These notes correspond to chapter 5 of the text. There is some additional information in these notes because I feel it is important that you all have some idea of where these concepts come from.

1 Curve fitting

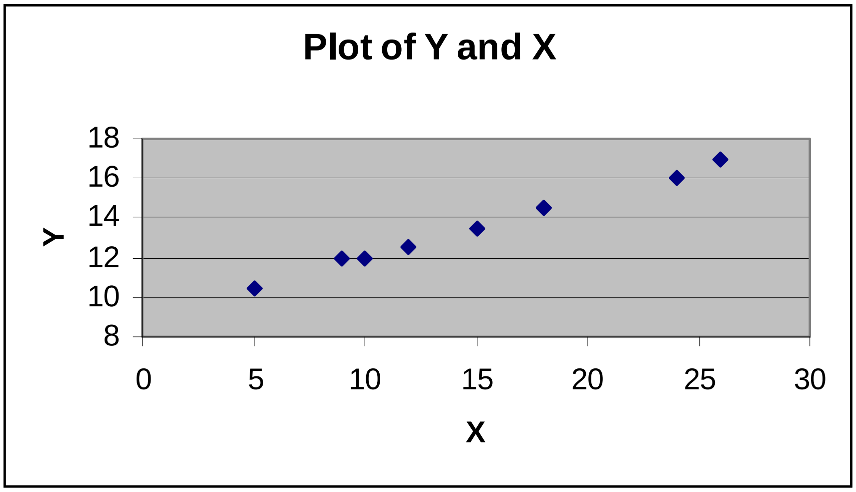
- Types of data

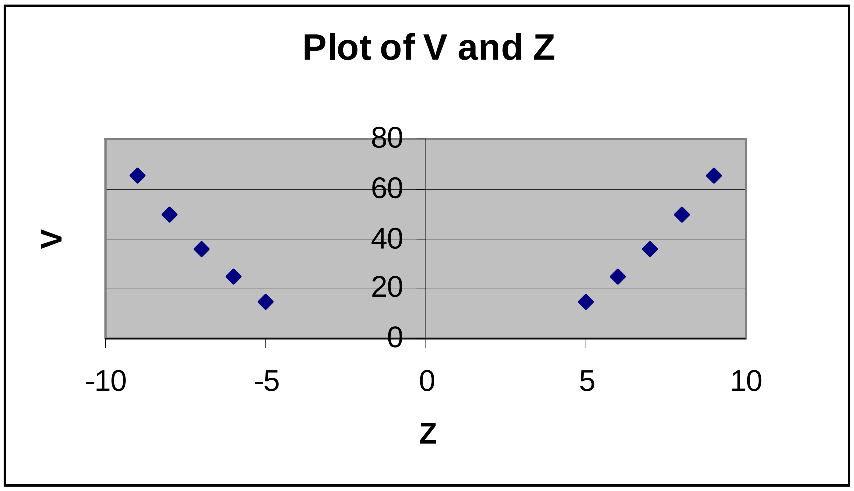
1. time series – describes the movement of a variable over time
2. cross-section – looks at the individual characteristics of firms or individuals at a point in time
3. panel (or pooled or longitudinal) data – combination of time series and cross-section data

Suppose we hypothesize that there is a relationship between two variables, X and Y. Economic theory would tell us what we should expect this relationship to be, but we will use econometrics to estimate this relationship. In order to estimate this relationship, we need data. If we were to have every possible observation on a variable, we would then have the population of that variable. However, most of the time we will have a sample of the available data. Thus, we will need to perform statistical tests to see if the estimates we obtain are significant (we will get to that later in the course). For now, suppose we have two variables, X and Y. I have created some numbers for X and Y and placed them in the table below; the scatterplot shows the graphical relationship between X and Y. I have also created two more variables, V and Z, and created a table and a scatterplot for them as well.

Y	X
13.5	15
12	9
14.5	18
12.5	12
10.5	5
17	26
14.5	18
12	10
16	24

V	Z
49.6	8
24.4	6
64.9	-9
24.4	-6
14.5	5
36.1	7
36.1	-7
64.9	9
49.6	-8
14.5	-5





When we attempt to fit "curves" to the data, what we are actually going to do is attempt to fit a straight line to the data. This is where the term "linear" comes into play in our analysis. In some cases, attempting to fit a straight line through the data may not be the best option. Looking at X and Y above, it seems that a straight line might yield a good approximation of the relationship between X and Y, but it does not look like a straight line (we are only going to try to fit one line to ALL of the data points) will yield a very good approximation of the relationship between V and Z. You should be aware that sometimes a linear approach is NOT the best approach to approximating the relationship between two variables – however, we will focus on linear regression techniques.

1.1 What line to fit?

We can attempt to fit an infinite number of lines through the X and Y data. We could connect the lowest point (5,11) and the highest point (26,17). We could try to draw a line in by hand that looks like it will fit and then try to measure the slope and intercept by hand. However, the method that we will use is the **least squares method**. We use the least squares method to find the line of best fit. The line of best fit is defined as the line which minimizes the sum of the squared (vertical) deviations of the points of the graph from the points of the straight line that we choose. So what we do is draw a line through the data, measure how far the data point is vertically from the line (that is the deviation), and square that value (that is the squared deviation). We do this for each data point and then sum the squared values. This gives us our "sum of squared deviations". What we would then do is draw a different line through the data and find the sum of the squared deviations of that line – if the sum of squared deviations of the second line was lower than the sum of the squared deviations of the first line, then the second line would be a better fit than the first line. Of course, we could draw a third line and repeat the process and see if it has a lower sum of squared deviations than the second line. Hopefully you get the idea.

1.2 Why use least squares?

We could use other methods of trying to find the line of best fit. Two possible alternative criteria are using the sum of the deviation values themselves (NOT squaring them) and using the sum of absolute value of the deviations. When using the sum of the deviations we would want to try to get the sum as close to zero as possible. One reason that we do NOT want to use the sum of the deviations without squaring them has to do with the following example. Suppose we have two data points. The X value of both data points is 10. The Y value of the first data point is 17 and the Y value of the second data point is 7. Clearly, the line that best approximates this relationship is a vertical line at $X = 5$. However, ANY line that passes through the point (5,12) will have a sum of deviations equal to zero, because one data point will be 5 units above the line and the other will be 5 units below the line. Thus it may be possible to find a line that has a sum of deviations equal to zero that does not give a good approximation of the data. As has also been suggested, we could try to minimize the sum of the absolute value of the deviations, as this would eliminate the problem of having two data points cancel each other out (since we are summing only positive values). There are two reasons we do not use absolute value. The first is that using the sum of the absolute value of the deviations puts less weight on a data point that is very, very far away from the line than the least squares method does while it puts more weight than least squares on data points that are fairly close to the line. The second reason that we use the least squares method rather than the absolute value method is mathematical. For those of you that have had calculus, whenever you see "minimize" or "maximize" you should think derivative. If you recall what the absolute value function looks like, it has a kink in it (it is not smooth), which makes it nondifferentiable, which makes it a mess to work with. The least squares method also has the added bonus in that it permits statistical testing of the estimates of the slope and intercept that we will obtain.

2 The Linear Regression Model

After we have looked at our data, we need to propose a model. As I have mentioned, we will work with LINEAR models in this class. Most of our models will look like:

$$Y = \beta_0 + \beta_1 X$$

Notice that when we write down our model, we put Y on the left-hand side and X on the right-hand side. We have, in effect, decided that X has an influence on Y. We call X (or, more generally, the right-hand side variable) the independent variable because we are assuming it is NOT influenced by anything. We call Y the dependent variable because we are assuming that X helps determine Y.

I will not hold you all responsible for knowing the derivation of least squares estimates, but it might provide some of you with some insight so I have provided a link **??** .[1] Keep in mind those are only the estimates for a simple model – the expressions for models with more than a constant and a single independent variable are much more involved (it's easier to use matrix algebra to represent those estimates).

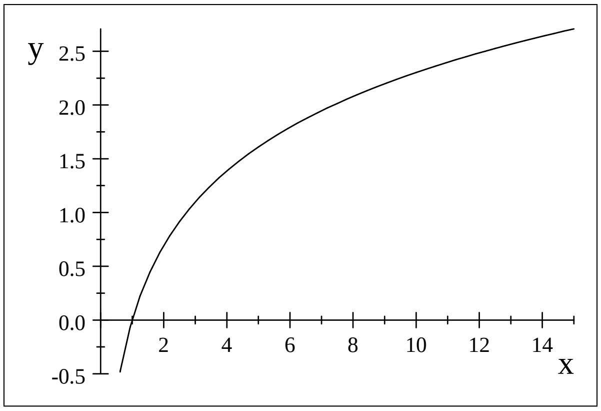
2.1 The General Linear Model

When we talk about the LINEAR regression model, we are talking about a model that is linear in the parameters (the parameters are $\beta_0, \beta_1, \beta_{2,}...$). There are models which are linear in the parameters that can test NONLINEAR relationships between X and Y. Consider the following 3 models:

1. $Y = \beta_0 + \beta_2 X_1 + \beta_2 (X_1)^2 + \varepsilon$
2. $Y = \gamma_0 + \gamma_1 \ln(X_2) + \varepsilon$
3. $\ln(Y) = \alpha_0 + \alpha_1 \ln(X_2) + \varepsilon$

Model 1 is linear in the parameters (the β'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 with (or who may have forgotten) the natural log function, the graph of $Y = \ln(X)$ is:

[1] In the derivation of the least squares estimates the parameter β_0 is represented as α, and the parameter β_1 is represented as β.



This function 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 function looks similar to production functions that we will discuss in chapter 7. If we take logs of both sides we get model 3 above. Some other models are:

- Exponential model: $\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- Reciprocal model: $\frac{1}{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- Interaction model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(X_1 X_2) + \varepsilon$

Again, all of the above models are linear in the parameters, though they may represent nonlinear relationships between Y and X. We can use multiple regression analysis to estimate the parameters for all of these models. Throughout this chapter we will discuss interpretation of these parameters.

3 Interpreting Regression Results

When we have an estimated coefficient we will denote it as $\widehat{\beta}$ (read as "beta hat"). Consider our basic linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$. For the slope coefficient, $\widehat{\beta_1}$, it means that if X increases by one unit, then Y will increase (or decrease if $\widehat{\beta_1}$ is negative) by $\widehat{\beta_1}$ units. If $\widehat{\beta_1} = -1.9$, then this means that a one unit increase in X will cause a 1.9 unit decrease in Y.

For $\widehat{\beta_0}$, think about what the intercept tells you in an equation of a line. It tells you what Y will equal if $X = 0$. There are two notes about the statistical significance of the intercept that you should be aware of. Even if the intercept is NOT statistically significant, we need to have the intercept in the regression equation, otherwise we will be forcing our regression line through the origin, which may not be very accurate. It is more important to have the freedom that the intercept provides than it is too worry about its significance. The second note about the intercept is that even if it IS statistically significant, it may not mean much to us if we do not have a lot of Xs that are close to zero. This last note is important: the intercept is assuming that $X = 0$. If we do not have a lot of observations in which $X = 0$, then it will be difficult for us to make any real claims about the economic importance of the intercept. We will discuss the interpretation of the intercept with some of the sample data sets we have in class.

3.1 Coefficient Estimates for Multiple Regression

Now suppose we have a model like $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. The interpretation of the estimated intercept coefficient, $\widehat{\beta_0}$, is still the same, only now we are assuming both X_1 and X_2 are equal to zero. The interpretation of the estimated slope coefficients, $\widehat{\beta_1}$ and $\widehat{\beta_2}$, are a little more complicated now. Now, while $\widehat{\beta_1}$ still tells us the change in Y given a one-unit change in X_1, we also have other variables in the model. In this case $\widehat{\beta_1}$ tells us the change in Y given a one-unit change in X_1, holding X_2 constant. That last part is key, and it will sometimes lead to confusion when signs do not match up with hypothesized directions.

Example using housing data I have a dataset provided by a former professor at FSU, Tim Sass, that has information on 980 houses sold in Santa Clara county in 1987. We can estimate two basic models to begin:

$$\begin{aligned} SalesPrice_i &= \beta_0 + \beta_1 LivingArea_i + \varepsilon_i \\ SalesPrice_i &= \alpha_0 + \alpha_1 Bedrooms_i + \varepsilon_i \end{aligned}$$

I am using different parameters (α and β) for the different models for clarity. Shortly we will discuss the mechanics of estimating the regression models using Excel and Stata, but for now I will just produce the results. Before viewing the results, we should hypothesize about the direction of the slope coefficients. I believe that both living area (which is measured in square feet) and number of bedrooms (which is simply a count of bedrooms) should be positively related to the price of a house, so that both β_1 and α_1 will be positive.

My estimation provides the following results:

$$\begin{aligned} \beta_0 &= 33903.48 \\ \widehat{\beta_1} &= 98.01 \\ \\ \alpha_0 &= 162476.90 \\ \alpha_1 &= 7057.869 \end{aligned}$$

Now, both coefficient estimates are positive, which is what we hypothesized. The coefficient estimate for living area ($\widehat{\beta_1}$) tells us that each additional square foot of living area yields an extra $98.01 in sales price, while the estimated coefficient for bedrooms ($\widehat{\alpha_1}$) tells us that each additional bedroom provides an additional $7057.87 in sales price. The magnitudes of the coefficients are vastly different, but think about how they are measured – houses might be 2000 or 3000 square feet, while number of bedrooms might reach a maximum at 7 or 8 (maybe 9). The intercepts tell us that if a house has 0 square feet, then it will sell for $33,903.48 (the estimated coefficient $\widehat{\beta_0}$), while if a house has zero bedrooms it will sell for $162,476.90 (the estimated coefficient $\widehat{\alpha_0}$). Neither of those is very reasonable, because the mean living area is 1555 square feet (the minimum is 544), and the mean number of bedrooms is 3.38 (the minimum is 1 – it turns out the maximum number of bedrooms in this data set is 30, which is almost certainly a typo because the house is is only 892 square feet; rule number 1, know your data).

Now, suppose we had the following model:

$$SalesPrice_i = \beta_0 + \beta_1 LivingArea_i + \beta_2 Bedrooms_i + \varepsilon_i$$

What might we hypothesize the sign of β_1 and β_2 would be, based on both intuition and our prior estimation results? It seems both should be positive, right? The estimated coefficients are:

$$\begin{aligned} \widehat{\beta_0} &= 49360.50 \\ \widehat{\beta_1} &= 105.00 \\ \widehat{\beta_2} &= -7778.45 \end{aligned}$$

These results seem a bit odd – the intercept is positive, the coefficient for square feet of living area is positive, but the coefficient for number of bedrooms is negative.[2] How do we explain these results? Is it a bad draw of data? It should not be because we saw "reasonable" results for the single variable regression model. Think about what the $\widehat{\beta_2}$ coefficient tells us – this coefficient represents the effect of adding one more bedroom, but now we are HOLDING LIVING AREA CONSTANT. What does this mean? It means that the house will have smaller bedrooms, which is typically viewed as a negative.

[2] Some might think the data point with 30 bedrooms is causing problems with the estimation. While it likely is, a huge benefit of using statistical packages other than Excel is that observations can be easily removed using various commands. Removing all observations with more than 19 bedrooms (which is only the single observation with 30 bedrooms), the estimated coefficients are $\widehat{\beta_0} = 80371.65$, $\widehat{\beta_1} = 123.4665$, and $\widehat{\beta_2} = -25637.91$. While the coefficient estimates change, the signs do not.

4 Statistical Testing

An additional benefit of multiple regression analysis is that it allows us to perform statistical tests on our estimated coefficients. With the housing data, $\widehat{\beta_1} = 105.00$, but is that estimate statistically different than zero? The same could be asked of $\widehat{\beta_2}$, even though it is a much larger (in absolute terms) number (-7778.45). In this section we will discuss various statistical tests that can be performed on individual coefficients, multiple coefficients, and the entire regression. What we will do is discuss the tests and their meanings – for a more detailed discussion of statistics see the statistics review **??** . Again, I am not going to hold you responsible for the statistics review on any exams, but for those of you who really want to know why we are using certain tests the information is there.

4.1 Testing Individual Coefficient Estimates

When we obtain our estimates for $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$, we do not know how reliable the estimates are, so we need to perform some statistical tests. We can show that:

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \tilde{\ } t_{N-k}$$

$$\frac{\hat{\beta}_2 - \beta_2}{s_{\hat{\beta}_2}} \tilde{\ } t_{N-k}$$

$$\frac{\hat{\beta}_3 - \beta_3}{s_{\hat{\beta}_3}} \tilde{\ } t_{N-k}$$

where $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ are our coefficient estimates; β_1, β_2 and β_3 are our null hypotheses; $s_{\hat{\beta}_1}, s_{\hat{\beta}_2}$ and $s_{\hat{\beta}_3}$ are the standard errors of $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$; and k is the number of independent variables, INCLUDING the constant (intercept) term. In this model, $Wage_i = \beta_1 + \beta_2 Tenure_i + \beta_3 Age_i + \varepsilon_i$, we would have $k = 3$. We can then construct hypotheses about our estimated coefficients. Typically we will want to test the hypothesis that the estimated coefficient is equal to zero, so essentially all we would have to do to calculate the test statistics is take the ratio of the coefficient estimate to the standard error of the coefficient (we have not really discussed how to calculate this but it is the standard deviation of the sampling distribution of a statistic, or an estimate of the standard deviation). HOWEVER, there are times when we want to test a hypothesis other than the estimated coefficient equals zero. In particular, there are times when we want to know if a slope coefficient is equal to 1. In these cases, you need to use the formulas above to calculate the test statistics.

Basic steps for hypothesis testing:

1. Set up the null and alternative hypotheses.

2. Construct your test statistic (remember to take the absolute value for a two-tailed test)

3. Pick a significance level

4. Look up the critical value in the table – remember how to count your degrees of freedom $(N - k)$

5. Reject or fail to reject the null hypothesis

We can use some shortcuts to testing hypotheses. Remember that the table for the t-distribution jumps from 120 degrees of freedom to ∞ degrees of freedom. When we have ∞ degrees of freedom (meaning more than 120), we can use the critical values for the normal distribution, which are 1.96 and 2.57 for the 5% significance level and the 1% significance level, respectively. If the absolute value of our test statistic is greater than 1.96, we can say the estimate is significant at the 5% level. If the absolute value of our test statistic is greater than 2.57, we can say the estimate is significant at the 1% level. While the shortcut method is useful for large data sets, we will need to know how to count degrees of freedom for small data sets. A huge benefit of using statistical software packages (other than Excel) is that the t-statistics for the hypothesis that the coefficient estimate is equal to zero are reported for you, as are the p-values (which are the exact level of significance).

4.1.1 One-tailed vs. two-tailed tests

For a two-tailed test we are testing whether or not the coefficient is significantly different than the hypothesized value. Thus, the coefficient can either be greater or less than the hypothesized value. In general, these are the types of statistical tests that we conduct.

At times, a one-tailed test may be useful. In these situations we are testing whether or not the estimated coefficient is specifically greater than (or specifically less than) the hypothesized value. The most common reason for conducting this test is to determine if an estimated coefficient is positive (greater than zero) or negative (less than zero). What this means is that our region of rejection is in a single tail of the distribution, and not both tails of the distribution.

4.2 Goodness of Fit/Regression Significance

We will discuss a few measures of goodness of fit. All we mean by "goodness of fit" is how well the model does in estimating the dependent variable.

4.2.1 R^2 and Corrected R^2 (or $\bar{R}^2$)

We can measure how well the regression line fits by looking at the residuals that are generated. Recall that the residuals tell how much the actual Y differs from the predicted Y. If the residuals are small, then the regression line is a good fit. If the residuals are large, then the regression line is not as good of a fit. Here's the problem with just looking at the residuals:

Suppose you have residuals that are in the hundreds of dollars. Is this residual small or large? If your dependent variable is measured in millions of dollars they might be small, but if the dependent variable is measured in thousands of dollars then they might be large. So just looking at the residuals will not tell you much because their "largeness" or "smallness" will depend on the units that the dependent variable is measured in.

In order to find a scale-free measure of goodness of fit, we divide the variation in Y into two parts, the explained variation and the unexplained variation. The variation in Y is given by:

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

This is known as the total sum of squares, or TSS, or Total Variation in Y. We can show that this can be decomposed into the residual sum of squares or error sum of squares (ESS), which is the portion of the variation in Y that is UNEXPLAINED by the model, and the regression sum of squares (RSS) which is the portion of variation of Y that is explained by the model.

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$$

The first term on the left side of the equation is the ESS, and the second term is the RSS. Note that if ALL of the variation in Y was explained by the model, then we would have $\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$, or perfect prediction.[3]

Now, we have an equation that breaks the variation in Y into explained and unexplained portion. What we need to do to get rid of the units of measurement (remember that is our goal) is to normalize the variation.

[3] ****IMPORTANT NOTE**** You may see the acronyms ESS and RSS in other sources used in a different way. As I have defined it, ESS is the error sum of squares. But notice that the error sum of squares is also called the **R**esidual variation. Also, as I have defined RSS it is the regression sum of squares. But notice that the regression sum of squares is also called the **E**xplained variation. Notice that I've capitalized and bold-faced the **R** and **E**. Other sources define RSS as the residual sum of squares and ESS as the explained sum of squares. Notice that this is the exact opposite of how I have defined them. The point is, if you look at another source and they are talking about the RSS, make sure that they have defined RSS as the regression sum of squares and NOT the residual sum of squares.

We do this by dividing through by the TSS. So we have our equation as:

$$\begin{aligned}\sum_{i=1}^{N}(Y_i-\bar{Y})^2 &= \sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{N}(\hat{Y}_i-\bar{Y})^2 \\ TSS &= ESS+RSS\end{aligned}$$

Now, divide through by TSS to get:

$$1=\frac{ESS}{TSS}+\frac{RSS}{TSS}$$

Define R^2 as $\frac{RSS}{TSS}$.

R^2 tells us how much of the variation in Y is explained by the regression model that we have estimated. It is unit-free and it will lie between 0 and 1. An $R^2 = 1$ tells us that ALL of the variation in Y is explained. An $R^2 = 0$ tells us that NONE of the variation in Y is explained. Generally, if R^2 is large (close to 1) we say that the model does well in explaining the variation in the dependent variable. If R^2 is small (close to 0) we say that the model does not do well in explaining the variation in the dependent variable. These are just general rules of thumb however. If the model uses time-series data, it is likely that the model will have a high R^2. Why? Because with time-series data most of the variables trend upward over time, so one variable typically "explains" a lot of the variation in the other just because they both increase over time. So R^2 may not be the best measure to use to check how well the model does when using time-series data. One other problem with R^2 is that the regression equation itself may not be significant. We will discuss this concept shortly.

Should our goal be to maximize R^2? While this seems like an appropriate goal, consider the following.

We draw a sample of the dependent variable, Y. The sample that we draw has a specific numerical value for its total variation (or total sum of squares). So let the total sum of squares of our sample of Y be TSS_Y. We know we can break TSS_Y into the portion of the variation explained by the model and the portion of the variation that is not explained by the model. Suppose $TSS_Y = 100$. We estimate a regression model with one independent variable, X_1. We find that the RSS of the model is equal to 20 with just the X_1 variable. In this simple model, $R^2 = .2$. Now, suppose we want to add another independent variable to our model, X_2. Suppose that X_2 has very little to do with Y. The question is, will the new model $Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \varepsilon$, explain LESS than the old model, $Y = \beta_1 + \beta_2 X_1 + \varepsilon$? The answer is no. The new model will explain at least 20% of the variation in Y, due to the fact that X_1 is included in the model. If X_2 has ZERO effect on Y, then it will explain ZERO of the variation in Y, which means that we will still only be explaining 20% of the model. We can never explain less of the variation in Y by adding more independent variables. So if our goal was to maximize R^2, then we would use what is called the "kitchen sink approach". The kitchen sink approach means we throw in every single variable that we can find and this will maximize R^2 because adding additional regressors to the model can NEVER lower the amount of explained variation in the model, and is likely to increase it (at least minimally).

Because R^2 can never decrease when we add additional regressors, we need a method of adjusting our "goodness of fit" when we are using multiple regression models. The statistic that we will use is called corrected R^2 or $\bar{R}^2$. We define $\bar{R}^2$ as: $\bar{R}^2 = 1 - \frac{\hat{V}ar(\varepsilon)}{\hat{V}ar(Y)}$, where $\hat{V}ar(\varepsilon)$ is the estimated error variance and $\hat{V}ar(Y)$ is the estimated variance of Y. Notice that this looks very similar to our definition of R^2. Recall that $R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$. We can show that R^2 is related to $\bar{R}^2$, but we will not get into those details. We have a simple formula for computing $\bar{R}^2$ that uses just the R^2 from the regression model, the number of observations, and the number of regressors:

$$\bar{R}^2 = R^2 - \left(\frac{k-1}{n-k}\right)\left(1-R^2\right)$$

Note the following:

1. If $k = 1$, $R^2 = \bar{R}^2$ (this is easy to see)
2. If $k > 1$, $R^2 \geq \bar{R}^2$ (this is a little more difficult to see)
3. $\bar{R}^2$ can be negative (also a little more difficult to see)

Again, we will not go into all the details. The main idea is that by adding unimportant regressors, $\bar{R}^2$ gets penalized more heavily. In other words, suppose you had two regression models for sales price of a home. In one regression model you included living area in square feet and the lot size that the home was on. Both of those likely influence the sales price of the home. Now suppose the second model had living area in square feet and the length of the hair of the purchaser (I'm trying to guess at something that we would think is ridiculous in determining home price). The penalty for including the hair length variable (assuming there is no relationship between hair length and sales price) would be greater than that of including lot size (assuming that lot size is important in determining sales price). I will say the same thing over and over: most statistical packages (other than Excel) report $\bar{R}^2$ so that you do not have to calculate it manually.

Final note on R^2 We can only use R^2 to compare models that have the exact same independent variable. That is, suppose we had 2 models, where one model used Y as the independent variable and the other model used $\ln Y$. Although it seems like we are using the same variable (after all, $\ln Y$ is just a transformation of Y), the models will have different total sums of squares and will NOT be comparable.

4.2.2 F-tests

We would like to know if the regression model is statistically significant. We can perform a statistical test to answer this question. Formally, we are testing:

$H_0 : \beta_2 = \beta_3 = \beta_4 = ... = \beta_k = 0$

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

Note that we do not include the intercept in our null hypothesis, meaning that we are testing to see if $k-1$ coefficients are equal to zero. What we wish to test is that all the regression coefficients are JOINTLY equal to zero. This is different than looking at each parameter estimate and seeing if it is (individually) different than zero, so we need a different statistical test than the t-test. The primary reason for this test might be called "I don't like my t-statistics." What I mean by that is that you may estimate a regression model and see that many (or all) of your estimated coefficients are individually INsignificant. However, this result does not mean the regression model is useless, as the independent variables may be jointly significant (which likely means that you have multicolinearity in your model).

The statistical test that we use is an F-test. We calculate our F-statistic as:

$$\frac{\frac{RSS}{k-1}}{\frac{ESS}{N-k}} \sim F_{k-1,N-k}$$

Alternatively, we could write:

$$\frac{\frac{R^2}{k-1}}{\frac{1-R^2}{N-k}} \sim F_{k-1,N-k}$$

You should convince yourself that you will obtain the same F-statistic regardless of which formula you use to calculate it.

Why do we use the F-distribution? We can show that our F-statistic is the ratio of 2 independent χ^2 random variables to their respective degrees of freedom.

To finish the test you just need to look up the critical value in the table in the back of the book. If your F-statistic is greater than the critical value then you reject the null hypothesis. As for choosing a significance level, realize that there are only tables in the back of the book for the 1% and 5% levels, so those are the only 2 significance levels you can test at unless you want to find tables for the other significance levels. Once again, if the F-statistic that you calculated was greater than the critical value, you reject the null hypothesis and conclude that at least one β is significant at the chosen significance level. Again, most packages (other than Excel) will give you the exact level of significance for this F-test.

Other F-tests Note that the F-test described above is simply a special case in which we are testing that all regressors are jointly equal to zero. However, we can also conduct F-tests to determine if a subset of regressors is jointly significant.

Suppose we wanted to include only independent variables that were significant at the 5% level in our regression model. One possible 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.

Suppose we have the following model:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon$$

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 β'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 β's are JOINTLY insignificant. We then need to run the RESTRICTED model:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_{k-q} X_{k-q} + \varepsilon$$

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:

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

An alternative test statistic, using R^2, is:

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

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 failing 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.

Again, the F-test for the significance of the regression is just a special case of the general F-test. In testing the significance of the regression, we just have $q = k-1$ because we are imposing that all coefficients are equal to zero, and $R^2_R = 0$ because there are no independent variables, so we are not explaining any of the variation in the dependent variable.

5 Dummy (or binary) Independent Variables

This section is not covered in the text but (1) it is fairly easy to understand and (2) could be quite useful.

Up until now we have focused strictly on using quantitative variables (price, 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 will not cover in this class), 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, in a regression to determine the impact of various factors on the sales price of a home, whether or not the home comes with a swimming pool impact the sales price, even controlling for all other factors. The question is how do we incorporate this knowledge about a swimming pool without having any quantitative measure of the pool (for instance, how large it is)?

We will consider a different set of data, using individual wages, for our discussion of dummy variables. Some important factors that impact an individual's wage might be age, years employed at current job, years of schooling, years of work experience in the field, and gender. Most of these variables are quantitative, but how do we incorporate a qualitative variable such as gender? 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:

$$wage = \beta_0 + \beta_1 age + \beta_2 tenure + \beta_3 school + \beta_4 experience + \beta_5 gender + \varepsilon$$

There is a slight problem. When we calculate our regression coefficients we have some formulas. Consider the intercept, $\beta_0 = \bar{Y} - \beta_1\bar{X}_1 - \beta_2\bar{X}_2 - \beta_3\bar{X}_3 - \beta_4\bar{X}_4 - \beta_5\bar{X}_5$. 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 transformation we create what are known as dummy variables. A dummy variable for gender (also called a binary variable) is just a variable that takes on the value 1 for one category and 0 for the other category (it could be 1 if gender is male and 0 if gender is female, or 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 expected value of the dependent variable (Y) if $X_2 = 1$? It is $\beta_1 + \beta_2$. And if $X_2 = 0$, then the expected value of Y is just β_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 β_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.

5.1 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.[4] 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

[4] Technically the estimation of regression coefficients involves inverting a matrix. With perfect colinearity, the matrix is not full column rank (there is some perfect linear relationship between two or more independent variables), and is thus not invertible. Many statistical packages won't let you estimate a model with perfect colinearity, and instead will choose a variable to omit for you (if you included both a male and female dummy variable it would exclude one of those).

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.

5.2 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:

$$\begin{aligned} \text{Model 1} &: \quad wage = \beta_1 + \beta_2 tenure + \beta_3 male + \varepsilon \\ \text{Model 2} &: \quad wage = \gamma_1 + \beta_2 tenure + \gamma_3 female + \varepsilon \end{aligned}$$

First, if it is true that all of your "gender" observations are male and female, then your coefficient on tenure (β_2) will remain unchanged. However, the intercepts (β_1 and γ_1) and the coefficients on your dummy variables (β_3 and γ_1) will change, but in a predictable fashion. The estimate that you get for β_3 will be the exact same as your estimate for γ_3, except that it will have the OPPOSITE sign. The intercept in model 1 (β_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 (γ_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 (β_1): In model 1 the intercept tells us how much a FEMALE worker with zero tenure will earn.

Model 1, coefficient on male (β_3): In model 1 this coefficient tells us how much more (or less) a MALE will earn 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 (γ_1): In model 2 the intercept tells us how much a MALE worker with zero tenure will earn.

Model 2, coefficient on female (β_3): In model 2 this coefficient tells us how much more (or less) a FEMALE will earn 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.

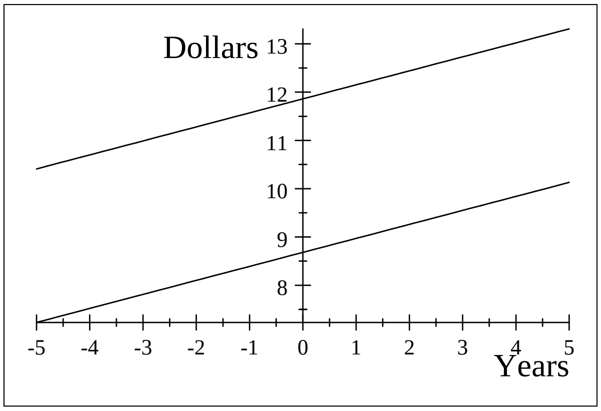
The main point is that the coefficient on the dummy variable tells us how much more (or less) the people with the characteristic captured by the dummy variable earn with respect to the group that has NOT been included as a dummy variable.

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 (because 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).

5.3 Other uses for dummy variables

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

5.3.1 Dummy variables as interaction terms

We can also make dummy variables act as interaction terms. Suppose we have the following model:

$$wage = \beta_1 + \beta_2 tenure + \beta_3(male * tenure) + \varepsilon.$$

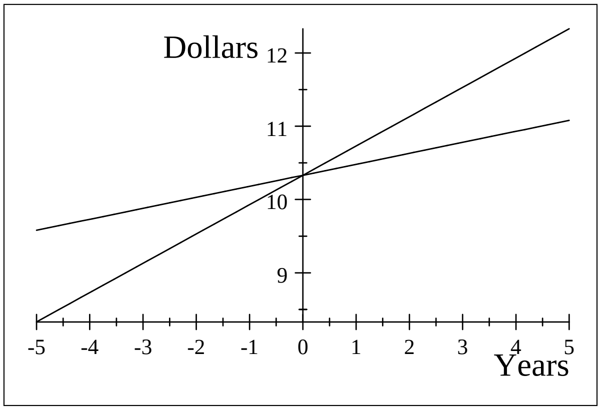
Now, if $male = 1$, the equation becomes: $wage = \beta_1 + \beta_2 tenure + \beta_3(tenure) + \varepsilon$ which is the same as $wage = \beta_1 + (\beta_2 + \beta_3)tenure + \varepsilon$. So we are allowing the slope coefficient to change. If $male = 0$, the equation becomes: $wage = \beta_1 + \beta_2 tenure + \varepsilon$. So the slope coefficient for males is equal to $\beta_2 + \beta_3$ while the slope coefficient for females is equal to β_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.

5.3.2 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:

$$wage = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) tenure + \varepsilon$$

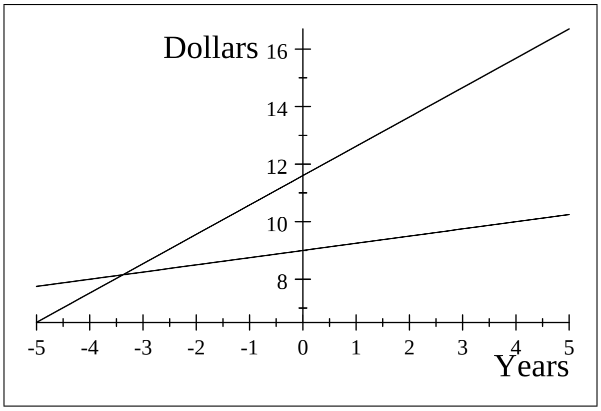
So the intercept for males becomes $\beta_1 + \beta_3$ and the slope becomes $\beta_2 + \beta_4$. For females, we have: $wage = \beta_1 + \beta_3 tenure$, so the intercept is β_1 and the slope is β_3.

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 β_3 was not statistically different than zero we could conclude that the intercept for male and female wages is the same. If β_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.

As mentioned, dummy variables are an easy way of allowing slope and intercept coefficients to vary for different groups. Also, they are easy to implement and interpret.