Stata Recitation - Week 8 - Chi-squared tests and correlation

McCourt School of Public Policy, Georgetown University

Key Ideas:

- Calculate statistics based on two variables: tab2 (chi2 option), corr, pwcorr
- Understand how missing data is treated for commands on multiple variables

Discrete/Categorical variables : Two-way tabulation and Pearson's chi-2 test

Two way tabulation

Produce counts of number of observations in each cell of a two-way table.

Examples

```
clear
sysuse nlsw88.dta
tab2 race married
```

- First variable goes in rows, second variable goes in columns.
- This is important when you have a variable with many categories:

  ```
  tab2 age married
  tab2 married age
  ```

- Many different options with tab:

  ```
  help tab2
  ```

- We've seen the missing option with one-way tabulations:

  ```
  tab2 union married
  tab2 union married , m
  ```

- Look at total number of observations for these two tables.
- Compare with Obs number from summarize:

  ```
  sum married union
  ```

Important

- We are moving into commands that take data from multiple variables
- If an observation has missing data for any of the variables, that observation is dropped from the calculation.
- With some commands, like `tab2`, we can avoid that behavior. But that will not be possible for other commands.

More options for tab2

- column : Gives percentage breakdown of row category within each column.

```
tab2 race union, column
tab2 race if union==0
tab2 race if union==1
```

- row : Gives percentage breakdown of column category within each row.

```
tab2 race union, row
tab2 union if race==1
tab2 union if race==2
tab2 union if race==3
```

- cell : Gives percentage of observations in each cell.

```
tab2 race union, cell
```

- expected : Gives the expected number of observations in each cell based on marginal distributions of each variable

```
tab2 race union, row column
```

- expected number of observations in the white, nonunion cell:

```
display 0.7204*0.7545*1878
tab2 race union, expected
```

Report Chi2 test statistic:

- Test for independence of two categorical variables:
- This will be covered in class, but you should know how to calculate and find test statistic.

```
tab2 race union, chi2
```

- test statistic: Pearson chi2(2) = 13.0814

- P-value: Pr = 0.001
- Components of chi2 test statistic can be reported using option: `cchi2`
- Students should read through these options after learning about the Chi-2 test in class.

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
```