Homework 19: ECO220Y – SOLUTIONS

Required Problems:

(1) (a) Assumption #1 (linearity) appears to hold in both time periods: there is no clear evidence of curvature in the scatter plots. Assumption #2 (no autocorrelation) cannot be checked with the given information and there is a potential concern as these are time series data. (However, if you are curious, there is no autocorrelation in these data.) Assumption #3 (homoscedasticity) is **very clearly violated** in the 1970 – 2016 period: notice that there is a lot more scatter in the 1970's and 1980's compared to later decades. For 1995 – 2016, Assumption #3 holds. Assumption #4 (Normally distributed errors) cannot be directly checked with the given information (however, given the big concern with Assumption #3, Assumption #4 is a lesser issue). Assumption #5 (errors have mean zero) has nothing to check: both regressions include a constant term. Assumption #6 (exogeneity of the x variable) is not relevant here: the question is inherently descriptive (how is forecast accuracy changing over time?) and not causal.

(b) We would use $\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}}$. Yes, heteroscedasticity will invalidate this approach: the value of s_e that we plug in is NOT appropriate. That s_e is an average of the scatter over the entire time period and hence grossly overstates scatter in the most recent years. Hence, our interval would be far too wide. In contrast, it would be fine to use that formula with the values obtained from the regression for the 1995 – 2016 period.

(c) These are measured in nautical miles. The value of the s_e for the later time period is much smaller because there is much less scatter about the OLS line in later years. This has NOTHING to do with the smaller sample size (n = 22 versus n = 47).

(d) While the numerator is much smaller in the second regression, the denominator is also smaller for two reasons: a smaller sample size and less variation in the x variable (year).

(e) Yes, these data are examples of summary values. There are many hurricanes and tropical storms in the Atlantic each year. The y-variable is average tracking forecast error over all storms each year. If we had data at the level of a storm (much less aggregate) rather than at the level of a year (much more aggregate) then we would see a lot more scatter (bigger s_e and smaller R^2). Almost anything can happen with an individual storm: they are far harder to predict than the average performance of NOAA in a given year.

(f) Use $\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}}$ with x_g of 2017 (and $\nu = 20$) to obtain the 90% prediction interval (30.06, 77.10). (Note that your answers will only be accurate to the *nearest integer*. The exact values above are computed with software whereas you only had access to rounded inputs to work with.)

(g) Use $\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}}$ with x_g of 2017 (and $\nu = 20$) to obtain the 99% confidence interval prediction of the mean (39.09, 68.07). (Note that your answers will only be accurate to the *nearest integer*. The exact values above are computed with software whereas you only had access to rounded inputs to work with.)

(h) Use $b_1 \pm t_{\alpha/2} s.e.(b_1)$ with $\nu = 20$ to obtain the 99.9% confidence interval estimate of the OLS slope coefficient that is (-5.95, -2.67). We are 99.9% confident that, from 1995 through 2016, NOAA has on average reduced its mean annual track forecast error for Atlantic tropical storms and hurricanes by between 2.67 and 5.95 nautical miles per year. It is clearly getting better each year (at least on average).

(2) The variance of the error term determines its relative importance (the amount of scatter, which reflects how much factors aside from x affect y). The underlying assumptions of the linear regression model tell us that $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, which says that it has mean zero, constant variance of σ_ε^2 (homoscedasticity), and is Normally distributed. When σ_ε^2 is large then there is a lot of scatter and the error term is relatively important compared to the x variable. In contrast, when σ_ε^2 is small then there is little scatter and the error term is relatively unimportant compared to the x variable. The mean of the error does NOT determine its relative importance because the mean is zero: the inclusion of the constant term implies that our line passes through the middle of the scatter diagram leaving the average residual to be zero (the positive errors cancel out with the negative errors).

While σ_ε^2 is the parameter that determines the relative importance of the error we cannot calculate it (without observing the entire population). Hence we calculate the standard error of the observed residuals (from our sample) instead. (Note: There is a degrees of freedom correction because we need two estimates to compute e: b_0 (the intercept) and b_1 (the slope).)

$$\sigma_\varepsilon^2 = \frac{\sum_{i=1}^{N}(\varepsilon_i - 0)^2}{N} \qquad s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

(3) In a simple regression analysis (i.e. only one x-variable) this formula quantifies how much sampling error affects the OLS slope estimate: $s_{b_1} = \frac{s_e}{s_x\sqrt{n-1}}$. It depends on: 1) the sample size (as $n \uparrow$, sampling error $\downarrow$), 2) the variability of the x-variable (as $s_x \uparrow$, sampling error $\downarrow$), and 3) the amount of scatter (as $s_e \uparrow$, sampling error $\uparrow$).

(4) (a) No, in practice you cannot observe epsilon. It is a Greek letter and hence is something associated with the population. We observe random samples and use the sample statistics to make inferences about the population parameters. However, Slide 4 of Lecture 19 is a HYPOTHETICAL case where we DO observe the entire population (notice how crowded the scatter diagram is). If we observe the entire population then we observe alpha and beta (40 and 5 in the cGPA and starting salary example). If we observe those parameters then we would observe epsilon: it is the distance between the line in Slide 4 and an individual observation (point) in that diagram. I.e. epsilon_i = y_i – (alpha + beta*x_i). However in REAL applications we only have a sample and an OLS intercept and slope (a and b and NOT alpha and beta). For a sample we can calculate e_i = y_i – y-hat_i = y_i – (a + b*x_i). You can think about e_i as an estimate of epsilon_i.

(b) Let's illustrate with another example. Suppose you wanted to investigate the relationship between study time and marks on a test. You are aware of all of the pitfalls of observational data (i.e. just observing marks and study time) and decide to collect experimental data. You randomly assign students to study for 10 hours and 1 second, 10 hours and 2 seconds, 10 hours and 3 seconds, … up to 10 hours and 60 seconds. You have very little variability in your x-variable. Even if more study time has a big effect on marks you will have trouble getting a precise estimate of that effect from your study because your participants hardly vary in their amount of study time. In other words, the standard error on the slope estimate will be large. Even keeping sample size and other factors constant you could get a much smaller standard error on your slope by setting up your experiment to vary study time (the x-variable) more: for example, randomly assigned some students to study for 1 hour, others to study for 5 hours, others to study for 10 hours, etc.. This is what the formula on Slide 6 tells us (i.e. the s.e. of the slope is inversely related to the variability of x as measured by its sample standard deviation).

(c) Answer: Yes, it is important. Assumption 6 formally says that COV(x, epsilon) = 0. What does that mean? Let's put it into the context of the cGPA and starting salary example to illustrate. We wrote a model as: Starting salary = alpha + beta*cGPA + epsilon. Assumption 6 says that there should be no relationship between epsilon and cGPA. Just like we did in class we need to brainstorm about what kinds of variables are in epsilon: in other words, what else affects graduates' starting salary aside from their cGPA? There are lots of good examples. Let's take field of study. Obviously students whose field of study is finance typically have higher starting salaries than students whose field of study is English literature. That means that field of study is in epsilon: it is another variable that affects the y-variable (salary). That is OK. A violation of Assumption 6 occurs when epsilon is correlated with the x-variable (cGPA). It is a well-known fact that at U of T and other universities there are "easy" fields of study with high marks and "hard" fields of study with low marks. I've seen student transcripts where the AVERAGE mark in a math course is a D and others (not math courses)

with an AVERAGE of A! Hence field of study is correlated with cGPA in violation of Assumption 6. There are many other examples that cloud any true causal relationship between cGPA and starting salary. Hence, in the end we must limit ourselves to descriptive statements when dealing with observational data where Assumption 6 is violated. (Another way to say this is that our estimate of the causal effect suffers from an endogeneity bias.) Another way to state Assumption 6 is to say that it is the assumption that there are no lurking variables. Remember lurking variables are variables correlated with BOTH the y-variable and the x-variable, which in our regression model notation means variables in epsilon that are correlated with the x-variable (y is a function of epsilon so of course it is correlated with epsilon). We do not encounter violations of Assumption 6 in experimental data precisely because we randomly set the values of the x-variable and hence assure that it is NOT systematically related to all of the other things that affect the y-variable.

(5) (a) Use prediction interval with $\nu = n - 2 = 50 - 2 = 48$. Formula: $\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}} = 40.47529 + 4.334865 * 3.2 \pm 2.011 * 11.487 \sqrt{1 + \frac{1}{50} + \frac{(3.2-2.74)^2}{49*0.3659988^2}} = 54.35 \pm 2.011 * 11.78 = 54.35 \pm 23.70$. Hence the lower limit of the 95% prediction interval is 30.7 and the upper limit is 78.0.

(b) Choice **(E)**

(c) Use confidence interval with $\nu = n - 2 = 50 - 2 = 48$. Formula: $\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}} = 40.47529 + 4.334865 * 3.2 \pm 2.011 * 11.487 \sqrt{\frac{1}{50} + \frac{(3.2-2.74)^2}{49*0.3659988^2}} = 54.35 \pm 2.011 * 2.625 = 54.35 \pm 5.28$. Hence the lower limit of the 95% confidence interval is 49.1 and the upper limit is 59.6.

(d) Choice **(C)**

(e) We are 95% confident that people with a cGPA that is 1 point higher (i.e. 3.00 versus 2.00) on average have a salary that is between $4,680 *lower* to $13,350 *higher*. In other words each additional point of cGPA is associated with an extra $4,335 but this is very imprecisely estimated and has a large margin of error (ME = $9,017 = 2.011*4.4837*1000), which means that we cannot be confident that the association is even positive.

(6) (a) Choice **(A)**

(b) Choice **(D)**

(7) (a) An individual patient that takes 6 mg of the drug should be 95% confident of sleeping between 4.28 and 8.37 hours. On average patients that take 6 mg of the drug will sleep between 5.79 and 6.87 hours with 95% confidence.

(b) No, absolutely not. In fact all of the patients may be dead from an overdose if they take 40 mg (i.e. infinite hours of sleep or zero hours depending on how you define “sleep”).

(c) Choice **(A)**

(d) Choice **(B)**

(8) (a) Choice **(D)**

(b) Don’t be fooled by all of the output. All we can say is that descriptively we see no (statistically significant) association between these variables. This does NOT rule out the possibility of a strong causal relationship. All else is NOT equal (i.e. not held constant) as we change the x variable (gdp growth in this case). Assumptions #1 - #5 seem to hold in this case

(no obvious violations). Unfortunately a violation of Assumption #6 is extremely problematic when it comes to properly interpreting results when we have a causal research question rather than a mere descriptive question.

(9) (a) We cannot say that if a student increases attendance by 1 percentage point that his/her mark will go up by 0.6889 percentage points. The reason is that we have a violation of Assumption #6: PCT_ATT is correlated with other factors that affect a student's mark like effort, motivation, etc. that we have not included in our model and hence are picked up by the error term. We can say that marks and attendance are positively correlated and that on average students whose attendance record is 10 percentage points higher (e.g. attending 95% of classes versus 85% of classes) have a mark that is on average 6.9 percentage points higher. This is a simple descriptive statement that is true. However, we must stop ourselves from going a step further and saying that higher attendance will lead to (i.e. causes) higher marks.

(b) We cannot say that if a student never attends class (PCT_ATT=0) their mark will be 20.65%. We can say that the intercept has no interpretation because it is way out of sample: no student is observed attending so few classes and hence we have no reliable basis for predicting what the mark would be for zero attendance.

(c) MARK-hat = 20.6517 + 0.6889006*PCT_ATT
MARK-bar = 69
20.6517 + 0.6889006*PCT_ATT-bar = 20.6517 + 0.6889006*70.18182 = 69 ✓

(d) Yes, this would be an outlier. No other student has a mark close to 0: the lowest mark observed in the data is 42%. In this instance, one could argue that this particular data point should be excluded because we have not observed an actual mark for this student. This outlier would be an influential data point in terms of the slope estimate if this student (who got a 0) had particularly high or a particularly low attendance. If this student was around average in terms of attendance then the outlier would not be influential in terms of the slope.

(10) (a) It is the variance of the change in mean happiness in rich countries from 2006 to 2011. It would be measured in units of happiness on a 10-point scale squared. More easily interpreted is the standard deviation 0.40 (= square root of 0.157597178). This says that among rich countries the s.d. of the change in happiness over that five year period is 0.4 units on a 10-point scale. That is a pretty big standard deviation indicating quite a bit of variability even among rich countries in how their happiness changed over this five year period. This is consistent with the scatter diagram (focusing on the vertical axis) which shows that some countries had a pretty big jump in happiness (almost a full point on a 10-point scale) to a pretty big decline (almost a full point drop).

(b) That is the standard deviation of the residuals which we denote as s_e. It is measured in the same units as the y variable: hence, change in happiness on a 10-point scale. It measures the amount of scatter about the line. It is pretty big at 0.36 (we just saw the s.d. of the y variable is 0.40). That means that there is quite a bit of scatter around the line: while there is generally a positive trend, it is pretty weak. Some countries that had a pretty good improvement (change) in GDP had happiness *drop*, which weakens the strength of the positive trend.

(c) That is the slope coefficient. It means that on average rich countries whose GPD increased by an extra $1,000 USD per capita over the five year interval (2006 to 2011) had average happiness increase by 0.1 units on a 10-point scale (e.g. comparing an average county whose GDP increased by $500 with one that increased $1500). Note that this statement is purely descriptive and does not imply causality.

(d) The standard error of the slope coefficient is 0.03, which means that it is estimated fairly precisely (the point estimate is 0.09). The t statistic and P-value for the hypothesis test $H_0: \beta = 0$ versus $H_1: \beta \neq 0$ mean that we can easily reject the null: in other words, we have a statistically significant slope estimate at even a 1% significance level.

(e) The intercept (constant) term is -0.16. In this case, an x value 0 is within the range of the data as the fourth scatter diagram makes clear. Hence, we can interpret it. On average, amongst rich countries (with GDP per capita of at least $20,000 USD in 2011) those whose per capita GDP did not change over the five year interval saw mean happiness decline by 0.16 on a 10-point scale.

(f) The R^2 of 0.21 means that 21% of the variation in the five-year change (from 2006 to 2011) in mean happiness among rich countries is explained by variation in the five-year change in GDP per capita.

(g) Any lurking variables that are constant over the five year period would be addressed by the proposed approach of looking at changes in the variables rather than levels. For example, in rich countries, which is what the analysis addresses, there have likely been very little changes in average educational attainment during these five years. Hence, while average educational attainment would be lurking variable in a regression like that considered in Lecture 19, Slide 19 (levels), it would "cancel out" in a regression like this one which looks at the change in happiness versus the change in GDP. Because education is constant it should not be correlated with the *change* in happiness. However, there can still be lurking variables that would affect this analysis. For example, suppose some countries launched a sizeable military response to terror attacks during the five year interval. This would affect both the country's GDP growth during this period and the happiness of its citizens and hence would be a lurking variable.

(11) The answer is **(C)**. If the P-value is 0.90 then the evidence is NOT in favor of the one-tailed research hypothesis. This is a situation where you got a negative test statistic for a right tailed test or a positive test statistic for a left tailed test. Pick one of those two scenarios and draw the Student t distribution (as suggested) to visually verify that the P-value for the two tailed test would be 0.20.