Homework 18: ECO220Y – SOLUTIONS

Required Problems:

(1) (a) Because these data are observational the interpretation of the OLS estimation results must be descriptive: describing patters that we see in these data. In this case a value of zero is basically within the range of the x-variable so we can interpret the intercept. On average OECD countries with no debt in 2003 had debt of about 3.66% of GDP in 2009. The slope we interpret descriptively. OECD countries with debt that is 1 additional percentage point of GDP in 2003 on average have debt that is 1.08 additional percentage points of GDP in 2009. In other words, countries with high debt in 2003 tended to have high debt in 2009: in the data we see that each extra percentage point of debt in 2003 is on average translating into an extra 1.08 percentage points of debt in 2009. The R^2 is 0.83 which is quite high indicating a strong relationship: 83 percent of the variation across OECD countries' debt in 2009 can be explained by variation in those countries' debt in 2003.

(b) No, that is not a coincidence. In a simple regression (i.e. a regression with only one x variable), the coefficient of determination (R^2) is simply the coefficient of correlation squared. (Note: Later this term we will look at multiple regression where we have multiple x variables and there this link necessarily breaks down because there are many correlations and not just one.) The interpretation of 0.91: countries with debt in 2003 that is one standard deviation higher have debt in 2009 that is 0.91 standard deviations higher on average. Hence there is some regression to the mean (i.e. not a perfect correlation), but the correlation is very strong: a country's debt in 2003 is a very good predictor of its debt in 2009. Still confused: review Sections 6.3 and ***especially*** Section 7.2.

(c) Review your old lecture notes for the calculations (i.e. where we just plugged into formulas). These data are experimental because the researcher randomly assigned each of the 25 patients a dosage of the drug (in milligrams) then recorded how much they slept. I.e. the x variable is randomly set and we see what happens to y. Because x is random it cannot be correlated with any other variables that also affect y (random by definition is not related to anything). Hence we have no concerns about lurking variables and endogeneity bias. Any upward pattern we see in the data could only reflect a causal relationship or sampling error. We cannot interpret the intercept because the experimental design did not include giving some patients zero mg of the drug (i.e. a placebo). We can interpret the slope and we use causal language. On average each additional mg of the drug increased how much a patient slept by 29 minutes (=0.48*60).

(2) <u>Assumption #1</u>: Regression equation is linear in the <u>error</u> and <u>parameters</u>; the variables (in boxes) are linearly related to each other. I.e. Linear relationship between variables (possibly non-linearly transformed)

$$\boxed{} = \alpha + \beta\boxed{} + \varepsilon_i$$

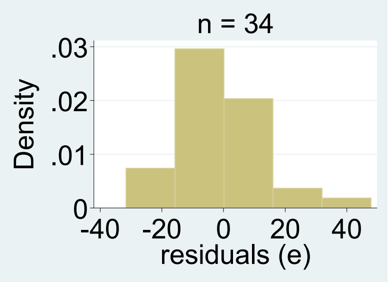
▶ This assumption clearly holds as the scatter diagram shown no evidence of a non-linear relationship between OECD countries' debt in 2009 versus 2003. I.e. the pattern appears straight.

<u>Assumption #2</u>: No autocorrelation / no serial correlation: $COV\left[\varepsilon_i, \varepsilon_j\right] = 0$ if $i \neq j$ I.e. No correlation amongst errors (no autocorrelation for time-series data)

▶ This assumption is not a concern because these data are cross-sectional and not time series. Don't get confused because two different years are compared. We are comparing these two years *across countries*. The <u>unit of observation</u> is a country. (In contrast, time series data would be limited to one country and watching that country's debt each year and the graph we would draw is NOT a scatter diagram but a line chart where the x-axis is year and the y-axis is debt.)

<u>Assumption #3</u>: Homoscedasticity: $V[\varepsilon_i] = \sigma_\varepsilon^2,\ i = 1, \ldots, n$. I.e. Homoscedasticity (single variance) of errors

▶ Given that these data are cross-sectional we need to check this assumption carefully as heteroscedasticity is common problem for cross-sectional data. However, the scatter diagram above does not show an obvious violation. The amount of scatter is roughly constant along the line. There is a bit of evidence that the amount of scatter may be increasing a bit as we move from low to high debt countries. You'd have to do a formal hypothesis test (which we do not learn in this course) to determine if that is enough evidence to reject the null hypothesis of homoscedasticity.



Assumption #4: Normality: ε_i is Normal I.e. Normally distributed errors

▶ It would be nice to see a histogram of the residuals to check properly. However, the scatter diagram itself does not raise any alarms: for example, an outlier would cause a violation of this assumption and there are no outliers in the scatter diagram. To the right is the histogram of the residuals for these data that you should use to check Assumption #4. You can see that it is very close to Normal.

Assumption #5: Error has mean zero: $E[\varepsilon_i] = 0,\ i = 1, \ldots,\ n$ I.e. Constant included (error has mean 0)

▶ This is easy: the OLS results clearly report a constant (i.e. do not force it to be zero) so this assumption will definitely hold. (Mathematically the mean of the residuals must be zero if a constant term is included.)

Assumption #6: x uncorrelated w/ error: $COV[x_i,\ \varepsilon_i] = 0$ I.e. No relationship between x and error

▶ This assumption will not hold for these observational data. There are many facts (variables) about each of the OECD countries in these data that would affect both the debt in 2003 and 2009. This is an extreme case because these are debt and not deficit figures. For example, highest marginal tax rate varies across OECD countries and contributes to the debt observed in 2003 and 2009. (There are *many* examples of lurking variables related to the tax code, structural labor market conditions, entitlement structures, etc.) Hence we would need to limit ourselves to descriptive or predictive statements. (Most people would not even try to make a causal statement in this example because the problems with doing so are more obvious in this example than in other examples where we natural have in mind a causal relationship between x and y (even if our data do not allow us to estimate it).)

(3) (a) These data are observational. Both the scatter diagram and the coefficient of correlation show (virtually) no association. We *cannot* conclude that these variables are unrelated even though we do not see an association here. We know that both are influenced by the same set of unobserved variables: hence we cannot measure any causal relationship from this diagram. The OLS slope and the coefficient of correlation suffers from a serious endogeneity bias. Hence, the lack of association does *NOT* mean that there is no causal relationship between these variables. We know that there is a "vast theoretical literature" talking about a link (and our gut instincts would suggest one as well). An endogeneity bias can make a correlation appear when there is no causal relationship *and* an endogeneity bias can make no correlation appear when there is a causal relationship. In other words, it can fabricate a false relationship or it can hide a real relationship. Hence the converse of the old cliché, which applies to observational data (but *not* experimental data), "correlation does not imply causation" is also true: "no correlation does not imply no causation" (again, when thinking about observational data that suffers from an endogeneity bias). Economists spend a tremendous amount of time worrying about endogeneity bias.

(b) No they do not match. Review pages 179 – 180 of your textbook: "One Correlation but Two Regressions."

(c) Lurking/confounding/unobserved/omitted variables must be systematically related to *both* the growth rate of countries and the inequality of countries. Some obvious examples would include variables measuring the tax/transfer system, educational policies/attainment, labor policies, and female labor force participation. These variables all vary considerably across OECD countries and should be related to *both* variables: growth rate and inequality.

(4) (a) http://homes.chass.utoronto.ca/~murdockj/eco220/TT220_1_NOV16_SOLN.pdf

(b) http://homes.chass.utoronto.ca/~murdockj/eco220/TT220_2_NOV17_SOLN.pdf

(c) http://homes.chass.utoronto.ca/~murdockj/eco220/TT220_2_NOV18_SOLN.pdf

(5) (a) The interpretation of the OLS intercept of 147.7342: The mean salary among the employees at Ryerson is $147,734 in 2016. (Notice that the Waterloo dummy is equal to zero for these employees so that we only look at the intercept.) The interpretation of the OLS slope of 2.838028: On average the employees at Waterloo have salaries that are $2,838 higher than the salaries at Ryerson in 2016. Putting the intercept and slope together, the mean salary for employees at Waterloo in 2016 is $150,572 (=$147,734 + $2,838).

(b) The tiny R-squared of 0.0016 means that only 0.16% of the variation in 2016 salaries among these 2,526 employees can be explained by whether they work at Ryerson or Waterloo. In other words, other factors (like years of employment, department, etc.) explain nearly all of the variation in salaries, rather than which university you work at.

(c) The interpretation of the s_e of 35.908: There is A LOT of scatter around the OLS line. This huge value of the s_e means that the standard deviation of the residuals is nearly $36,000! This is consistent with the tiny R-squared discussed in the previous part. There is a lot of variation in salaries among employees at Waterloo and among employees at Ryerson.

(d) $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{(1{,}189-1)34.97342^2+(1{,}357-1)36.70673^2}{1{,}189+1{,}357-2} = \frac{3280143.188}{2{,}544} = 1289.364$, which means $s_p = \sqrt{1289.364} = 35.908$. Remember the s_e is a measure of the s.d. of the residuals assuming equal variances (i.e. assuming constant scatter about the OLS line). Because in this simple case where x is dummy variable the OLS line passes through the mean of salary for each of the two categories, the concept of variation around the mean and variation around the OLS line are the same concept. This is one of the reasons it was important for you to study the seemingly unimportant special case in Section 14.5 of the textbook: the routine assumption in regression analysis is equal variances. Further, in this special case where the x variable is dummy we can really see what the homoscedasticity assumption (Assumption #3) means and how it relates to material from earlier chapters.

(e) While the standard deviations of salaries at the two universities are not equal (they differ by almost $2,000), they are not really that different: around $35,000 versus around $36,700. Hence, there is not a serious violation.

(f) Salary-hat = 150.5722 – 2.838028*Ryerson; n = 2,546; R-squared = 0.0016; s_e = 35.908. (Note: Make sure you get the NEGATIVE sign on the OLS slope coefficient.)

(g) Salary-hat = 154762.3 – 4190.085*Waterloo

(6) (a) How are annual real GDP growth rates affected by changes in annual government spending?

(b) The underlying data are panel (longitudinal) data. There are 132 observations (i.e. dots in the scatter diagram) that correspond to 33 countries each followed for 4 years (132=33*4).

(c) If we standardized both x and y then the slope of the OLS line would be the coefficient of correlation. Professor Krugman reports that the R-squared is 0.31, which means that the coefficient of correlation is 0.56. This means that for every one standard deviation increase in the percentage change in real government purchases, the annual growth rate in real GDP is 0.56 standard deviations higher on average.

(d) Yes. Looking at the graph, where the variables have not been standardized so we can assess their magnitudes, we see that we are talking about potentially big changes in the GDP growth rates that policy makers would surely care about.

(e) The article reports a t test statistic of 7.7. That is huge. Hence, we'll be able to easily reject the null hypothesis and infer that the research hypothesis is true: there is a relationship between GDP growth and changes in government purchases. We can argue about whether it is causal or not, but there is something there that is not just random noise.

(f) 31 percent of the variation across the 33 countries and across the four years (2010-2013) in the annual percentage change in real GDP can be explained by variation in the annual percentage change in real government purchases.

(g) The linearity assumption (Assumption #1) is fine (no evidence of curvature in the scatter diagram that would require a non-linear transformation of either the x or y variable). There is likely autocorrelation (violation of Assumption #2) because we are tracking each country over four years and macroeconomic variables typically have positive autocorrelation. There is no evidence of heteroscedasticity in the scatter diagram so Assumption #3 looks good. Ideally we would draw a histogram of the residuals to best visualize the check for Assumption #4, but there is no clear evidence of a violation of Normally distributed errors from the scatter diagram (i.e. Assumption #4 looks OK). There is nothing to worry about for Assumption #5: it holds whenever a regression analysis includes a constant term (which is the default). Professor Krugman recognizes that economists can raise valid concerns about Assumption #6: in other words, can we treat changes in government spending as exogenous or is it endogenous (i.e. influenced by factors that also affect real GDP growth rates)? While it is not hard to think of variables that would affect both the changes in government spending and GDP growth rates across countries and over time, Professor Krugman argues that the evidence does suggest a causal link. It is important to remember that checking assumptions is NOT a yes or no exercise: there are nearly always grey areas. The question is whether we have very serious violations that stop us from answering our research questions. In real-life applications there are at least modest violations of the assumptions. In this particular case, you could easily find economists who would argue that there is a serious issue with Assumption #6 and we cannot use this evidence to support a case for government spending as a stimulus for economic growth. Professor Krugman argues that the violations are not so extreme that the endogeneity bias is entirely the cause of the positive relationship observed in the scatter diagram: he argues that at least part of what we see is a underlying positive causal relationship between government stimulus spending and economic growth.