Homework 5: ECO220Y – SOLUTIONS

Required Problems:

(1) (a) These data are observational, cross-sectional, and contain two variables with interval data.

(b) These results show that there is a negative association between the number of children women have and years of education. In particular, each extra year of education is associated with 0.2 fewer children on average. This is NOT to say that we can conclude that education is causing women to have fewer children. What we have observed is simply a CORRELATION between these variables. Correlations do not imply causal relationships. All we can do is describe what we see in the data, which is simply that the average number of children more educated women have is lower than the average number of children less educated women have. We cannot conclude that the education itself is causing the measured difference (as per the slope of the OLS line). (To convince yourself, think of reasons that women choose to obtain more education and ask if any of those underlying factors would also affect their choice about having children.) The coefficient of correlation (-0.3132) shows that there is a relatively weak negative relationship between these two variables. The "intercept" of 5.1 is meaningless. No Canadian women in our sample had zero years of education (not even close to zero years of education). We should not extrapolate back because it would take us well out of our sample and hence we'd have NO evidence to support an interpretation of 5.1 as the number of children women with zero years of education have on average.

(c) The scatter diagram is not a good summary of these interval variables because in this example both variables are integers and take on relatively few unique values. An alternate way to summarize these data is with a cross-tabulation. Next is a Stata cross-tabulation of the original data (note that it is very long and extends over two pages):

	NUM_KID						
EDU_YRS	0	1	2	3	4	5	Total
7	0	0	0	0	1	0	1
8	0	0	0	1	2	1	5
9	0	2	5	5	2	2	16
10	1	6	15	21	18	6	75
11	2	19	53	58	23	11	174
12	7	28	94	91	52	19	302
13	13	72	121	106	55	21	397
14	29	99	148	104	42	16	458
15	28	115	126	87	45	7	417
16	43	103	78	66	26	6	328
17	40	57	49	24	15	5	196
18	21	34	34	13	4	1	111
19	12	23	14	4	0	2	56
20	10	8	5	3	1	0	28
21	4	6	0	2	1	0	13
22	1	2	0	1	0	0	4
23	3	0	0	0	0	0	3
24	2	1	0	0	0	0	3
25	0	1	0	0	0	0	1
27	1	0	0	0	0	0	1
Total	217	576	742	586	287	97	2,589

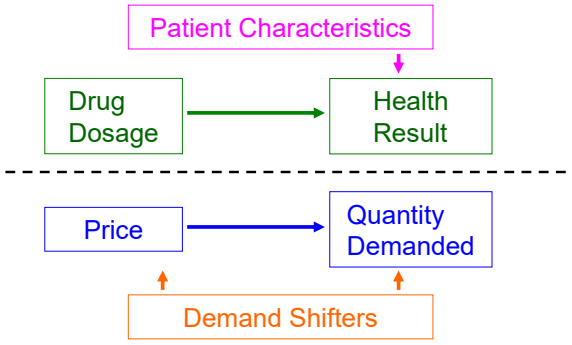
	NUM_KID					
EDU_YRS	6	7	8	9	11	Total
7	0	0	0	0	0	1
8	1	0	0	0	0	5
9	0	0	0	0	0	16
10	5	1	0	1	1	75
11	4	3	1	0	0	174

12	7	3	1	0	0	302
13	6	2	0	1	0	397
14	12	6	0	2	0	458
15	6	2	0	1	0	417
16	3	3	0	0	0	328
17	4	0	1	1	0	196
18	4	0	0	0	0	111
19	1	0	0	0	0	56
20	1	0	0	0	0	28
21	0	0	0	0	0	13
22	0	0	0	0	0	4
23	0	0	0	0	0	3
24	0	0	0	0	0	3
25	0	0	0	0	0	1
27	0	0	0	0	0	1
Total	54	20	3	6	1	2,589

(2) (a) An example of time series data would be tracking the prices and tickets sold each day for non-stop Air Canada flights from Toronto to Chicago on Monday afternoons for 3 years: there would be roughly 156 observations (=52*3) and two variables (price and quantity). An example of cross-sectional data would be recording the prices and liters of water sold across 50 randomly selected medium-sized municipalities: there would be 50 observations and two variables (price and quantity).

(b) Observational. Experimental data would require that someone had randomly set prices and then watched how consumers responded. Not surprisingly firms are interested in maximizing their profits and hence have little appetite for letting some researcher come in and randomly set their prices.

(c)



(d) The data would be observational and time series. It would have 3 observations and two variables (price and quantity). The coefficient of correlation would be positive but that does not mean that demand is upward sloping. We can see from the diagram that the market equilibrium prices and quantities do NOT trace out the demand curve. This is because of changes over time in demand shifters (such as income, tastes, prices of substitutes, prices of compliments, etc.). If there were no demand shifters (of course not possible) and only supply shifters then we could recover demand. Of course in the very periods when demand is soft the equilibrium price tends to be lower: this does not mean that the low price caused low quantity demanded but rather a negative demand shock cause both the low equilibrium price and the low quantity demanded.

(e) Yes we could and it would not suffer from an endogeneity bias. (Note: We would need to be careful to consider the total number of actual customers rather than just the sample size. For example, for a linear demand as the number of consumers increases – other things equal – demand rotates out and hence the slope become flatter.) There would be no endogeneity bias because these are experimental data. Price has been randomly set and is not influenced in any way by demand shifters (that also affect the quantity sold). It does not matter at all that our sample of customers is

heterogeneous. Unlike observational data is not the case that consumers in markets with higher income, for example, are offered the product for a higher price in market equilibrium: according to the description of the data collection price was *randomly* set and hence was not influenced by demand shifters like income.

(3) (a) The covariance is 41.3 mark-hours.

(b) The coefficient of correlation 0.88, which means that a one standard deviation increase in hours of sleep is associated with a 0.88 standard deviation increase in mark. There is a strong positive linear association between study time and marks. The coefficient of correlation is 0.88, which is not too far from its maximum possible value of one for a perfect positive linear association.

(c) The coefficient of determination is 0.78, which means that 78 percent of the variation in marks in our sample of 10 students can be explained by variation in hours of sleep.

(d) These data are observational.

(e) The regression line equation is: mark-hat = 36.7 + 6.1*hours. In these data we see that students who spend an extra hour of sleeping have test marks that are on average 6.1 points higher on a test worth 120 points. The intercept of 36.7 has no interpretation because no one in our sample has reported close to zero hours of sleep. It is simply a shifter of the least squares line. Saying something like "Sleeping zero hours means a grade of 36.7" is incorrect for two reasons: (1) We have absolutely no data to support that estimate of what marks would be if a student slept zero hours and (2) It wrongly implies causality, which is discussed next. We cannot conclude that the increased sleep time is causing the higher marks. In other words, we *cannot* say things like "if a student sleeps an extra hour then their mark would go up by 6.1 points" or "each extra hour of sleep will increase the mark by 6.1 points on average." Statements like these are wrong because they imply that we have estimated the causal effect, which we have not. What we can do is *describe* the data. It is true that in the data students that slept more hours have higher marks and that in the data we see that students that slept an extra hour had marks that are 6.1 points higher. These are correct descriptive statements and are different from causal statements. Unfortunately you cannot use these results to answer the interesting causal research question posed at the start of this problem.

(f) These replication results are quite different: for example, the coefficient of correlation and determination are quite a bit lower and the slope is steeper. All of these – covariance, standard deviations, correlation, R^2, intercept, slope – are sample statistics and not population parameters. Hence they are all subject to sampling error. Given the small sample sizes, there is a lot of sampling error. It is not surprising that the results look different.

(g) percentage_mark-hat = 33.25 + 5.83*hours, n = 30, R-squared = 0.42

(h) mark-hat = 39.9 + 0.12*minutes, n = 30, R-squared = 0.42

(4) (a) The OLS line is NOT an estimate of the U.S. supply curve for corn during this period. A supply curve (like you learned in introductory level economics) is a *causal* relationship: it tells how the quantity supplied responds to the market price. Remember that these are observational data and will suffer an extremely severe form of endogeneity bias: every variable that is a supply shifter is a lurking/unobserved/confounding/omitted variable (these affect both the market equilibrium price each year and the quantity supplied each year). Hence, the slope of the OLS line is an extremely biased estimate of the slope of supply. In fact, as shown above, it is quite possible to even get a *downward* sloping OLS line when trying to estimate supply! That didn't happen in this case, but we still know it is extremely biased.

(b) Our interpretation must be descriptive, not causal. Also, we're talking about such big numbers, it is useful to change from millions to billions to help make our interpretation clear. From 1926 through 2012, in years when the price of corn is one dollar per bushel higher, on average the U.S. corn production is about 2.14 billion bushels higher.

(5) (a) His inference is faulty: he does not grasp the concept of regression to the mean.

(b) Yes, absolutely necessary. If there were absolutely no random elements (hard to imagine when thinking about human behavior) then there would be no regression to the mean.

(c) No, he is pessimistic about our ability as humans to overcome our own biases even when we are made aware of them (e.g. by being shown a convincing experiment).

(6) (a) The R^2 is moderate in the first graph: across all 156 countries about 43 percent of the variation in mobile-cellular subscriptions per 100 inhabitants is explained by variation in the HDI (Human Development Index). The R^2 in the second graph is basically zero: when looking at OECD member nations there is no correlation between HDI and mobile-cellular subscription rates. While there are three somewhat usual points, even if we remove them the last graph shows the R^2 is still tiny. The reason for these differences is that the OECD countries must be systematically different from other countries (as we know they are). Hence, with OECD member nations there is basically no relationship between mobile-cellular subscription rates and the HDI. However, looking at all countries – including some very poor countries – there is a clear positive correlation between these variables. Note: The difference is NOT due to the difference in sample sizes (i.e. the number of observations). If we pulled out 34 nations at random (instead of the 34 OECD member nations), we would expect a comparable R^2 (i.e. around 0.43 aside from any sampling error).

(b) The SST, SSR, and SSE are sums of squares. As such, they do depend on the sample size: squares are positive so the bigger the sample, the bigger the SSs are. That is why they are all much bigger in the first graph and smaller in the others. However, once we look at the R^2, which is a ratio of the SSR to the SST, then we get a measure that does not depend on the sample size (or the original units of measurement). This is why the R^2 is usually reported along with the OLS results, whereas the SSs would usually only be given in the full software output.

(7) (a) Cross-sectional, 1,250 observations, three quantitative variables (and also some identifier variables for make and model).

(b) The shape is roughly Normal (more density in the centre and less in the tails). A vehicle with a GHG rating of 3 would have a standardized value of $-1.4 = (3 - 5.3)/1.6$: this means that it is almost one and a half standard deviations below average in terms of GHG ratings (not good!).

(c) The diagonal elements are the variance (in MPG-squared). These are best interpreted by converting them to standard deviations: 5.5 MPG for city and 6.2 MPG for highway. Hence the vehicles are a bit more variable in terms of highway fuel efficiency. The off-diagonal is the covariance (in MPG-squared). This is best interpreted by computing the coefficient of correlation: $r = cov/(sd\ x * sd\ y) = 31.1/(5.5 * 6.2) = 0.91$. This means these are strongly positively correlated: vehicles with good city fuel economy typically have good highway fuel economy.

(d) The intercept has no interpretation because clearly no vehicle gets zero miles per gallon in highway driving. The slope coefficient means that vehicles in these data (1,250 different makes and models) that get an extra MPG in highway driving on average get an extra 0.81 MPG in city driving. The R^2 is pretty high: 83 percent of the variation in city-driving fuel economy is explained by variation in highway-driving fuel economy.