Homework 4: ECO220Y – SOLUTIONS

Reminder: For all solutions to textbook exercise, visit the Readings page in Quercus.

Required Problems:

(1) The answer is **(A)**. Notice that the question asks how the graphs *differ*. They both show no relationship of any kind. The only difference is the amount of variation of the x and y variables (causing one to *look* more "horizontal" and the other to *look* more "vertical").

(2) (a) These would be observational data. Both the x-variable (also referred to as the independent variable or the explanatory variable) and the y-variable (also referred to as the dependent variable) are endogenous. Both are affected by other factors (unobserved variables) because the Bank of Canada sets the interest rate in response to everything it knows about the direction of the economy and many of these factors obviously affect the growth rate of GDP. The Bank of Canada does not randomly set the x-variable (interest rates) but rather systematically chooses it taking into consideration factors that also affect the y-variable (growth rate of GDP). It would be highly unethical for the Bank of Canada to try to collect experimental data by randomly setting the interest rate and seeing what happens to the growth rate of GDP even if such data would be much better able to establish the causal link.

(b) These would be experimental data. Ethics with respect to animals are currently far more lax than with human subjects. Researcher would randomly assign mice to be fed certain diets and then they watch what happens to them. These experimental data are highly informative because significant differences in outcomes can be attributed to the different diets: in other words we can infer causality. The differences in diets are entirely random and hence cannot be influence by unobserved variables like the individual traits of the mice. However there is a price to pay for these informative data: the well-being of some mice, those given no choice but to consume an inferior diet, is sacrificed. In contrast, observational data would allow the mice to select their own diets (we would simply observe what they choose) and their individual characteristics, such as propensity to exercise, may affect their choice of diet and their health. While collecting these observational data would allow the mice free choice, we would not be able infer causality from any association we observe between diet and health.

(c) These would be observational data. Firms set (choose) prices taking into account the market demand conditions that also affect the quantities that they sell. Everything that affects demand (demand shifters) would affect both price and quantity (which are the unobserved variables in the terminology of Lecture 4). This is the classic economic example of observational data. It is also is the central problem in empirical economics: how do we measure the responsiveness of quantity demanded to a change in price? Unfortunately because only observational data is available--we've yet to see many firms randomly set their prices--we cannot easily measure the causal relationship that is very much of interest. More advanced courses in econometrics would develop sophisticated methods to try to extract a measure of the causal relationship between quantity demanded and price from observational data. This is not easy. We must continually be on guard for endogeneity bias.

(3) (a)

```
. corr reading math science;
(obs=65)

             |  reading     math  science
-------------+---------------------------
     reading |   1.0000
        math |   0.9479   1.0000
     science |   0.9817   0.9708   1.0000
```

(b) Strongly *correlated* because there is no significant evidence of a non-linear relationship in any of the six graphs. Neither are there concerns about outliers unduly influencing the coefficient of correlation statistic.

(c) These are cross-sectional, observational data. We can interpret these strong positive correlations descriptively, not causally. Countries with strong reading scores also typically have strong math scores. Further, this correlation is very strong. (We *cannot* say that countries who work to teach their students to read well we can expect to boost their math scores. That would be a causal interpretation and is incorrect.)

(d) The variation in scores across countries on all three tests – reading, math, and science – declined a little bit from 2009 to 2012: for example, the s.d. of science scores dropped from 56 test points to about 51 (the mean test points in science are roughly around 475). That said, there is still considerable variation. Further, the variation on the math scores across countries remains the highest in both years (the s.d. is about 55 versus 47 test points for reading, in 2012). Also, in both years, the correlation among all scores is very high.

(4.1) What does the phrase "empirical evidence" mean? It refers to evidence based on ____. **(B)**

(4.2) This is an example of what kind of data? **(B)**

(4.3) What happens to the 50 plots in the control group? **(B)**

(4.4) What is the distinguishing feature of a randomized controlled experiment in the farming example? **(A)**

(4.5) observational **(A)**

(4.6) experimental **(B)**

(4.7) cross-sectional **(B)**

(4.8) time series **(B)**

(4.9) longitudinal (panel) **(A)**

(4.10) By using observational data that shows that in schools with smaller class sizes the learning outcomes are typically better than in other schools with larger classes, why is it difficult to answer "Does reducing class size improve elementary school education?" It is difficult because ____. **(D)**

(4.11) While the results will not by systematically wrong, there will be a fair bit of sampling noise and this will limit your ability to answer your research question. **(A)**

(4.12) You have made no attempt to ensure that the two groups are *otherwise identical* and this means that your data should not be used to answer your research question. **(B)**

(4.13) You have failed to ensure that other factors are *held constant* across the two groups and this will lead to an overestimate of the causal effect of the questionnaire format. **(B)**

(4.14) You should have conducted a randomized controlled experiment rather than relying on observational data. **(B)**

(5) (1) Bank of Canada interest rate and inflation rate example: These are observational, time-series data. The research question is what is the effect of changes in policy interest rate (X) on inflation (Y). Because the Bank of Canada CHOOSES the policy rate after considering many macroeconomic variables these data are observational. The Bank is not randomly setting the interest rate. There are many unobserved/lurking/confounding variables that affect BOTH the interest rate and the inflation rate: to brush up on this check out http://www.clevelandfed.org/research/commentary/2010/2010-17.cfm. For example, the unemployment rate may directly affect future inflation AND the Bank's action. Hence we would say that the x variable in this case, the lagged Target Overnight Rate, is ENDOGENOUS, because it is related to variables that ALSO affect the y variable in this case (CPIX). Further, the positive correlation that we see between the interest rate and inflation of 0.31 suffers from an ENDOGENEITY BIAS. In fact we know from ECO100 that raising the interest rate should decrease inflation down the road: hence this is an example of sever endogeneity bias. (Someone asked me if the bias is always positive and the answer is no. It depends on the circumstances.)

(2) Chocolate consumption and Nobel Laureate production: These data are observational, cross-sectional data for a bunch of countries. The research question is: What is the effect of eating chocolate on intellectual ability (as measured by Nobel prizes)? The X variable -- chocolate consumption -- is clearly ENDOGENOUS. That means that it is related to other variables that also affect the Y variable -- Nobel Laureate production. The other variables are called unobserved variables and include things like the GDP per capita, which would be associated with BOTH the chocolate consumption (an expensive luxury good) and Nobel Laureate production (an extreme luxury good that requires huge investments in education and research infrastructure sustained over many decades). Hence the reported correlation of 0.79 suffers from a severe ENDOGENEITY BIAS: it does NOT reflect the real (if any), causal relationship between eating chocolate and brain power. It vastly overstates this relationship. For the simple correlation to be an unbiased measure of the causal relationship between these two variables we would need a massive world-wide experiment sustained over many years where countries are randomly assigned levels of chocolate consumption (which they must follow). Not going to happen.

(3) Drug dosage and hours of sleep. These are cross-sectional, experimental data. The drug dosage each person got was randomly assigned. That means that drug dosage is EXOGENOUS. As such the correlation we obtained of 0.61 between sleep and drug dosage suffers no endogeneity bias: it is an unbiased measure of the strength of the true causal relationship between these two variables. (There is only sampling error to worry about and generally sampling error does not cause bias, just noise.)

(6) First, find mean and s.d. of each variable. (This is review: see Lecture 3, Slides 18 – 19 if you are unsure how to do this.) For chosedom, the mean is 0.4922481 and the s.d. is 0.5002632. For male, the mean is 0.5335917 and the s.d. is 0.4991929. Next, compute the sample covariance:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{n-1} =$$

$$\frac{178(0-0.4922481)(0-0.5335917)+215(0-0.4922481)(1-0.5335917)+183(1-0.4922481)(0-0.5335917)+198(1-0.4922481)(1-0.5335917)}{774-1} =$$

$$-0.0068544.$$

Next, compute the coefficient of correlation: $r = \frac{s_{xy}}{s_x s_y} = \frac{-0.0068544}{0.5002632*0.4991929} = -0.0274.$

Hence, the coefficient of correlation between male and chosedom is a very slight negative correlation of -0.0274, which means these variables are essentially unrelated. There is no clear association or correlation between sex and making the best credit card choice. NOTE: There can be no non-linear relationship between dummy variables because dummy variables can take only two possible values.

(7) (a) While the scatter diagram shows some curvature and a bit of a break in the pattern between 2006 and 2008, the correlation still works fairly well as a summary statistic here. In particular, 0.9683 means that undergraduate enrolments have been rising quite steadily over the period from 2000 – 2016: i.e. this strong positive correlation between year and enrolments means enrolments have been increasing quite steadily.

(b) It wouldn't change at all: it would still be 0.9683.

(c) It wouldn't change at all: it would still be 0.9683.

(d) While enrolments of 420,000 would not be an outlier in 2016, it would be a major outlier in 2000. That particular outlier would cause the positive correlation of 0.9683 to drop considerably as it would dramatically weaken the overall positive trend observed in the scatter diagram. (In fact, the correlation would drop to 0.7237.) The coefficient of correlation is not robust to outliers: it can be quite sensitive, depending on where the outlier is.

(8) (a) The correlation is very close to zero. When x is small (0), y is small (0) about 80% of the time (=949/1182) and when x is big (1) y is small (0) about 80% of the time (=2268/2818). Alternatively, you could notice that when y is small (0), x is small (0) about 30% of the time (=949/3217) and when y is big (1), x is small (0) about 30% of the time (=233/783). Hence, there is no evidence of a relationship between x and y: when one is big or small the other is not reacting by either following or doing the opposite. [If you did compute the correlation, like in Required Problem (6) above, it is -0.0022.]

(b) The correlation is negative and very strong. When x is small (0), y is small (0) about 7% of the time (=28/414) and when x is big (1) y is small (0) about 96% of the time (=562/586). Alternatively, you could notice that when y is small (0), x is small (0) about 5% of the time (=28/590) and when y is big (1), x is small (0) about 94% of the time (=386/410). Hence, there is clear and strong evidence of a negative relationship between x and y: when one is big or small the other is usually doing the opposite. [If you did compute the correlation, like in Required Problem (6) above, it is -0.8927.]

(c) The correlation is positive but not particularly strong. When x is small (0), y is small (0) about 37% of the time (=382/1041) and when x is big (1) y is small (0) about 16% of the time (=152/959). Alternatively, you could notice that when y is small (0), x is small (0) about 72% of the time (=382/534) and when y is big (1), x is small (0) about 45% of the time (=659/1466). Hence, they are more likely to move together than to move in opposite directions. [If you did compute the correlation, like in Required Problem (6) above, it is 0.2354.]

(d) The correlation is exactly zero. When x is small (0), y is small (0) exactly 70% of the time (=1638/2340) and when x is big (1) y is small (0) exactly 70% of the time (=462/660). Alternatively, you could notice that when y is small (0), x is small (0) exactly 78% of the time (=1638/2100) and when y is big (1), x is small (0) exactly 78% of the time (=702/900). Hence, the variables are not reacting to each other at all: regardless of whether the other is big or small, the chance of being big or small is totally unaffected. [If you did compute the correlation, like in Required Problem (6) above, it is exactly 0.]

(9) The reason for that big discrepancy is that debt as a percent of GDP in 2005 is *strongly positively correlated* with debt as a percent of GDP in 2010. Countries that had debt issues in 2005 are very likely to continue to have debt issues in 2010. In the weird hypothetical world where those two variables had nothing to do with each other then we would be able to simply add variances to get the variance of the difference (i.e. we would have found the variance of the difference is 2413 which does equal the sum of the variances).