Chapter 3 outline, Econometrics

Recall that in chapter 1 we began working with the least squares procedure. We hypothesized a linear relationship between our X and Y variables, and then we derived the estimators for a and b, our intercept and slope coefficients. We then did a detour into the world of statistics in chapter 2, which should prove useful for the rest of the course. We want to return to the least squares procedure now.

1 The two-variable regression model

Once again we hypothesize a linear relationship between X and Y. Formally, we would write our regression model as:

$Y_i = \alpha + \beta X_i + \varepsilon_i$, where Y_i is our observation on Y, X_i is our corresponding observation on X, α is our intercept coefficient, β is our slope coefficient, and ε_i is our random error term that has some probability distribution

1.1 The error term

We need the error term in the regression for a few reasons.

1. We will never include all of the relevant variables in our regression model; as such, our model is misspecified and contains some errors
2. There is error in measurement of the variables

1.2 Assumptions of the classical linear normal regression model (CLNRM)

1. The relationship between X and Y is linear
2. The Xs are nonstochastic variables (meaning they have a fixed value)
3. $E[\varepsilon] = 0$
4. $Var[\varepsilon] = \sigma^2$; the error term has constant variance
5. $E[\varepsilon_i \varepsilon_j] = 0$ for all $i \neq j$; this means the error terms are independent
6. The error term is normally distributed

Note that we can change most of these assumptions – it just means that we have to alter our regression procedure slightly. We will alter assumption 3 below. If our error term is NOT normally distributed it complicates statistical testing of the model – however, if we have the first 5 assumptions we still have the classical linear regression model (notice that the word normal is dropped).

1.2.1 Altering assumption 3

Suppose $E[\varepsilon] = \lambda$

The question is, can we find a “different” regression equation that satisfies $E\left[\varepsilon\right] = 0$? Yes.

Our initial regression equation is:

$Y_i = \alpha + \beta X_i + \varepsilon_i$, where $\varepsilon_i\tilde{}N(\lambda, \sigma^2)$

Assume all assumptions other than number 3 hold. How can we rewrite this equation so that it satisfies assumption 3?

Suppose that we define a new error variable, γ, as $\gamma_i = \varepsilon_i - \lambda$. It is easy to verify (and you should do this) that $\lambda_i\tilde{}N(0, \sigma^2)$.

Now, subtract λ from both sides of the regression equation. We have:

$Y_i - \lambda = \alpha + \beta X_i + \varepsilon_i - \lambda$

But $\varepsilon_i - \lambda = \gamma_i$, so:

$Y_i - \lambda = \alpha + \beta X_i + \gamma_i$, which gives us an error term that has mean zero. However, we still have the extra λ hanging around. What do we do with it?

$Y_i = (\alpha + \lambda) + \beta X_i + \gamma_i$

Now, let $(\alpha + \lambda) = \delta$

$Y_i = \delta + \beta X_i + \gamma_i$ – notice that this model is in the same form as $Y_i = \alpha + \beta X_i + \varepsilon_i$, only now we have the expected value of our error term as zero. What this means is that if our error term has $E\left[\varepsilon\right] \neq 0$, all that happens is the mean of the error term gets “consumed” (I can’t think of a better word right now) by the intercept. The slope estimate will still be unbiased, but now the intercept term will be biased.

1.2.2 The other assumptions

We will discuss how to handle problems with the other assumptions throughout the course. There are a few details to be aware of however. If assumption 4 is violated, it means that our model is heteroscedastic (heteroscedastic just means the error variance is not constant). If we were to plot our error terms, we would expect to see a funnel shaped plot. If assumption 5 is violated, it means that we have serial correlation (this just means that when one error term occurs it influences the outcome of the next error term) in our model. If we were to plot our regression line and the error terms, positive serial correlation would be shown as error terms below the regression line (usually) being followed by error terms below the line and the opposite for error terms above the regression line. If we had negative serial correlation, this would be shown as error terms below the regression line (usually) being followed by error terms above the regression line. We have also assumed that $E[X_i\varepsilon_i] = X_i E[\varepsilon_i] = 0$ because we assumed the X_is are nonstochastic (nonrandom).

You should also note that by assuming $\varepsilon\tilde{}N$, we have assumed that $Y\tilde{}N$. If we take the expected value of Y, we get:

$E\left[Y\right] = E\left[\alpha + \beta X + \varepsilon\right] = \alpha + \beta X + E\left[\varepsilon\right] = \alpha + \beta X$

Also, the variance of Y is constant:

$Var\left[Y\right] = Var\left[\alpha + \beta X + \varepsilon\right] = Var\left[\varepsilon\right] = \sigma^2$

Think about why the variance of Y is equal to σ^2. (What expectations operator rule are we using?)

2 Best Linear Unbiased Estimation (BLUE)

When we run our regression models, we will estimate α and β. The estimates we obtain for α and β are RANDOM VARIABLES. They have a probability distribution, and the value we obtain from the regression procedure depends on the sample of Xs and Ys that we draw. We will write the estimators of α and β as $\hat{\alpha}$ and $\hat{\beta}$.

Recall that we estimate $\hat{\alpha}$ and $\hat{\beta}$ by using the least-squares procedure in chapter 1. The next question is, how good are our least-squares estimators of $\hat{\alpha}$ and $\hat{\beta}$? Well, we have a theorem called the Gauss-Markov theorem that states:

Theorem 1 *Gauss-Markov Theorem: Given assumptions 1–5, the estimators $\hat{\alpha}$ and $\hat{\beta}$ are the most efficient linear unbiased estimators of α and β.*

Recall that if an estimator is the most efficient estimator then that means the estimator has the lowest variance of all unbiased estimators. As a result, this is a powerful theorem.

So what do we do if assumptions 1–5 do not hold? We will attempt to modify our regression equation so that they do hold. This is basically the main point of the rest of the course (along with actually running some regressions and interpreting results). We have a very powerful theorem. Sometimes we don't have all the assumptions the theorem requires. How do we change our model so that we have these assumptions? That's basically it. We have already seen that one of the assumptions, number 3, is not that important, and we have already derived a modified version of the model that satisfies assumption 3.

In chapter 1 we saw what our estimators of the intercept and slope coefficients are.

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$

$\hat{\beta} = \frac{\sum_{i=1}^{N}\left[\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)\right]}{\sum_{i=1}^{N}\left[\left(X_i-\bar{X}\right)^2\right]}$

We can show that both $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators. That is, $E\left[\hat{\alpha}\right] = \alpha$ and $E\left[\hat{\beta}\right] = \beta$. We can also show that $Var\left(\hat{\beta}\right) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{N}\left[\left(X_i-\bar{X}\right)^2\right]}$ and $Var\left(\hat{\alpha}\right) = \sigma_\varepsilon^2\frac{\sum_{i=1}^{N}\left(X_i^2\right)}{N\sum_{i=1}^{N}\left[\left(X_i-\bar{X}\right)^2\right]}$. So,

$\hat{\alpha}\tilde{}N\left(\alpha, \sigma_\varepsilon^2\frac{\sum_{i=1}^{N}\left(X_i^2\right)}{N\sum_{i=1}^{N}\left[\left(X_i-\bar{X}\right)^2\right]}\right)$ and $\hat{\beta}\tilde{}N\left(\beta, \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{N}\left[\left(X_i-\bar{X}\right)^2\right]}\right)$

We still have one more thing to estimate. Our variances for $\hat{\alpha}$ and $\hat{\beta}$ both contain σ_ε^2, which is the error variance. When we wrote down our regression equation we assumed the error variance was constant, but we did not make an assumption about the numerical value of the error variance. So we need to estimate the error variance. We can estimate the error variance as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\left(\hat{\varepsilon}_i^2\right)}{N-2} = \frac{\sum_{i=1}^{N}\left(Y_i-\hat{\alpha}-\hat{\beta}X_i\right)^2}{N-2}$$

Note that $\hat{\varepsilon}_i$ is called the residual of the regression. It is equal to $Y_i - \hat{Y}_i$, where $\hat{Y}_i$ is the predicted value of Y_i. Also note that we divide by $N-2$. We do this because this gives us an unbiased estimator. An intuitive way to think about this is that we N observations or N degrees of freedom. However, we need to estimate two pieces of information ($\hat{\alpha}$ and $\hat{\beta}$) so we lose 2 degrees of freedom. We can now obtain estimates of the variance of $\hat{\alpha}$ and $\hat{\beta}$ by inserting the estimated error variance, $\hat{\sigma}_{\varepsilon}^2$, in for the true error variance, σ_{ε}^2.

3 Hypothesis testing and confidence intervals

Our goal now is to use our knowledge of hypothesis testing and confidence intervals to determine how "good" our estimates of α and β are. The steps for testing hypotheses about α and β are identical to the steps for testing hypotheses with unknown variances in chapter 2. Start by setting up your hypothesis. Construct your test statistic. Choose a significance level. Check the table in the back of the book. Fail to reject or reject the null hypothesis.

When setting up our hypotheses about regression coefficients we will typically set up the null hypothesis as, $H_0 : \beta = 0$ and the alternative hypothesis as $H_A : \beta \neq 0$. There is a reason for this particular choice of the null hypothesis. Our regression equation is: $Y = \alpha + \beta X + \varepsilon$. If we cannot reject the null hypothesis of $\beta = 0$, then we may as well say that Y does not depend on X because by NOT rejecting the null hypothesis we are saying that the estimate for β is NOT statistically different than zero.

Next, we construct our test statistic. Our test statistic will be: $\left|\frac{\hat{\beta}-\beta}{\hat{\sigma}_{\hat{\beta}}}\right|$ (we will perform two-tailed tests), where $\hat{\beta}$ is the estimated value of β, β is the null hypothesis value (usually zero), and $\hat{\sigma}_{\hat{\beta}}$ is the standard error of $\hat{\beta}$ (this is just the square root of the estimated variance of $\hat{\beta}$). This test statistic will be distributed t_{N-2}. Why will it be t_{N-2}? Intuitively, we can say that we are estimating two parameters, α and β, and thus we lose 2 degrees of freedom. However, we know that a statistic has the t distribution if it is equal to a standard normal divided by the square root of a chi-square divided by its degrees of freedom. We can show that:

$$\frac{\frac{\hat{\beta}-\beta}{\sigma_{\varepsilon}}}{\sqrt{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}} \,\tilde{}\, N\left(0,1\right), \text{ and } \frac{(N-2)\hat{\sigma}_{\varepsilon}^2}{\sigma_{\varepsilon}^2} \,\tilde{}\, \chi_{N-2}^2$$

Then,

$$\frac{\frac{\frac{\hat{\beta}-\beta}{\sigma_{\varepsilon}}}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2}}}{\sqrt{\frac{\frac{(N-2)\hat{\sigma}_{\varepsilon}^2}{\sigma_{\varepsilon}^2}}{N-2}}}\,\tilde{}\,t_{N-2}$$

It's actually fairly simple algebra. I'll rewrite this as:

$$\frac{\frac{\hat{\beta}-\beta}{\sigma_{\varepsilon}}}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2}} \div \sqrt{\frac{\frac{(N-2)\hat{\sigma}_{\varepsilon}^2}{\sigma_{\varepsilon}^2}}{N-2}} = \frac{\left(\hat{\beta}-\beta\right)\sqrt{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}{\sigma_{\varepsilon}} \div \frac{\hat{\sigma}_{\varepsilon}}{\sigma_{\varepsilon}}$$

Now,

$$\frac{\left(\hat{\beta}-\beta\right)\sqrt{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}{\sigma_{\varepsilon}} \div \frac{\hat{\sigma}_{\varepsilon}}{\sigma_{\varepsilon}} = \frac{\left(\hat{\beta}-\beta\right)\sqrt{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}{\sigma_{\varepsilon}} * \frac{\sigma_{\varepsilon}}{\hat{\sigma}_{\varepsilon}} = \frac{\left(\hat{\beta}-\beta\right)\sqrt{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}{\hat{\sigma}_{\varepsilon}}$$

This looks very similar to our t-statistic, except for two things. The first is easy to spot: our t-statistic does not have $\sqrt{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}$ in the numerator. The second is a little more subtle. Notice that the estimated standard deviation (or the standard error) in the denominator of the formula that I just derived is $\hat{\sigma}_{\varepsilon}$, the estimate of the standard deviation of the error term. However, in the t-statistic the estimated standard deviation (or the standard error) in the denominator is $\hat{\sigma}_{\hat{\beta}}$, which is the standard error of $\hat{\beta}$. However, note that:

$\frac{\sqrt{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}{\hat{\sigma}_{\varepsilon}} = \frac{1}{\hat{\sigma}_{\hat{\beta}}}$, since $\hat{\sigma}_{\hat{\beta}}^2 = \frac{\hat{\sigma}_{\varepsilon}^2}{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}$, so we can now substitute in $\frac{1}{\hat{\sigma}_{\hat{\beta}}}$ to get our t-statistic.

We can perform tests of hypotheses on α in the same manner, only we use:

$\left|\frac{\hat{\alpha}-\alpha}{\hat{\sigma}_{\hat{\alpha}}}\right|\tilde{}\,t_{N-2}$

Now, our third step in testing hypotheses is to choose a level of significance, usually 5% or 1%.

Our fourth step is to find the critical value in the table that corresponds to our chosen level of signficance and the number of degrees of freedom of our test statistic ($N-2$ in the examples above).

The fifth step is to fail to reject the null hypothesis if our test statistic is less than the critical value from the table or reject the null hypothesis if our test statistic is greater than the critical value from the table. Typically we would like to be able to reject the null hypothesis that we formulate because we want our slope coefficient to be statistically different than zero.

As an example, suppose we estimate $\hat{\beta}$ as -1.9. Also suppose the standard error (this is just the term used for the estimated standard deviation) of $\hat{\beta}$ is 0.82. Finally, we need to know the number of observations. Let $N = 209$. We want to test:

$H_0 : \beta = 0$

$H_a : \beta \neq 0$

The test statistic is: $\left|\frac{\hat{\beta}-\beta}{s_{\hat{\beta}}}\right|\tilde{}\,t_{N-2}$

Substituting in our values for $\hat{\beta}$, β, and $s_{\hat{\beta}}$, we get: $\left|\frac{-1.9-0}{.82}\right| = \left|\frac{-1.9}{.82}\right| = 2.$ 3171 1

Suppose we want to test at the 1% level.

Now, look at the t-table for the critical value. Look down the .01 column and look across the row for ∞ degrees of freedom. The critical value that you should see is 2.576. This should look very familiar, as the critical value for the 1% level of the normal distribution is 2.57. Recall that when the degrees of freedom for the t-distribution get large the t-distribution approximates the normal.

Since our test statistic is less than the critical value, we fail to reject the null hypothesis at the 1% level. However, if we wanted to test at the 5% level, the critical value for the 5% level is 1.96. Since our test statistic is greater than the critical value at the 5% level, we reject the null hypothesis at the 5% level. If we wanted to we could also look at the 2% level of significance (since it is given in the table). The critical value (2.326) is slightly greater than the test statistic, so we fail to reject at the 2% level.

4 Interpretation of the $\hat{\alpha}$ and $\hat{\beta}$[1]

The book doesn't really go into detail about how to interpret $\hat{\alpha}$ and $\hat{\beta}$, so I've decided to add these notes. I did not discuss hypothesis testing for $\hat{\alpha}$, but the process is the same as it is for $\hat{\beta}$. There is one thing you should look at before you attempt to interpret your results, and that is the significance of the coefficient ($\hat{\alpha}$ or $\hat{\beta}$). If the coefficient is NOT significant, don't waste your time trying to explain what it means. If it is not significant, it is not significant. However, you may want to explain WHY you think the coefficient is not significant. There could be any number of reasons why the coefficient is not significant. Two possibilities are the variable that the coefficient corresponds to is not as important as you thought, or you just happen to have a "bad" data set where the expected relationship between X and Y doesn't hold.

As for interpretation of the results, think about what $\hat{\beta}$ and $\hat{\alpha}$ are: $\hat{\beta}$ is the slope and $\hat{\alpha}$ is the intercept. So what does $\hat{\beta}$ mean? It means that if X increases by one unit, then Y will increase (or decrease if $\hat{\beta}$ is negative) by $\hat{\beta}$ units. If $\hat{\beta} = -1.9$, as it does in the example above, then this means that a one unit increase in X will cause a 1.9 unit decrease in Y.

For $\hat{\alpha}$, 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

[1] These are notes I've added that the book doesn't have.

mean much to us if we do not have a lot of Xs that are close to zero. We will see this in the example with the housing data set in class.

5 Analysis of variance and goodness of fit

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. What we will do now is add zero to the term in brackets. This gives us: $\sum_{i=1}^{N}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$. Now, we FOIL (recall foiling from algebra).

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

We can show that the last term on the RHS is equal to 0. That is:

$\sum_{i=1}^{N} 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$, which means:

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

We already know that $\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ is the total variation (or total sum of squares) in Y. The first term on the RHS, $\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$, is known as the residual variation in Y. It is also called the "error sum of squares" or ESS. Notice what $\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$ is; it is just the sum of the squared residuals, since $Y_i - \hat{Y}_i = \hat{\varepsilon}_i$. This is the portion of the variation in Y that is UNEXPLAINED by the model. The second term on the RHS, $\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$, is known as the explained variation in Y. It is also called the "regression sum of squares" or RSS. Take a closer look at the term. It shows how much $\hat{Y}_i$ deviates from the mean of Y. This is the portion of the variation in Y that is EXPLAINED by the model. 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.

****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.

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

$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$, or if we write it in acronym form, $TSS = ESS + RSS$

Now, divide through by TSS to get:

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

Define R^2 as $\frac{RSS}{TSS}$ (alternatively we could say that $R^2 = 1 - \frac{ESS}{TSS}$)

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. See the next section.

5.1 Testing the regression equation

We can perform a statistical test to see if the estimated relationship between X and Y is statistically significant. This test is not particularly useful when we have only one independent variable because we can use the t-test to determine the significance of the coefficient that corresponds to our independent variable. However, when we add more independent variables we may have some variables that are not significant, some that are marginally significant, and some that are highly significant. In this case, we would need to look at the test statistic for the regression equation. This statistic will have the F-distribution, and I will introduce it in chapter 4. The important point to note is that if your R^2 is high but your regression equation is not significant, then you shouldn't put much faith in the value that you obtained for R^2. However, if you have a low R^2 but it is statistically significant, then this suggests that the independent variable(s) that you have included in your regression equation help explain the variation in the dependent variable, but that other independent variables may help more.

5.2 Correlation and Causation

A final note concerns the notions of correlation and causation. We know from principles of economics that association (correlation) does not imply causation. Two variables may be correlated but one neither one may explain much of the variation in the other if there is a third variable that actually causes both of the other variables. R^2 seems to be a measure of correlation, but it really is a measure of causation. When we write down our regression equation, we have implicitly assumed that the independent variables cause the dependent variables. You should note that if we were to let Y be the INDEPENDENT variable and X be the DEPENDENT variable (usually it is the other way around), you would get different estimates for the slope and intercept.