Econ 184b Professor Brainerd

Practice problems for the second midterm exam

Part I: Multiple Choice

1. Under imperfect multicollinearity
   a. The OLS estimator cannot be computed.
   b. two or more of the regressors are highly correlated.
   c. the OLS estimator is biased even in samples of $n > 100$.
   d. the error terms are highly, but not perfectly, correlated.
   e. none of the above.

2. Imagine you regressed earnings of individuals on a constant, a binary variable ("Male") which takes on the value 1 for males and is 0 otherwise, and another binary variable ("Female") which takes on the value 1 for females and is 0 otherwise. Because females typically earn less than males, you would expect
   a. the coefficient for Male to have a positive sign, and for Female a negative sign.
   b. both coefficients to be the same distance from the constant, one above and the other below.
   c. none of the OLS estimators to exist because there is perfect multicollinearity.
   d. this to yield a difference in means statistic.
   e. none of the above.

3. The adjusted R^2, or $\overline{R}^2$, is given by

   a. $1-\frac{n-2}{n-k-1}\frac{SSR}{TSS}$.
   b. $1-\frac{n-1}{n-k-1}\frac{ESS}{TSS}$.
   c. $1-\frac{n-1}{n-k-1}\frac{SSR}{TSS}$.
   d. $\frac{ESS}{TSS}$.
   e. none of the above

4. If you reject a joint null hypothesis using the F-test in a multiple hypothesis setting, then
   a. a series of t-tests may or may not give you the same conclusion.
   b. the regression is always significant.
   c. all of the hypotheses are always simultaneously rejected.
   d. the F-statistic must be negative.
   e. none of the above.

5. The interpretation of the slope coefficient in the model $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$ is as follows:
   a. a 1% change in X is associated with a β_1% change in Y.

b. a 1% change in X is associated with a change in Y of 0.01 β_1 .

c. a change in X by one unit is associated with a 100 β_1 % change in Y.

d. a change in X by one unit is associated with a β_1 change in Y.

6. To decide whether $Y_i = \beta_0 + \beta_1 X + u_i$ or $\ln(Y_i) = \beta_0 + \beta_1 X + u_i$ fits the data better, you cannot consult the regression R^2 because

a. ln(*Y*) may be negative for 0<*Y*<1.

b. the *TSS* are not measured in the same units between the two models.

c. the slope no longer indicates the effect of a unit change of *X* on *Y* in the log-linear model.

d. the regression R^2 can be greater than one in the second model.

7. The components of internal validity are

a. a large sample, and BLUE property of the estimator.

b. a regression R^2 above 0.75 and serially uncorrelated errors.

c. unbiasedness and consistency of the estimator, and desired significance level of hypothesis testing.

d. nonstochastic explanatory variables, and prediction intervals close to the sample mean.

e. (a) and (c)

8. Simultaneous causality

a. means you must run a second regression of X on Y.

b. leads to correlation between the regressor and the error term.

c. means that a third variable affects both Y and X.

d. cannot be established since regression analysis only detects correlation between variables.

9. In the fixed effects regression model, using (n - 1) binary variables for the entities, the coefficient of the binary variable indicates

a. the level of the fixed effect of the ith entity.

b. either 0 or 1.

c. the difference in fixed effects between the ith and the first entity.

d. the response in the dependent variable to a percentage change in the binary variable.

e. (a) and (b)

10. Which of the following does NOT result in the inconsistency of the OLS estimator?

a. Omitted variable bias

b. Simultaneous causality

c. Using a linear regression when the expectation of Y is quadratic in X:

$E[Y \mid X] = \beta_0 + \beta_1 X_1 + \beta_2 X^2_2$ with $\beta_2 \neq 0$.

d. Using homoskedasity-only standard errors when the errors are heteroskedastic

e. None of the above

Part II: True/False/Uncertain

1. To test whether or not the population regression function is linear rather than a polynomial of order *r*, one should use the test of *(r*-1) restrictions using the *F*-statistic.

2. If the true model is $Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + u$ but you estimate $Y = \beta_1 + \beta_2 X_1 + u$, your estimate of β_2 will always be biased.

3. If $\hat{\beta}_1$ is significantly different from zero at a 5% significance level, then X has a causal impact on Y.

4. Suppose $Var(u_i)$ is positively correlated with X in the population. In this case, the OLS estimators are biased. (Unless otherwise stated, assume all relevant assumptions hold.)

5. $\ln(Y) = \beta_1 + \beta_2 \ln(X) + u$ is linear in both the parameters and the variables.

6. Suppose we have the following model: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$. The effect on Y of a change in X will depend on the level of X.

7. Suppose you are working with a cross-sectional sample of 50 states. For each state you have information on the beer tax rate and the number of drunk driving arrests in the past year. T/F/U: If you are trying to estimate the causal impact of beer taxes on drunk driving, you should include state fixed effects in your model.

Part III: Short answer problems

1. Earnings functions attempt to find the determinants of earnings, using both continuous and binary variables. One of the central questions analyzed in this relationship is the returns to education.

a. Collecting data from 253 individuals, you estimate the following relationship

$$\underset{}{ln(Earn)} = \underset{(0.14)}{0.54} + \underset{(0.011)}{0.083}\ Educ, \quad R^2 = 0.20,\ SER = 0.445$$

where *ln(Earn)* is the log of average hourly earnings and *Educ* is years of education.

What is the effect on earnings of an additional year of schooling? If you had a strong belief that years of high school education were different from college education, how would you modify the equation? What if your theory suggested that there was a "diploma effect"?

b. You read in the literature that there should also be returns to on-the-job training. To approximate on-the-job training, researchers often use a potential experience variable, which is defined as Exper = Age – Educ – 6. You incorporate the experience variable into your original regression

$$ln(Earn) = \underset{(0.16)}{-0.01} + \underset{(0.012)}{0.101}\, Educ + \underset{(0.006)}{0.033}\, Exper - \underset{(0.0001)}{0.0005}\, Exper^2, \quad R^2 = 0.34,\ SER = 0.405$$

What is the effect of an additional year of experience for a person who is 40 years old and had 12 years of education? What about for a person who is 60 years old with the same education background?

c. At what number of years of experience are log Earnings maximized?

d. Test for the statistical significance of each of the coefficients of the added variables. Why has the coefficient on education changed so little?

2. Consider the following regression output for an unrestricted and a restricted model.

Unrestricted model:

Dependent Variable: TESTSCR
Included observations: 420

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	658.47	7.68	85.73	0.00
STR	-0.76	0.23	-3.27	0.00
EL_PCT	-0.19	0.03	-5.62	0.00
LOG(AVGINC)	11.69	1.74	6.71	0.00
MEAL_PCT	-0.37	0.04	-9.53	0.00
CALW_PCT	-0.07	0.06	-1.21	0.23

R-squared	0.80	Mean dependent var	654.16
Adjusted R-squared	0.79	S.D. dependent var	19.05
S.E. of regression	8.64	Akaike info criterion	7.16
Sum squared resid	30888.4	Schwarz criterion	7.22

Restricted model:

Dependent Variable: TESTSCR
Included observations: 420

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	593.48	6.96	85.32	0.00
STR	-0.39	0.27	-1.42	0.16

EL_PCT	-0.43	0.03	-14.34	0.00
LOG(AVGINC)	28.36	1.40	20.32	0.00

R-squared	0.71	Mean dependent var	654.16
Adjusted R-squared	0.71	S.D. dependent var	19.05
S.E. of regression	10.26	Akaike info criterion	7.50
Sum squared resid	43792.4	Schwarz criterion	7.54
Log likelihood	-1571.8	F-statistic	342.98
Durbin-Watson stat	1.30	Prob(F-statistic)	0.00

Calculate the homoskedasticity only F-statistic and determine whether the null hypothesis can be rejected at the 5% significance level.

3. Using data from 93 countries, a researcher estimates the following model *(heteroskedasticity-robust t-statistics in parentheses)*

$GRAFT_i = -4.702 + 2.456\ FEMPAR_i + 0.478 \ln GNP_i + 0.003\ EDUC_i - 0.281\ CATHOLIC_i$
(-8.177) (3.270) (5.311) (0.086) (-1.767)

$-0.152\ MUSLIM_i + 0.481\ BRITCOL_i + 0.312\ NOTCOL_i + 0.141\ ETHNIC_i + 0.092\ FREEDOMS_i$
(-0.792) (3.672) (1.950) (0.705) (2.788)

adj. $R^2 = 0.75$, n = 93

where
$GRAFT_i$ = index of the amount of perceived corruption in country i, with higher values meaning less corruption
$FEMPAR_i$ = proportion of the Parliament that is female in country i
ln GNP_i = logarithm of GNP in country i
$EDUC_i$ = average years of schooling in country i
$CATHOLIC_i$ = proportion of population Catholic in country i
$MUSLIM_i$ = proportion of population Muslim in country i
$BRITCOL_i$ = 1 if former British colony
$NOTCOL_i$ = 1 if never colonized
$ETHNIC_i$ = proportion of population in largest ethnic group in country i
$FREEDOMS_i$ = index of political freedoms in country i, with higher values meaning more political freedoms

a. Briefly explain what heteroskedasticity is, and why the researcher used heteroskedasticity-robust standard errors in his estimation.

b. Which of the coefficients are significantly different from 0 at the 5% significance level (two-tailed test)?

c. Why might the researcher have chosen to leave variables in the regression that are not statistically significant?

d. Briefly explain how you could test whether the effect of **proportion of the parliament that is female** on the corruption measure differs for European countries.

e. Is multicollinearity a problem for this regression? Why or why not?

4. This dataset contains data on 420 school districts in California:

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     testscr |       420    654.1565    19.05335     605.55     706.75
    comp_stu |       420    .1359266    .0649558          0   .4208333
      avginc |       420    15.31659     7.22589      5.335     55.328
```

`testscr` is the average test score of 5th graders in the district
`comp_stu` is the average number of computers per student in the district
`avginc` is the average income in the district

Regression 1:

```
. reg testscr comp_stu

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  1,   418) =   33.05
       Model |                 1  11146.6436           Prob > F      =  0.0000
    Residual |   140962.95   418  337.231938           R-squared     =
-------------+------------------------------           Adj R-squared =  0.0711
       Total |  152109.594   419  363.030056           Root MSE      =  18.364

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    comp_stu |   79.40485   13.81145     5.75   0.000
       _cons |   643.3633              309.28   0.000     639.2743    647.4523
------------------------------------------------------------------------------
```

Regression 2:

```
. reg testscr comp_stu avginc

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  2,   417) =  231.05
       Model |  79955.8549     2  39977.9274           Prob > F      =  0.0000
    Residual |  72153.7387   417  173.030549           R-squared     =  0.5256
-------------+------------------------------           Adj R-squared =  0.5234
       Total |  152109.594   419  363.030056           Root MSE      =  13.154

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    comp_stu |   40.22145   10.08643     3.99   0.000     20.39487    60.04803
      avginc |   1.808115   .0906701    19.94   0.000     1.629887    1.986342
       _cons |   620.9952   1.865068   332.96   0.000     617.3291    624.6613
------------------------------------------------------------------------------
```

a. How does the interpretation of *comp_stu* change when we add *avginc* to the regression?

b. Which regression is better? Why?

c. The coefficient on *comp_stu* went down significantly with the inclusion of *avginc* to the model. What does this suggest about the correlation between *comp_stu* and *avginc*? Explain.

NEW VARIABLE:
. gen avginc2=avginc*avginc

```
. reg testscr comp_stu avginc avginc2

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  3,   416) =
       Model |  88100.6286     3  29366.8762           Prob > F      =  0.0000
    Residual |   64008.965   416  153.867704           R-squared     =  0.5792
-------------+------------------------------           Adj R-squared =
       Total |  152109.594   419  363.030056           Root MSE      =  12.404

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    comp_stu |   45.50465   9.539196     4.77   0.000     26.75362    64.25569
      avginc |   3.874927   .2966648    13.06   0.000     3.291778    4.458076
     avginc2 |  -.0445311   .0061206    -7.28   0.000    -.0565623   -.0324998
       _cons |   601.3871    3.21818   186.87   0.000     595.0612     607.713
------------------------------------------------------------------------------
```

d. What is the interpretation of the coefficient on *avginc2*?

e. Would you accept or reject the null hypothesis that there is a linear relationship between *avginc* and *testscr*? Explain.

NEW VARIABLE:
. gen comp_inc=comp_stu*avginc

```
. reg testscr comp_stu avginc comp_inc

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  3,   416) =  163.94
       Model |  82406.4226     3  27468.8075           Prob > F      =  0.0000
    Residual |   69703.171   416    167.5557           R-squared     =  0.5418
-------------+------------------------------           Adj R-squared =  0.5385
       Total |  152109.594   419  363.030056           Root MSE      =  12.944

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    comp_stu |   109.9386   20.75689     5.30   0.000     69.13712    150.7401
      avginc |    2.59762   .2248998    11.55   0.000     2.155539    3.039702
    comp_inc |  -4.558676   1.192024    -3.82   0.000    -6.901818   -2.215535
       _cons |    609.333   3.559197   171.20   0.000     602.3367    616.3292
------------------------------------------------------------------------------
```

f. What is the interpretation of the coefficient on *comp_inc*?

5. Suppose you were interested in estimating the impact of HIV on school enrollment in South Africa. Consider the following model:

$$School_i = \beta_0 + \beta_1 HIV_i + u_i$$

Where $School_i$ is a dummy variable indicating whether or not individual *i* is enrolled, HIV_i is a dummy variable indicating whether the individual had contracted HIV, and u_i is an error term. You decide to estimate the model using OLS, and estimate the following coefficients: $\widehat{\beta_0} = 0.80$; $\widehat{\beta_1} = -0.08$. (When the dependent variable is binary, the coefficient gives the change in probability that Y=1 for a unit change in X).

Can you interpret these estimates as causal? Provide TWO reasons why these estimates may not be internally valid. Give a brief explanation of the potential biases and please be specific in your explanation.

6. Does studying improve your grades? You will examine this question using a panel data set from a small college which surveyed 210 students randomly selected from the first-year class. These students were re-surveyed at the end of every semester for eight semesters. This analysis looks at grades (semester GPA, from the Registrar), average study hours per day, and demographic data. Study hours were calculated from daily time diaries kept by the students. The data are summarized in Table 1.

The regressions in Table 2 are estimated using only data for the first semester of freshman year; this is a cross-sectional data set with n = 210 observations.

Table 1. Variable definitions and summary statistics, first semester data only
Unit of observation: individual college student, n = 210

Variable	Definition	Mean	Std. dev.
GPA1	first semester college GPA (0-4 scale)	3.004	.652
Study	study time, in hours per day (semester average)	3.427	1.631
male	=1 if male, =0 if female	0.480	0.034
black	=1 if black, =0 otherwise	0.171	0.026
ACT	score on ACT standardized admissions test (max=36)	23.38	3.71

Major variables: these are "most likely" majors reported during the first-semester survey, divided into 5 mutually exclusive and exhaustive groups:

major_ag	=1 if agriculture major, =0 otherwise	0.116	0.018
major_hum	=1 if humanities major, =0 otherwise	0.233	0.029
major_math	=1 if math or science major, =0 otherwise	0.252	0.028
major_prof	=1 if pre-professional major (nursing, pre-law)=0 otherwise	0.189	0.022
major_ss	=1 if social sciences major, =0 otherwise	0.210	0.018
health_bad	=1 if health is bad (self-reported), =0 otherwise	0.067	0.017
health_exc	=1 if health is excellent, =0 otherwise	0.371	0.033
health_avg	=1 if health is average, =0 otherwise	0.562	0.036

Table 2. Study Hours and GPA: Regression Results (n = 210)

	(1)	(2)	(3)
Dependent variable	*GPA1*	*GPA1*	*GPA1*
Study	.038 (.025)	–	.038 (.025)
ln(Study)	–	.114 (.086)	–
$Study^2$	–	–	-.012 (.054)
$Study^3$	–	–	.0076 (.0064)
Male	-.132 (.084)	-.133 (.082)	-.130 (.082)
Black	-.220+ (.122)	-.224+ (.121)	-.224+ (.123)
ACT	.062** (.013)	.060** (.015)	.061** (.014)
major_ag	.834** (.298)	.830** (.301)	.831** (.299)
major_hum	.796** (.283)	.794** (.281)	.792** (.280)
major_math	.643* (.280)	.641* (.285)	.642* (.285)
major_prof	.664* (.292)	.667* (.293)	.667* (.291)
health_bad	.019 (.166)	.020 (.165)	.019 (.167)
health_exc	.127 (.086)	.123 (.085)	.128 (.087)
constant	.719+ (.408)	.589 (.411)	.678 (.401)
F-statistics testing the hypothesis that all coefficients are zero for:			
$Study^2$, $Study^3$	–	–	2.03
All major variables (major_ag,...major_prof)	3.62	3.64	3.61
Regression summary statistics: R^2	.273	.274	.276
Adjusted R^2	.251	.252	.250

Heteroskedasticity-robust SEs in parentheses; statistical significance is at the +10%, *5%, **1% levels.

a. A student is considering increasing studying from 2 ½ hours per day to 3 hours per day. Compute a 95% confidence interval for the predicted effect of this increase using Table 2, regression (1).

b. Suppose you believe that the relationship between studying and grades is nonlinear. Which specification is the preferred specification for modeling this nonlinear relationship, (1), (2), or (3)? Briefly explain your reasoning.

c. Using Regression (1), test the hypothesis (at the 5% significance level) that humanities majors and social sciences majors have the same GPA on average, holding constant hours studying, sex, race, ACT score, and health.

d. Provide an example of a variable omitted from regression (1) that plausibly would result in omitted variable bias in the estimated effect on *GPA1* of *Study*, controlling for the other variables in regression (1). In what direction do you expect the coefficient on *Study* to be biased, and why?

e. Besides omitted variable bias, what is the most important threat to the internal validity of the regressions in Table 2? Explain your reasoning. (Choose <u>one</u> threat only.)

Now consider the full panel data set, with T=8 observations for each student. Of the 210 entering freshman in the study, 176 completed all 8 semesters. Thus for the panel data set n = 176, and the total number of observations is 8 × 176 = 1408.

The questions below consider regressions using the full panel data set (n = 176, T = 8). There is no table of results for these questions.

Let $\widehat{\beta_1^{FE}}$ denote the fixed effects regression estimator of β_1 in the regression model:

$$GPA_{it} = \beta_1 Study_{it} + \alpha_i + u_{it} ,$$

where i = 1,..., 176 are individual fixed effects for each student in the data set, and t = 1,...,8.

f. Restate the omitted variable you provided in your response to part (d) of this question. Would $\widehat{\beta_1^{FE}}$ be subject to omitted variable bias due to the omission of this variable from the regression? Explain.

g. Consider the following statement: "Although the number of students in the panel data set (n = 176) is smaller than in the first semester data set (n = 210), this loss of 34 students will not affect the internal validity of the estimate of $\widehat{\beta_1^{FE}}$." Do you agree or disagree? Explain.

h. Taking everything into account, which of the two estimators of the effect on *GPA* of *Study* would you prefer: the OLS estimator in Table 2, or the fixed effects estimator $\widehat{\beta_1^{FE}}$? Explain.