Homework 22: ECO220Y – SOLUTIONS

Required Problems:

(1) (a) Two points determine a line. (Note: If this question is not very easy for you, study the "Math Review with a Diagnostic Quiz" http://homes.chass.utoronto.ca/~murdockj/eco220/MATH_REV_ECO220.pdf from Week 1.)

For the Weatherized line: Using a ruler and reasonable approximation, two points are (500, 4.3) and (1000, 7.6), which yields: $\widehat{MMBTU} = 1 + 0.0066HDD$.

For the Unweatherized line: Using a ruler and reasonable approximation, two points are (500, 4.3) and (1000, 8.7), which yields: $\widehat{MMBTU} = -0.1 + 0.0088HDD$.

(b) $MMBTU_{it} = \alpha + \beta HDD_{it} + \delta Unweatherized_{it} + \gamma Unweatherized_{it} * HDD_{it} + \varepsilon_{it}$, where $MMBTU_{it}$ is household i's natural gas consumption measured as MMBtu in winter month t, HDD_{it} is the measure of heating degree days (for the years covered by the research study) for household i in winter month t, and $Unweatherized_{it}$ is a dummy variable equal to 1 for an unweatherized home and 0 otherwise. (Note: It is equally correct to have included a dummy for weatherized rather than a dummy for unweatherized.)

(c) $\widehat{MMBTU}_{it} = 1 + 0.0066HDD_{it} - 1.1Unweatherized_{it} + 0.0022Unweatherized_{it} * HDD_{it}$

Note: If you wrote your model with unweatherized as the omitted category, then you should get:
$\widehat{MMBTU}_{it} = -0.1 + 0.0088HDD_{it} + 1.1Weatherized_{it} - 0.0022Weatherized_{it} * HDD_{it}$

(d) Unsurprisingly, homes that have been weatherized to help deal with cold winter temperatures, use less additional natural gas when confronted with an increase in cold winter temperatures compared to homes that have not been weatherized. The difference in slopes is considerable: for every 100 unit increase in heating degree days (which range from about 200 to 1500 in these data), a weatherized home uses an additional 0.7 MMBtu per month whereas an unweatherized home uses an additional 0.9 MMBtu per month, which is about one-third more natural gas use. It is a bit puzzling that these two lines cross at lower levels of heating degree days: it is not clear why a weatherized home should perform worse (i.e. consume more gas) than an unweatherized home in mild winter weather. Also, see the excerpt given with the question.

(2) Research question: "What is the effect of using marijuana as a teenager on hard drug use as an adult?" Regression model to assess the question: $harddrug_i = \beta_0 + \beta_1 mari_i + \varepsilon_i$ where $harddrug_i$ measures amount of hard drug use as an adult and $mari_i$ is a dummy variable that equals 1 if the person ever used marijuana as a teenager and 0 otherwise. To test for statistical significance in this simple regression we can test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 > 0$ with a t test. We would expect to reject the null in favor of the research hypothesis because it is well known that there is a positive correlation. However, these are observational data. Hence, we cannot conclude that the marijuana use causes the hard drug use. Other factors such as personality, upbringing, geographic location, etc. affect both of these variables. Since we've not controlled for these other variables, we would be mistaken if we attributed the positive effect marijuana seems to have on hard drug use to the actual marijuana use. Hence, despite the likely "statistical significance" we would find (i.e. that we would likely reject the null in favor of the research hypothesis), we CANNOT actually conclude that marijuana use increases hard drug use.

(3) (a) Let's call the variable research output O and grant money G. We can create dummy variables to pick up the different titles. AST = 1 if faculty member is an Assistant Professor, = 0 otherwise; ASC = 1 if faculty member is an Associate Professor, = 0 otherwise; $ASTT$= 1 if faculty member is an Assistant Professor, Teaching Stream, = 0 otherwise, $ASCT$= 1 if faculty member is an Associate Professor, Teaching Stream, = 0 otherwise. The omitted category is Full Professor. j indexes the sampled faculty members.

$$O_j = \beta_0 + \beta_1 G_j + \beta_2 AST_j + \beta_3 ASC_j + \beta_4 ASTT_j + \beta_5 ASCT_J + \varepsilon_j$$

Interpretation of the dummies is relative to the omitted category: Full Professor. If we thought that a Full Professor should have the highest productivity in terms of research output then we would expect the parameters on all of the dummies to be negative.

(b) Again we would have observational data. Grant money is not randomly assigned to faculty members. In fact, faculty members with the most promising lines of research are the most likely to be awarded lucrative grants. Hence we would expect that G will be positively correlated with the error. This violates Assumption #6 and will lead our OLS estimates to be biased: not equal to the true parameters is expectation. We could also argue that title is endogenous. Faculty members are not randomly promoted to Full Professor. Instead faculty members with the most promising research careers are promoted. Hence, the dummy variables are also related to the error. This type of regression analysis can be used for descriptive purposes but not to make any policy decisions about the effectiveness of grant money in promoting research. Hence, you will be unable to answer the research question by estimating any of the theoretical models you have proposed.

(c) To do this, we add interaction terms between grant and title.

$$O_j = \beta_0 + \beta_1 G_j + \beta_2 AST_j + \beta_3 ASC_j + \beta_4 ASTT_j + \beta_5 ASCT_J + \beta_6 AST_j * G_j + \beta_7 ASC_j * G_j + \beta_8 ASTT_j * G_j + \beta_9 ASCT_j * G_j + \varepsilon_j$$

In this specification, β_1 measures the marginal impact of grant money on the research output of a Full Professor. $(\beta_1 + \beta_6)$ measures the marginal impact of grant money on the research output of an Assistant Professor. $(\beta_1 + \beta_7)$ measures the marginal impact of grant money on the research output of an Associate Professor. And so on. We could test whether there is a differential effect of grant money between Full Professors and Associate Professors by simply testing whether β_7 is statistically significant. However, as described in Part (b) we could only use the estimates in a descriptive way. While the parameters measure the impact of grant money on research output, the parameter estimates (what we observe) are biased due to the classic problems of observational data (which are definitely present in this example).

(4) See: http://homes.chass.utoronto.ca/~murdockj/eco220/TT220_4_MAR17_SOLN.pdf

(5) See: http://homes.chass.utoronto.ca/~murdockj/eco220/TT220_5_APR18_SOLN.pdf

(6) (a) We notice a clear day of week pattern (with drops on Sundays). We also notice a clear increase during the playlist inclusion period. The most basic model specification (that ignores the upward time trend clearly visible before the beginning of the playlist inclusion date) is:

$$streams_t = \beta_0 + \beta_1 PlayIncl_t + \beta_2 Mon_t + \beta_3 Tue_t + \beta_4 Wed_t + \beta_5 Thu_t + \beta_5 Fri_t + \beta_6 Sat_t + \varepsilon_t$$

It includes a full set of day of week dummies with Sunday serving as the reference (omitted category). (You could have made any day of the week the reference category.) Also, we have a dummy for the Playlist Inclusion period with the non-Playlist Inclusion period serving as the reference category. Notice the t observation index given that the data featured in Figure 1 are time series data. Of course, generally the authors have lots more songs (not just "What Ifs") so they have panel data. More generally, we could include song fixed effects as well as a variable measuring time since release. (The time trend variable is not a dummy so that is not required to answer the specific question asked.)

(b) $E[\beta_0] > 0$: This measures the average streams (in millions) on Sundays outside of the Playlist Inclusion period. This is clearly positive given Figure 1. (While Sunday is lower than the other days, it still has positive streams.)

$E[\beta_1] > 0$: This measures how much streams (in millions) *differ* during the Playlist Inclusion period: Figure 1 clearly shows it is on average *higher* compared to outside of the Playlist Inclusion period.

$E[\beta_2] \dots E[\beta_6] > 0$: These measure how each day of the week *differs* from Sundays on average. Figure 1 shows that streaming drops off (spikes down) on Sundays: all of the other days of the week are on average *higher* than Sundays.