Problem set 4

2020-09-24

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 quiz.

See the exercises below, or you can download them as a pdf. You can download the data file you need for question 6 on the website, 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?
    The answer is not entirely obvious, but one must properly interpret the coefficient on alcohol in either case. If we include $attend$, then

we are measuring the effect of alcohol consumption on college GPA, holding attendance fixed. Because attendance is likely to be an important mechanism through which drinking affects performance, we probably do not want to hold it fixed in the analysis. If we do include *attend*, then we interpret the estimate of as measuring those effects on $colGPA$ that are not due to attending class. (For example, we could be measuring the effects that drinking alcohol has on study time.) To get a total effect of alcohol consumption, we would leave attend out.

b. Should SAT and $hsGPA$ be included as explanatory variables? Explain.
We would want to include SAT and $hsGPA$ as controls, as these measure student abilities and motivation. Drinking behavior in college could be correlated with one's performance in high school and on standardized tests. Other factors, such as family background, would also be good controls.

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?
   We could imagine many factors that are correlated with both crime rates and the size of a county's police force, which would introduce omitted variable bias. For example, areas with lower property values may have higher crime rates but also have fewer police (if property values either directly or indirectly predict revenue to fund police departments). Conversely, areas with higher incomes may have more funding for police and lower crime rates.

   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 $\hat{\beta}_1 < \beta_1$?)
   Suppose that higher-income areas have lower crime rates ($crime = \beta_0 + \beta_1 police + \beta_2 income + u$, $\beta_2 < 0$ and that counties with high incomes tend to hire more police ($police = \alpha_0 + \alpha_1 police + u$, $\alpha_1 > 0$. In this case, an equation that omits income would estimate a downward-biased relationship between police and crime rates, so 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

The proposed research in assessing the presence of gender bias in setting wages is too limited. There might be some potentially important determinants of salaries: type of engineer, amount of work experience of the employee, and education level. The gender with the lower wages could reflect the type of engineer among the gender, the amount of work experience of the employee, or the education level of the employee. The research plan could be improved with the collection of additional data as indicated and an appropriate statistical technique for analyzing the data would be a multiple regression in which the dependent variable is wages and the independent variables would include a dummy variable for gender, dummy variables for type of engineer, work experience (time units), and education level (highest grade level completed). The potential importance of the suggested omitted variables makes a "difference in means" test inappropriate for assessing the presence of gender bias in setting wages.

Note, however, that even in the absence of differences in mean wages between male and female engineers after controlling for these factors, there may be additional factors driving *those* differences. For example, if women earn less than men among all engineers, but that difference vanishes after controlling for the type of engineers (if, say, men are more likely to be senior engineers), there may still be bias in determination of those promotions, which we are not set up to detect here.

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.

There are many potential issues! The main one is that incarceration is (thankfully) not random. As one example, people are incarcerated are more likely to face environmental factors like poverty, childhood trauma, neighborhood crime rates. Their labor market outcomes in the absence of

incarceration may have been worse off relative to the never-incarcerated population regardless.

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.
    Regardless of which column you choose, the answer is yes. Suppose you choose column (1). The t-statistic is $t = \frac{5.90}{0.23} = 26.7$, which is statistically significant at any conventional level. You can also calculate a 95% confidence interval: $5.90 \pm 1.96 * 0.23 = 5.90 \pm 0.4508$

    b. Do there appear to be important regional differences in hourly earnings? Use an appropriate hypothesis test to explain your answer.
    $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$
    H_a : at least one equality does not hold (any one coeffcient or more is non-zero)
    To answer this, we can look at the F-statistic for regional effects begin zero at the bottom of column 3. We see that F is 6.24. The 5% critical value from a distribution of $F_{3,\infty} = 2.6$, so we can reject at the 5% level (or any standard level, for that matter.

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

Regressor	(1)	(2)	(3)
College (X_1)	5.90 (0.23)	5.92 (0.23)	5.88 (0.23)
Female (X_2)	− 2.85 (0.22)	− 2.83 (0.22)	− 2.83 (0.22)
Age (X_3)		0.31 (0.06)	0.31 (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 (0.15)	4.75 (1.13)	4.05 (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:
   a. Construct the R^2 for each of the regressions

(a) Using the expressions for R^2 and $\bar{R}^2$, algebra shows that

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2), \text{ so } R^2 = 1 - \frac{n-k-1}{n-1}(1-\bar{R}^2).$$

$$\text{Column 1: } R^2 = 1 - \frac{420-1-1}{420-1}(1-0.049) = 0.051$$

$$\text{Column 2: } R^2 = 1 - \frac{420-2-1}{420-1}(1-0.424) = 0.427$$

$$\text{Column 3: } R^2 = 1 - \frac{420-3-1}{420-1}(1-0.773) = 0.775$$

$$\text{Column 4: } R^2 = 1 - \frac{420-3-1}{420-1}(1-0.626) = 0.629$$

$$\text{Column 5: } R^2 = 1 - \frac{420-4-1}{420-1}(1-0.773) = 0.775$$

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?

(b)

$$H_0 : \beta_3 = \beta_4 = 0$$
$$H_1 : \beta_3 \neq, \beta_4 \neq 0$$

Unrestricted regression (Column 5):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4, \; R^2_{\text{unrestricted}} = 0.775$$

Restricted regression (Column 2):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \; R^2_{\text{restricted}} = 0.427$$

$$F_{HomoskedasticityOnly} = \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}})/q}{(1-R^2_{\text{unrestricted}})/(n-k_{\text{unrestricted}}-1)}, \; n = 420, \; k_{\text{unrestricted}} = 4, \; q = 2$$
$$= \frac{(0.775-0.427)/2}{(1-0.775)/(420-4-1)} = \frac{0.348/2}{(0.225)/415} = \frac{0.174}{0.00054} = 322.22$$

5% Critical value form $F_{2,00} = 4.61$; $F_{HomoskedasticityOnly} > F_{2,00}$ so H_o is rejected at the 5% level.

c. Construct a 99% confidence interval for β_1 for the regression in column (5)

$-1.01 \pm 2.58 * 0.27$

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.

6. Download the dataset growth.dta from the website, 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 values 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?

See do-file