

17.871
Spring 2002
Problem Set # 2: Descriptive Statistics and Bivariate Regressions

Handed out: February 21, 2002
Due back: February 28, 2002

Part A.

Do the following review exercises in Freedman, et al., **3rd edition**. When an exercise calls on you to explain something, do so legibly, using complete sentences, paragraphs, etc. Please type if the explanation takes more than a sentence.

Chapter 3 review exercises (pp. 50-55), # 1, 2, 4, 8, 9, 10
Chapter 4 review exercises (pp. 74-76) # 1, 2, 3, 5, 6, 7, 10
Chapter 8 review exercises (pp. 134-139) # 1, 2, 3, 6, 7, 11

Part B.

The data on the page 3 is a random sample taken from responses to the Spring 1995 Subject Evaluation Guide at MIT.

1. What is the correlation coefficient between the overall rating of the subject and the subject's enrollment?
2. What is the correlation coefficient between the overall rating of the subject and the *log* of the subject's enrollment?
3. Discuss which correlation is the better summary of the relationship between subject size and class quality.

For this Part, show the calculations that produce your results for B-1 and B-2.

Part C.

Create a STATA data set of the data you hand-analyzed in Part B.

1. What does STATA calculate as the correlation between overall rating of the subject and the subject's enrollment (to 3 decimal places)?
2. What does STATA calculate as the correlation between the overall rating of the subject and the *log* of the subject's enrollment (to 3 decimal places)?
3. Generate two graphs:
 - a. Overall subject rating against enrollment
 - b. Overall subject rating against the *log* of enrollment.Does this graph cause you to change your answer to B-3? Why or why not? (Hint 1: When I say "generate a graph of y against x," y is the dependent variable and x is the

independent variable. Hint 2: In drawing graph b. using the ",xlog" option to the graph command instead of generating the log of enrollment and plotting against that.)

For this part, turn in a log file that shows (1) the variable names in the data set, (2) the averages of the variables, and (3) the correlations among the relevant variables. Also, turn in printed versions of the graphs in C-3.

Part D.

There are two data sets in /mit/course_number/Examples:

fla_precinct_subset.dta: Contains basic election data, at the precinct level, from the 2000 Florida presidential election.

fla_county_subject.dta: Contains basic data about the voting technology used in each county in the 2000 Florida presidential election.

There is a description of the variables in each data set on page 4.

Using these two data sets, show the answers to the following questions:

1. What was the correlation coefficient between % of the registered voters in a precinct who were Black and the percentage of ballots in each precinct that registered either an "overvote" or an "undervote." Does the correlation coefficient you calculated fairly describe this relationship?
2. Part of the policy debate currently raging about voting reform is whether to require optical scanning voting machines to scan ballots in the precincts (in the presence of voters) or in the central office (at the end of Election Day). What is the average overvote+undervote rate in optical scan counties that use central scanning *versus* counties that use precinct scanning?

For this Part, turn in the STATA "do-file" and output that illustrate the answers to these questions. Write two very short paragraphs that use your results from STATA to answer these questions.

Here are some hints:

1. You're going to need to use the "merge" command.
2. An elegant way to find the average number of overvotes per county broken down by the type of voting technology would be to type:

```
table technology, c(mean overvote). If you want to weight each precinct's contribution to the mean by the total number of ballots cast type  
table technology [fweight=total_ball], c(mean overvote). To find out more about the "table" command and using weights, either confer with Hamilton or use the STATA "help" command.
```

Estimate the amount of time it took to do this problem st: ____ hours.

Subject Evaluation Data

The following are (real) data, randomly-sampled, from the Spring 1995 Subject Evaluation Guide. Students were asked to rate, on a 7-point scale (1=poor, 7=excellent) the overall quality of the subject. Subject enrollment is what was reported by the Registrar.

ID #	Rating	Enrollment
1	6	56
2	5.9	52
3	4.7	18
4	5.4	45
5	6.5	13
6	3.8	501
7	5.8	10
8	6	50
9	6	17
10	5.5	25
11	4.9	22
12	5.2	29
13	6	39
14	5.3	14
15	6.3	22
16	6.6	15
17	5.3	116
18	6.4	13
19	5.4	30
20	4.2	65
21	6.1	20
22	5.5	28
23	6.2	22
24	5	7
25	5.7	25
26	6.6	19

Variables in fla_county_subset.dta

variable name	description	coding	
county	county name	Natural county name	
technology	type of voting technology used.	Datavote	a type of punch card ballot (non-pre-scored cards)
		Hand	hand-counted paper ballots
		Lever	mechanical lever machines
		Optical	optical scan
		Votomatic	a type of punch card ballot (pre-scored cards)
centraltab	are optically scanned ballots scanned in the precinct or in the central office?	1	Precinct
		2	Central
		9	Not applicable

Variables in fla_precinct_subset.dta

variable name	description
county	county name
precinct	precinct name/number
total_ball	total number of ballots cast in the precinct
undervote	total number of ballots cast in the precinct that recorded no vote for president (note: I don't know why Broward County is listed as having - 1 undervotes.)
overvote	total number of ballots cast in the precinct that recorded more than one vote for president
blackrv	percentage of registered voters in the precinct who are African American