Homework 6: ECO220Y – SOLUTIONS

Required Problems:

(1) Answers are on pages 18 – 21 of "Logarithms in Regression Analysis with Asiaphoria."

(2) (a) The raw data are panel: the same variable Real GDP per Capita is observed for two different time periods for many different countries. The figure depicts cross-sectional variation (i.e. unit of observation in the figure is a country).

(b) No. That is a 45 degree line and clearly not an OLS line: there is no way the line shown in the figure minimizes the sum of the squared errors: it clearly lies below most of the observations and does not fit the data well.

(c) The logarithmic transformations straightened the scatter plot. The axes show a common trick to help people understand data that have had a logarithmic transformation. Instead of showing logged units they show original units (in this case thousands of US dollars). However, they move them to reflect the logarithmic transformation (i.e. the numbers are not spaced the same as the real number line). In the figure, 5 and 10 are the same distance apart as 25 and 50. That looks like a huge graphing mistake, but it is not. To see why, ln(5) = 1.61 and ln(10) = 2.30 with a difference of 0.69 (= 2.30 – 1.61) and ln(25) = 3.22 and ln(50) = 3.91 with a difference of 0.69 (= 3.91 – 3.22). Hence if a logarithm has been applied these should be equidistant. The nice thing about the figure (once you get used to it) is that the units on the axis are in thousands of dollars (and not the natural logarithm of thousands of dollars).

(d) See the paper for a discussion (if needed).

(3) The *additional* two graphs show the *diagnostic plots* for the two regressions (the first regression using the real Fortune 500 data and the second regression using the fake Fortune 500 data). Recall that, for a regression, a diagnostic scatter plot graphs the residuals on the y-axis versus the predicted values of y (y-hat) on the x-axis. While technically this is just transforming and replotting the information already available in the top two graphs (where we simply plotted the y variable versus the x variable), visually it is easier for humans to notice problems – like non-linearity and heteroscedasticity – in the diagnostic plot. Of course, severe issues will be obvious in the original scatter plot of y versus x: the diagnostic plot helps you notice more subtle issues. Also we'll see later in the course that the diagnostic plot is helpful in multiple regression when we are not able to graph the relationship between y and a bunch of x variables.

The first diagnostic plot (real Fortune 500 data) shows a clear pattern. This reflects the very extreme positive skew of revenues in the original data: even the natural log is not powerful enough to eliminate the skew and fully address the non-linearity. (Recall that there are TWO forces creating the positive skew: one is the natural positive skew of revenues across firms and two is the fact that your membership in the Fortune 500 is determined by your revenue, which creates truncation (no firms below the 500 are included).) Of course, you can see these problems even in the original plot (although it is even more obvious in the diagnostic plot).

The second diagnostic plot (fake Fortune 500 data) shows a no pattern at all: a random cloud of dots. That is what we hope to see in a diagnostic plot and means we do not have issues with non-linearity or heteroscedasticity.

(4) The intercept has no interpretation (no vehicle gets 1 MPG in city driving). The coefficient 6.02 means that vehicles in these data that have 1 percent better fuel economy in city driving have GHG ratings that are approximately 0.06 points higher (on a 10 point scale) on average. The R^2 is very high: 94% of the variation in GHG ratings can be explained by variation in the natural log of fuel economy in city driving.