Homework 23: ECO220Y – SOLUTIONS

Required Problems:

(1) (a) (B)

(b) Cross-sectional, observational data. The unit of observation is a country. There are 7 observed x variables: log GDP,…, Divorce etc. There are many unobserved variables that affect mean happiness in a country ranging from climate to war.

(c) $H_0: \beta_{GDP} = \beta_{health} = \beta_{educ} = \beta_{soc} = \beta_{free} = \beta_{corr} = \beta_{div} = 0$ versus $H_1: not\ all\ coefficients\ are\ zero.$ Compute the test statistic: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{0.80/7}{(1-0.80)/(139-7-1)} = 74.9$. Using the F table we see that this test statistic is well above even the critical value for $\alpha = 0.001$ (a c.v. of about 3.77), which means that the multiple regression model in Panel B is overall highly statistically significant.

(d) To assess statistical significance of each coefficient we would use a t test. However, Table 3.1 does not report the standard errors of each coefficient. It only gives stars to measure the statistical significance *BUT* it very unusually does one-tailed tests. This is almost unheard of: *the* test of statistical significance is, by default, two-tailed. However, if a coefficient is not statistically significant for a one-tailed test (and the coefficient is of the expected sign), it will not be statistically significant for a two-tailed test. Hence, it looks like the coefficients on education and divorce are not statistically significant for sure.

(e) These data are observational, panel data. The unit of observation is an individual person in a specific year. (The same person appears as many different observations in these data because each person answered the survey every year for many years.) There are many dummy variables: you should list them.

(f) Table 1: Income is logged because these data (as other happiness data) show a non-linear relationship between happiness and income (specifically, diminishing marginal returns). The omitted category for each grouping of dummy variables is sometimes easy to identify. For example for the grouping of *Single*, *Widowed*, *Divorced*, and *Separated* the omitted category is clearly *Married*. For some other groupings you would need to see the original report. For example, the question regarding *Health* ranged from "Very Poor" to "Excellent." With this information in hand the omitted category for the health dummies is *Very Poor.* This is consistent with all of the health dummies having large positive and statistically significant coefficients: compared to very poor health the others are all better in terms of happiness (holding other included x variables fixed). There are 26 years of data (= 2009 – 1983). Hence there would be 25 time dummy variables included in the model*: remember to subtract one for the omitted category.* Standard errors are reported in parentheses. You should check which of the coefficients are statistically significant and interpret those that are.

Table 2: The *Fixed effects* refer to including a dummy variable for each person in the data. There would be a lot of these. The total number of observations is100,945 and we know that each person is followed for 26 years. Hence there are roughly 3,900 fixed effect variables. We can only say roughly because some people would certainly be missing values in some years. If you divide 100,945/26 it comes out to 3882.5, which is not even an integer. Remember one person would be the omitted category and not have a fixed effect variable. There is a "--" in the space for the coefficient on *Female* because that coefficient is not mathematically identified once you include a fixed effect for each person. Over time a person's sex is constant and the fixed effect picks up anything about that person (including sex). In contrast the other variables can change over time so including a constant term for each person does not remove all useful variation and preclude estimating the other coefficients. You should check which of the coefficients are statistically significant and interpret those that are. You should also discuss differences across the tables: the coefficients certainly change (e.g. income) when you control for differences across people with the fixed effects.

(g) (C)

(h) (E)

(2) Robust standard errors allow for a violations of some of the underlying assumptions. Most often they allow for a violation of the homoscedasticity assumption (i.e. they are still correct even if you have heteroscedasticity) or a violation of the no autocorrelation assumption (i.e. they are still correct even if you have autocorrelation). Robust standard errors are very commonly used: you will often see notes in tables of results from empirical papers that say things like "robust standard errors given in parentheses below the coefficient estimates." For regression, you do not need to know the formula for calculating them (which I did not give you) or the theory of how they are able to adjust for heteroscedasticity or autocorrelation: we did not cover these things. (You would learn about this in an upper level course.) However, you DO need to understand the link between the homoscedasticity assumption in regression analysis and the equal variances assumption in testing the difference between means (Section 14.5) and the link between robust standard errors in regression analysis and the general test of the difference between means that does not assume equal variances (Section 14.2). With any robust standard errors you can go ahead and conduct the usual t tests using the regular formulas given on our Aid Sheets.

(3) (a) 58.9% of the sample is female. 17.9% of the sample is in Program B.

(b) Use Regression #1: inter_grade-hat = 28.7 + 0.52*80 + 4.3 = 74.6. For the second question use Regression #2 (because not told gender and program): inter_grade-hat = 17.3 + 0.72*80 = 74.9. (Note: You could use Regression #1 plugging in the means of the unknown variable values but that is more work.)

(c) To answer use Regression #3 (because it is a question about differences across programs and not a question about differences across programs holding gender and intro grades fixed). The difference in the mean intermediate grades between Programs A and B is 8.11 marks (=14.68 – 6.57)

(d) NO. Females do not tend to earn lower grades in the intermediate course. In fact, Regression #4 shows that females earn slightly higher grades than males (but the difference is not statistically significant). The negative (and statistically significant) on females in Regression #1 means after we hold intro marks and the program constant. Regression #1 suggests that females do worse in the intermediate course than their intro grades and program would predict.

(e) The coefficient on introductory grade of 0.52 means that after controlling for program and sex, we observe that students who earned a grade that is one point higher in the introductory course on average earn a grade that is 0.52 points higher in the intermediate course. We use a descriptive interpretation because these data are clearly observational. The coefficient on Program A of 8.25 means that after controlling for introductory grades and sex, students in Program A on average earn grades that are 8.25 points higher compared to students in Program D. (The coefficients on Program B is interpreted similarly.) The coefficient on Program C of -0.48 is not statistically significant, which means that after controlling for introductory grades and sex, we have insufficient evidence of any systematic difference in the mean grades between students enrolled in Program C versus Program D. The coefficient on the dummy variable for females means that after controlling for introductory grades and program of study, females on average earn marks that are 2.3 points lower than males in the intermediate level course.

(f) The constant term in Regression #1 has no interpretation at all: no student earned a grade of 0 (or anything close) in the introductory course. Same answer for Regression #2. In contrast, the constant for Regression #3 has a clear interpretation: students in Program D had an average grade of 64.35 in the intermediate course. The constant in Regression #4 means that the average grade of males in the intermediate course is 74.9.

(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.