Stata Recitation - Week 9 - Correlation and Regression

McCourt School of Public Policy, Georgetown University

MAIN IDEAS:

- Corr and pwcorr
- Basic regression and location of relevant output
- Understand the role of missing data in analysis sample

Correlation

Quantifies the way two variables move together. Can be calculated for any numeric variables, but, like other summary statistics, interpretation is not always clear for categorical variables.

Example: Wage and Tenure

- Wage tends to increase with tenure

```
twoway (scatter wage tenure) (lfit wage tenure)
```

- Correlation is positive:

```
correlate wage tenure
```

- Average levels of education are lower for older people in this sample:

```
correlate age grade
```

- Correlate help page

```
help correlate
```

- Correlate can be abbreviated to cor
- Correlate takes a varlist

```
corr age grade wage tenure union
```

- When using correlate with multiple variables, what happens to observations with missing data?
- Compare sample size of previous command with observation counts from:

```
sum age grade wage tenure union
```

- If we don't want to use a single sample for entire correlation matrix, we can use pwcorr
- pwcorr is an alternative version of corr

```
help corr
pwcorr age grade wage tenure union
```

- But, we want to see the number of observations for each comparison: Look at option list

```
pwcorr age grade wage tenure union, obs
```

In Class Activity 1

Correlate Practice Questions:

Use highschool and beyond data with this command:

```
use http://www.ats.ucla.edu/stat/stata/notes/hsb2
```

Write stata command and interpret the result in a do-file.

1. Create a scatter plot with a linear fitted line to examine the relationship between write and read.
2. What is the correlation between write and read? Is it significant? Report the observation number in this case.
3. Generate a new variable called `id_odd` which is equal to 1 if the student id number is odd numbers and 0 if it is even numbers. (Hint: try Google `stata modulus`)
4. Count the number of `id_odd` equals to 1. Replace the `write` to missing value if `id_odd` equals to 1.
5. What is the correlation between write and read now? Is it significant? Report the observation number in this case. Compare result with question 2.
6. Count the number of female students if their student id is odd. Replace the `read` to missing value if the respondent is female with an odd student id.
7. What is the correlation between write and read now? Is it significant? Report the observation number in this case. Compare result with question 2 and 5.

Answers:

```
*1
clear
use http://www.ats.ucla.edu/stat/stata/notes/hsb2
twoway (scatter write read) (lfit write read)

*2
pwcorr write read, obs sig

*3
gen id_odd=mod(id,2)

*4
count if id_odd ==1
replace write=. if id_odd==1

*5
pwcorr write read, obs sig

*6
count if id_odd==1 & female==1
replace read=. if id_odd==1 & female==1

*7
pwcorr write read, obs sig
```

Regression

We will cover how to run regressions and where to find output components. Some output components have not been covered in class, and won't be covered until Quant II. You don't have to know what the output means yet.

Regression example:

```
clear
sysuse auto

twoway (scatter mpg weight) (lfit mpg weight)
```

Single-variable regression

```
regress mpg weight
```

- Review components of regression output
- Many of these components won't be covered until Quant 2.
- Overall output: Number of obs, R-squared, Sum of squares
- Variable specific output: Coefficients, standard errors, p-values
- Note: P-values are not actually zero, they just may have too many leading zeros to display.

Regression on a restricted sample: domestic cars only

```
regress mpg weight if foreign==0
```

- Note number of observations

Multiple-variable regression

```
regress mpg weight length
```

Regression with missing values

- So far the variables that we've used have no missing values:

```
summarize mpg weight length
```

- What if mpg has some missing values?
- Go into data editor and make some observations equal to .
- Copy the resulting commands into your do-file:

```
replace mpg = . in 9
replace mpg = . in 18
replace mpg = . in 30
replace mpg = . in 50
replace mpg = . in 62
```

- Run regression on modified data

```
regress mpg weight length
```

- What happens to number of observations? Why?
- What if other variables also have missing data in different observations?

```
replace weight = . in 7
replace weight = . in 22
replace weight = . in 27
replace weight = . in 41
replace weight = . in 60
```

- Run regression on modified data

```
regress mpg weight length
```

- What happens to number of observations? Why?
- What if other variables also have missing data in the same observations?
- Change the length variable to missing for the same observations that are missing mpg.

```
replace length = . if mpg==.
```

- Run regression on modified data

```
regress mpg weight length
```

- What happens to number of observations? Why?

In Class Activity 2

Use high school and beyond data with this command:

```
use http://www.ats.ucla.edu/stat/stata/notes/hsb2
```

Write stata command and interpret the result in a do-file.

1. Regress the writing score (dependent variable) on gender. What does the relationship seem to be?
2. Regress the writing score (dependent variable) on gender, type of school, type of program, and student id. What does the relationship seem to be?
3. Try regression with using if statement
4. Try Regression with missing data

Answers

```
regress write gender
regress write female schtyp prog id
```