Econ 184b Professor Brainerd

## Solutions to the second set of midterm problems

### Part I: Multiple Choice

1. b
2. c
3. c
4. a
5. b
6. b
7. c
8. b
9. c
10. d

### Part II: True/False/Uncertain

1. True. Suppose you estimate a regression of the form $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3$ . This is a polynomial of order 3. To test whether the population regression function is linear, you would perform an *F*-test using 3-1 = 2 restrictions, i.e. you are testing $H_0$: $\beta_2 = 0 \; and \; \beta_3 = 0$ which has two restrictions.

2. False. Bias from an omitted variable only occurs if both the omitted variable belongs in the regression and the omitted variable is correlated with the included variable of interest. If the omitted variable $X_2$ is uncorrelated with $X_1$, there is no correlation between u and $X_1$, so the zero conditional mean assumption holds.

3. False/Uncertain. To infer a causal impact, we need to be convinced that all of the OLS assumptions hold, particularly the zero conditional mean assumption that X and u are uncorrelated.

4. False. This represents heteroskedasticity, which means that the variance of the regression error term $u_i$, conditional on the regressors, is not constant. Heteroskedasticity means that the standard errors are too small (if uncorrected for heteroskedasticity) but it does not cause the OLS estimates to be biased.

5. False. The regression is linear in the parameters but is nonlinear in X, since X is in logs.

6. True. This is a quadratic regression function; the effect of a change in X on Y is given by $\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$. This shows that as X changes, the effect of a change in X on Y will change as well. (One could also answer this question by illustrating the change in Y with an example.)

7. False. Since you have cross-sectional data, each state is observed only once. You can only include state fixed effects if you have panel data, i.e. each state is observed at least twice.

Part III: Short answer problems

1. a. One additional year of education increases earnings by 8.3 percent.

To test whether years of high school education have a different effect than years of college education, the regression would include separate variables for years of each type of schooling (for example, *YearsHS* and *YearsCollege).* Then one would test the equality of the coefficients: $H_o$: $\beta_1 = \beta_2$ vs $H_1$: $\beta_1 \neq \beta_2$ (either by using the "test" command in Stata or by transforming the regression). A "diploma effect" could be studied by creating a binary variable for a high school diploma, a junior college diploma, a B.A. or B.Sc. diploma, and so forth.

b. For the first person, the *Exper* variable increases from 40-12-6 = 22 to 23, and results in a 1.1 percent earnings increase. You can calculate this either by plugging in 22 and 23 into the regression and calculating the change: $(.033\text{x}23-.0005\text{x}23^2)-(.033\text{x}22-.0005\text{x}22^2)$ or by calculating ∂Y/∂Exper = .033 – 2x.0005Exper and plugging in Exper=22 and Exper = 42. For the 60 year old, there is an expected decrease of 1 percent.

c. To find the years of work experience associated with maximum log earnings, calculate calculating ∂Y/∂Exper = .033 – 2x.0005Exper, set equal to zero and solve for Exper:

.033 – 2x.0005Exper = 0
Exper = 33 years

d. The t-statistic for *Exper* is .033/.006 = 5.5 and for *Exper*$^2$ it is -.0005/.0001 = 5.0. Both coefficients are significant at the 1% significance level since they exceed the critical value of 2.58. The fact that the coefficient on the education variable hardly changed suggests that education and experience are not highly correlated.

2. There are two restrictions, namely $H_0 : \beta_{meal\_pct} = 0, \beta_{calw\_pct} = 0$. The *F*-statistic is

$$F = \frac{(.80- .71)/2}{(1-.80)/(420-5-1)} = 93.2$$

The 5% critical value from the $F_{2,\infty}$ distribution is 3.00. Hence we easily reject the two restrictions at the 5% level of significance.

3. a. Heteroskedasticity is nonconstant error variance. If homoskedasticity is incorrectly assumed and homoskedasticity-only standard errors are used, they will be incorrect (usually too small) and inference will be faulty. Heteroskedasticity-robust standard errors are used to ensure that the standard errors are correctly calculated.

b. T-statistics are given in the question, so comparing these with the critical value 1.96 we see that FEMPAR, lnGNP, BRITCOL, and FREEDOMS are statistically different than zero (and NOTCOL is very close).

c. The researcher might leave statistically insignificant variables in the regression for a number of reasons, including: economic theory suggests their inclusion; omitting relevant variables can lead to omitted variable bias; and it is often better to leave in variables to show the reader their effect.

d. Define a dummy variable D = 1 if the country is a European country, = 0 otherwise. Then interact FEMPAR with D: D*FEMPAR and include it in regression (along with the dummy variable itself). Then test the coefficient on the interaction term using a t-test. If the null hypothesis is rejected, this provides evidence that the effect of proportion female does differ for European countries.

e. Multicollinearity is suggested by insignificant t-statistics and incorrect signs. Here, a number of variables are insignificant (EDUC, CATHOLIC, MUSLIM, and ETHNIC) . For multicollinearity to be affecting the results, one must make a case that one or more of these insignificant variables is highly correlated with the other regressors, and/or that the signs are incorrect. A variety of answers are possible here; as long as the reasoning is clearly explained "yes" and "no" answers would both be acceptable.

4. a. In the first regression, the coefficient on *comp_stu* gives the change in test scores when the number of computers per student in the district is increased by one. In the second regression, the coefficient gives the gives the change in test scores when the number of computers per student in the district is increased by one, *holding average income in the district constant*.

b. The second regression is better for a number of reasons. First, on theoretical grounds *avginc* should be included in the regression since it is correlated with test scores. Second (as discussed in part c), omitting *avginc* causes omitted variable bias. Third, the adjusted R-squared is significantly higher in the second regression, indicating that the model provides a better fit of the data.

c. The large change in the size of the coefficient on *comp_stu* indicates that *compu_stu* and *avginc* are positively correlated (school districts with more computers per student are wealthier). Since *avginc* is likely to be positively correlated with test scores, when it is omitted from the regression the coefficient on *comp_stu* is picking up the effect of both computers and income on test scores.

d. The negative coefficient on *avginc2*, combined with the positive coefficient on *avginc*, tells us that the relationship between average income and test scores is concave: as average income increases, test scores rise but at a decreasing rate. The coefficient on *avginc2* can't be interpreted directly without including the effect of *avginc* itself.

e. To test whether there is a linear relationship between *avginc* and *testscr*, test the null hypothesis $H_o$: $\beta_3 = 0$ vs $H_1$: $\beta_3 \neq 0$. The t-statistic is -7.28 > |1.96| so we reject the null hypothesis that the relationship is linear.

f. The coefficient on the interaction term *comp_inc* tells you that the effect on test scores of adding one more computer per student depends on the average income level of the school district. Since the coefficient is negative this indicates that the increasing computers raises test scores, but this effect is smaller as average income increases: $\partial$Test score/$\partial$ Comp_stu = 109.9 – 4.56 x Avginc.

5. Clearly omitted variable bias is likely to be an issue here. For example, one would want to include a measure of family resources (income) in the regression, because this is likely to be correlated with HIV status (individuals from poorer families are more likely to be HIV positive) and individuals from poorer families are less likely to be in school. Omitting this variable would cause the estimate of $\beta_1$ to be biased downward, since HIV is picking up the effect of family income ( in other words, because family income and HIV are negatively correlated and family income and school enrollment are positively correlated). A second potential bias is measurement error, in which individuals may not know their HIV status. This will cause the estimate to be biased; if one assumes classical measurement error the estimate will be attenuated, i.e. biased toward zero. Again, other answers are possible and acceptable as long as the case is presented clearly and correctly regarding the potential sources of bias.

6. a. 95% confidence interval : (.038 ± 1.96 × .025) = (.038 ± .049) = (-.011, .087)

Effect of an increase of 0.5 hours of study time: 0.5 × (-0.11, .087) = (-.0055, .0435) is the 95% confidence interval for the change in GPA from 0.5 hours more of study time.

b. Regressions (2) and (3) are both well-suited to modeling declining marginal returns to studying, whereas the linear specification in regression (1) is not. In Regression (1) the effect of one hour of studying does not depend on the number of hours of studying; in other words the linear model has constant marginal returns to studying.

The linear-log specification (Regression (2)) and the cubic specification (Regression (3)) both allow for declining marginal returns to studying. Comparing (2) and (3), Regression (2) is preferred because it has a slightly better fit as indicated by the higher adjusted R2. The coefficient is also easier to interpret in Regression (2) than Regression (3).

c. Comparing the list of majors in Table 1 and Table 2 indicates that the omitted major in Table 2 is social sciences majors. Therefore the coefficient on the major_hum binary variable is the mean difference in GPA1 between social science and humanities majors, holding constant the other variables in the regression. To test the hypothesis that the two majors have the same GPA1 on average (holding constant the other variables) we need to test whether the coefficient on major_hum is zero, against the alternative that it is nonzero. The double asterisk in Table 2 indicates that the hypothesis is rejected at the 1% significance level (because the asterisks are included in the table it was unnecessary to compute the t-statistic).

d.  Student ability is an omitted variable that could bias the coefficient.  A student's ability is plausibly related to studying; for example perhaps more able students need to study less.  Ability is also likely to be positively correlated with GPA1 since more able students likely get higher grades, all else equal.  Therefore the omission of ability would cause omitted variable bias.  In this case, with Corr(ability, study time) < 0 and Corr(ability, GPA1) > 0 we expect   to be biased downward.

Other omitted variables (e.g. income, sleep, campus job) are possible; a full-credit answer explained both correlations and the likely direction of bias in  .

e.  As we discussed in class there are five possible threats to internal validity:  (i) omitted variables bias; (ii) incorrect functional form (iii) errors-in-variables; (iv) sample selection bias; and (v) simultaneous causality (it was not necessary to list these in the answer to this question).

The investigations in Regressions (2) and (3) suggest that incorrect functional form is unlikely to be a significant source of bias in this situation, i.e. it is not a significant threat to internal validity.

Possibly measurement error in Study time resulting from students not keeping accurate records of their study time (or perhaps multitasking while studying).  This would lead to bias in  , or attenuation bias if the measurement error was random.

Sample selection bias arises when there is a selection process related to the dependent variable. In this case, the first-semester data consists of all 210 randomly selected students so there is no sample selection bias.

Simultaneous causality bias is the most serious threat to internal validity (after OVB) in the regressions in Table 2.  This would occur if GPA1 affects study time, which seems likely: students who are doing poorly might be motivated to study more.  This reverse effect would suggest a negative coefficient (worse grades induce more studying).  The original effect of studying on GPA1 is plausibly positive (more studying leads to better grades), which suggests that the effect estimated by OLS combines these two and could lead to a coefficient near zero, with these two effects offsetting each other in the data.  This is a major threat to internal validity and is a plausible explanation for the puzzlingly small effect of Study estimated in Regressions (1) and (2).

f.  Including individual fixed effects would eliminate omitted variable bias arising from any omitted variable that varies across students, but not over time.  If the proposed omitted variable was an omitted student characteristic, such as student ability, that plausibly does not change over time, including the student fixed effects would eliminate this source of OVB.

A full-credit answer explained whether or not the omitted variable varies over both students and time, and therefore whether or not the fixed effects estimator eliminates this source of OVB.

g.  Disagree.  The students were not lost at random.  Presumably an important reason that they did not complete the 8 surveys is that they dropped out, and one reason for dropping out is poor grades. So there was a selection process (dropping out) that determined the data that is related to the dependent variable (GPA). This means that there will be selection bias in .

h. The OLS estimator in Table 2 is subject to two important threats to internal validity: OVB arising from unobserved individual characteristics, and simultaneous causality bias.

The FE estimator solves the OVB problem that results from omitted personal characteristics that do not vary over time (such as ability) and thus improves upon the estimates in Table 2. However, the problem of simultaneous causality still remains. In addition there is the additional problem of sample selection bias in the panel data. Whether the FE estimator is preferred over the OLS estimator in this case is a matter of judgment about the relative importance of OVB in the OLS estimates versus sample selection bias in the FE estimates; the judgment could go either way.