Stata Recitation - Week 9 - Correlation and Regression

McCourt School of Public Policy, Georgetown University

MAIN IDEAS:

- Run basic regression and locate relevant output
- Understand the concept of estimation and postestimation commands
- Find help on postestimation commands for any given estimation command
- Run basic regression postestimation commands

RESOURCES

PDF Manuals [U] User's Guide, Section 20: Estimation and postestimation commands.

Warm Up: Use `IPEDSpull_recitation.dta`, which can be found on the Blackboard site or attached

1. Create a variable, percent_women, that gives the percentage of full-time students that are women.
2. Determine the correlation of percent_women with the graduation rate.
3. Using pwcorr, find the correlation between retention rate, graduation rate, transfer rate, and percent of first-time full-time students with financial aid. Give the number of observations for each correlation.
4. Tabulate the number of colleges in each region by size category. What percentage of colleges in the Far West are "Very Large"?
5. Tabulate the number of colleges in each region by whether the college is a land-grant college. Does it seem likely that the proportion of land-grant colleges is similar in each region?

Ans:

```
use "C:\Users\gppilab\Desktop\IPEDSdata_recitation.dta", clear
gen percent_women = ftwomen / allfttotal
pwcorr percent_women gradrate
pwcorr ftretention gradrate transfer percentftftanyaid, obs sig
tab  region sizecat, row
tab region landgrant, chi2
```

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 takes a varlist

```
help correlate
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

```
help pwcorr
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
```

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.

We should avoid interpretation of regression output now.

Regression example:

```
clear
sysuse auto

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

Single-variable regression

```
regress mpg weight
```

- Components of regression output
- P-values are not actually zero
- (They just have too many leading zeros to display)
- Predicted MPG for a car weighing 2020 lbs based on estimated model:

  ```
  display   39.44028-0.006008*2020
  ```

- Compared to actual MPG for the car with this weight:

  ```
  list make mpg weight if weight == 2020
  ```

- Difference between actual and predicted is the residual:

  ```
  display 35-27.30412
  ```

  ### Multiple-variable regression

  ```
  regress mpg weight length
  ```

- Note the changes in degrees of freedom and R-squared

Regression on a restricted sample: domestic cars only

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

- Note number of observations

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?

```
replace length = . in 9
replace length = . in 18
replace length = . in 27
replace length = . in 41
replace length = . in 60
```

- Run regression on modified data

```
regress mpg weight length
```

- What happens to number of observations? Why?

Postestimation

After any estimation command, such as regress or mean, the results are stored in Stata's memory until a new estimation command is run. The currently stored results can be re-displayed by retyping the estimation command without any arguments. For example, type regress in command prompt.

Stored results can be used for additional calculations, such as: - calculating predicted values - creating diagnostic plots - running additional statistical tests

```
* Reload original data
clear
sysuse auto
* Run basic regression
regress mpg weight length
```

Example: Predicted values

- Look at the results from the previous regression:

```
regress
```

- What is the predicted mpg for a car weighing 2,000 lbs that is 200 inches long?
- One way is to use the calculated coefficients manually:

```
display (-.0038515)*2000 + (-.0795935)*200 + (47.88487)
```

- There is a short-cut to accessing these coefficients:

```
display _b[weight]*2000 + _b[length]*200 + _b[_cons]
```

- For more information, see:

```
help _variables
```

- What if we wanted to get predicted values for every observation?

```
gen pred_mpg1 = _b[weight]*weight + _b[length]*length + _b[_cons]*1
```

- Show output:

```
browse mpg pred_mpg1
```

- We can do this even easier with a post estimation command:
- First go to regress help page, then navigate to regress postestimation

```
help regress
```

- see predict option

```
predict pred_mpg2
```

- New variables are the same:

```
sum pred_mpg1 pred_mpg2
br  pred_mpg1 pred_mpg2
```

Example: Diagnostic Plots

- Back to regress postestimation help page

```
help regress postestimation
```

- Many plots are available after regression.
- When to use these plots and how to interpret them will be covered in class.
- Here, we will discuss how plots are created and what information they contain.

`rvpplot`

- Plots regression residual against an X-variable of your choice.
- Used to estimate the plausibility of regression assumptions.

```
rvpplot weight
```

- Let's construct the same graph using predict and twoway scatter.

```
predict mpg_resid , residuals
twoway scatter mpg_resid weight , name(Graph2)
```

- Note: Using the name option allows us to have multiple graph windows open simultaneously
- For more information see: help graph name
- drop graphs from memory when finished:

```
graph drop _all
```

`rvfplot`

- Plot regression residuals against fitted or predicted values.

```
rvfplot
```

- Again, we can construct the same graph from components

```
twoway scatter mpg_resid pred_mpg1, name(Graph2)
```

- drop graphs from memory when finished:

```
graph drop _all
```

Example: Statistical Tests

- The t-statistic and p-value reported in regression results are results from statistical tests about the estimated coefficient. Specifically, the null hypothesis being tested is that the true value of the coefficient is equal to zero. We can get the same results from a related test using a post-estimation command.

`test weight=0`

- Note only a single equals sign is required for test commands. You haven't covered this type of test yet, but note that the p-value is very close to the p-value reported above.

Optional example: Statistical Tests

- Another estimation command estimates population means from sample data
- Note: this is different from the sample mean calculated by the summarize command.

```
clear
sysuse bpwide
```

- Estimation command: mean

```
mean bp_before bp_after
```

- Test for equality of these two population means:

```
test bp_before=bp_after
```

- Again, F-test has not been covered yet, but look at the p-value:
- Compare p-value with that from a two-tailed t-test:

```
ttest bp_before=bp_after
```

Practice Problems: Use IPEDSpull_recitation.dta, which can be found on the Blackboard site or attached.

```
use "C:\Users\gppilab\Desktop\IPEDSdata_recitation.dta", clear
```

1. Regress the graduation rate (dependent variable) on the the number of full-time students. What does the relationship seem to be?
2. Predict the graduation rate for a community college with 750 full-time students.
3. What is the actual graduation rate and name of the college with 750 full-time students?
4. Regress the graduation rate (dependent variable) on the number of college employees, the percent of students with aid, and the full time retention rate.
5. Create a variable that is the predicted graduation rate from the regression in #4. Look at summary stats to compare the mean actual and predicted grad rates.

Ans:

```
regress gradrate allfttotal

predict gradpred if allfttotal ==750

list gradrate if allfttotal == 750

regress gradrate ftretention totemp percentftftanyaid

predict gradpred2
sum gradpred2 gradrate
```