Problem Set 3

Welcome

We've been so busy with labs and research ideas that we have *three* chapters of coverage to work through! Note that this is your last problem set before the next exam

You can download the data file you need for question 6 here, along with information on the variable definitions here

What do I submit?

- Your written up answers to exercise questions. If you work on a piece of paper, please scan using some sort of phone software (like Microsoft Lens or Adobe Scan) rather than just taking a picture.
- A do-file that runs your Stata analysis (for question 4).
- A log file that includes the output from running your do-file (for question 4).

Exercises

1. Suppose that we want to estimate the effects of alcohol consumption ($alcohol$) on college grade point average ($colGPA$). In addition to collecting information on alcohol consumption and grade point averages, we also obtain attendance information (say, percentage of lectures attended, $attend$). A standardized test score (say, SAT) and high school GPA ($hsGPA$) are also available.

   a. Should we include $attend$ along with alcohol as explanatory variables in a multiple regression model? What would be the interpretation of $\beta_{alcohol}$ if we did?
   b. Should SAT and $hsGPA$ be included as explanatory variables? Explain.

2. A research plans to study the casual effect of police on crime, using data from a random sample of U.S. counties. She plans to regress the county's crime rate on the (per capita) size of the county's police force.

   a. Explain why this regression is likely to suffer from omitted variable bias. Which variables would you add to the regression to control for important omitted variables?
   b. Use your answer to (a) and the expression for omitted variable bias (from the slides or textbook) to determine whether the regression will likely over- or underestimate the effect of police on the crime rate. (That is, is $\hat{\beta}_1 > \beta_1$, or that $\hat{\beta}_1 < \beta_1$?)

3. Critique each of the following proposed research plans. Your critique should explain any problems with the proposed research and describe how the research plan might be improved. Include discussion of any additional data that needs to be collected, and the appropriate statistical techniques for analyzing those data.

a. A researcher is interested in determining whether a large aerospace firm is guilty of gender bias in setting wages. To determine potential bias, the researcher collects salary and gender information for all of the firm's engineers. The researcher then plans to conduct a "difference in means" test to determine whether the average salary for women is significantly less than the average salary for men

b. A researcher is interested in determining whether time in prison has a permanent effect on a person's wage rate. He collects data on a random sample of people who have been out of prison for at least 15 years. He collects similar data on a random sample of people who have never served time in prison. The data set includes information on each person's current wage, education, age, ethnicity, gender, tenure (time in current job), occupation, and union status, as well as whether the person has ever been incarcerated. The researcher plans to estimated the effect of incarceration on wages by regressing wages on an indicator variable for incarceration, including in the regression the other potential determinants of wages such as education, tenure, union status, and so on.

4. Consider a dataset that contains information on 4700 full-time full-year workers. The highest educational achievement for each worker was either a high school diploma or a bachelor's degree. The worker's ages ranged from 25 to 45 years. The data set also contains information on the region of the country where the person lived, marital status, and number of children. See below for variable definitions.

a. Is the college-high school earnings difference estimated from this regression statistically significant at the 5% level? Construct a 95% confidence interval of the difference.

b. Do there appear to be important regional differences in hourly wearnings? Use an appropriate hypothesis test to explain your answer.

Dependent Variable: average hourly earnings (*AHE*).

Regressor	(1)	(2)	(3)
College (X_1)	5.90	5.92	5.88
	(0.23)	(0.23)	(0.23)
Female (X_2)	−2.85	−2.83	−2.83
	(0.22)	(0.22)	(0.22)
Age (X_3)		0.31	0.31
		(0.06)	(0.05)
Northeast (X_4)			0.75
			(0.32)
Midwest (X_5)			0.65
			(0.30)
South (X_6)			−0.29
			(0.28)
Intercept	13.71	4.75	4.05
	(0.15)	(1.13)	(1.14)

Summary Statistics

F-statistic for regional effects = 0			6.24
SER	6.77	6.72	6.71
R^2	0.190	0.205	0.210
n	4700	4700	4700

Variable	Definition
AHE	average hourly earnings (in 2005 dollars)
College	1 if college, 0 if high school
Female	1 if female, 0 if male
Age	age (in years)
Ntheast	1 if Region = Northeast, 0 otherwise
Midwest	1 if Region = Midwest, 0 otherwise
South	1 if Region = South, 0 otherwise
West	1 if Region = West, 0 otherwise

5. Consider the regression results below and do the following:

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28**	−1.10*	−1.00**	−1.31*	−1.01*
	(0.52)	(0.43)	(0.27)	(0.34)	(0.27)
Percent English learners (X_2)		−0.650**	−0.122**	−0.488**	−0.130**
		(0.031)	(0.033)	(0.030)	(0.036)
Percent eligible for subsidized lunch (X_3)			−0.547*		−0.529*
			(0.024)		(0.038)
Percent on public income assistance (X_4)				−0.790**	0.048
				(0.068)	(0.059)
Intercept	698.9**	686.0**	700.2**	698.0**	700.4**
	(10.4)	(8.7)	(5.6)	(6.9)	(5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\overline{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix (4.1). Heteroskedasticity-robust standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

a. Construct the R^2 for each of the regressions
b. Construct the homoskedasticity-only F-statistic for testing $\beta_3 = \beta_4 = 0$ shown in column (5). Is the statistic significant at the 5% level?
c. Construct a 99% confidence interval for β_1 for the regression in column (5)

6. Download the dataset , which contains data on average growth rates from 1960 through 1995 for 65 countries, along with variables that are potentially related to growth. You can download a detailed description of all variable names is available here. For all questions, exclude Malta, which has an extremely high trade share. Estimate a regression of growth on tradeshare, yearsschool, rev_coups, assassinations, and rgdp60, with heteroskedasticity-robust standard errors.

a. What is the value of the coefficient on rev_coups? Interpret the value of this coefficient. Is it large or small in a real-world sense?
b. Use the regression to predict the average annual growth rate for a country has average valeus for all regressors.
c. Construct a 90% confidence interval for the coefficient on tradeshare. Is the coefficient statistically significant at the 10% level? d.Test whether the political variables rev_coups and assassinations, taken as a group, can be omitted from the regression. What is the p-value of the F-statistic?