected at one point (the kink point in the function). How do we estimate such a model? Suppose we have data on wages and years of schooling. We create a dummy variable (COLLEGE) and set it equal to one if the person has 16 or more years of schooling, zero otherwise. Our piecewise regression equal would then be:

$wage = \beta_1 + \beta_2 school + \beta_3((school - 16) * college) + \varepsilon$

Using our data set, the output I get is:

$wage = -2.64 + 1.11 school + .97((school - 16) * college)$

What does this mean? Suppose I was someone with 17 years of school. My predicted wage would be: $-2.64 + 1.11 * 17 + .97((17 - 16) * 1) = 17.2$

Suppose I was someone with 12 years of school. My predicted wage would be:

$-2.64 + 1.11 * 12 + .97((12 - 16) * 0) = 10.68$

The main point of the piecewise regression is to keep the function continuous. Continuity is a useful mathematical property, and easy to understand intuitively (a function is continuous if I don't have to lift my pencil off the paper when I am drawing it) but it is a difficult concept to define mathematically. However, suppose we estimate:

$wage = \beta_1 + \beta_2 school + \beta_3 college + \beta_4(school * college) + \varepsilon$

This will also allow the slope and intercept to change for someone who has graduated from college. The results I get are:

$wage = 1.22 + .77school - 13college + .96(school * college)$

Suppose I have someone with 17 years of school now. This person's predicted wage is:

$1.22 + .77 * 17 - 13 * 1 + .96(17 * 1) = 17.63$

The person with 12 years of school has a predicted wage of:

$1.22 + .77 * 12 - 13 * 0 + .96(12 * 0) = 10.46$

They are very similar to the wages estimated with the piecewise model. Now look at someone with 15.99 years of school, 16 years of school, and 16.01 years of school.

- Piecewise
  1. 15.99: $-2.64 + 1.11 * 15.99 + .97((15.99 - 16) * 0) = 15.109$
  2. 16.00: $-2.64 + 1.11 * 16 + .97((16 - 16) * 1) = 15.12$
  3. 16.01: $-2.64 + 1.11 * 16.01 + .97((16.01 - 16) * 0) = 15.131$
- Dummy variable approach
  1. 15.99: $1.22 + .77 * 15.99 - 13 * 0 + .96(15.99 * 0) = 13.532$
  2. 16.00: $1.22 + .77 * 16 - 13 * 1 + .96(16 * 1) = 15.9$
  3. 16.01: $1.22 + .77 * 16.01 - 13 * 1 + .96(16.01 * 1) = 15.917$

Notice that with the piecewise approach we have a fairly smooth transition – 15.109, 15.12, 15.131. With the dummy variable approach we get a big jump – 13.532, 15.9, 15.917. This is the basic idea of continuity – you shouldn't have big jumps. Now, which approach should be used given the problem we have, concerning wages for college graduates? I would go with the dummy variable (non-continuous) approach because there IS (or at least should be) a big difference between college graduate wages and non-college graduate wages. Now, if we thought that there was some magical age that occurred that caused the slope of wages to change then I would suggest using the piecewise approach – there shouldn't be a big jump in wages from the time you are 25.99 years old to the time you turn 26.01 years old (holding all else constant of course).