Chapter 4: Basic Statistical Routines

Instructor: Sergi Quintana
Notes from: Manuel V. Montesinos

Statistics Brush-Up Course

Fall 2022

Descriptive Statistics and Percentiles

We can get descriptive statistics including the **mean** and the **standard deviation** of the sample using the `summarize` command.

Let's get the descriptive statistics first for the 0/1 (dummy) variable *female*, and next for a continuous variable *earnings* in a detailed way.

The `mean` command produces estimates of **means** along with *standard errors*.

If you need to calculate **summary statistics conditioned by other variables**, try `table` or `tabstat`.

Correlation

The **correlation coefficient** measures the **strength** and **direction** of the linear relationship between two variables.

The `correlate` command displays the **correlation matrix**. If there is no list of variables specified, Stata computes it for all the variables. It does a listwise deletion of missing data by default (if an observation has a missing value for any of the variables listed in the command, that observation is not taken into account).

You can also use the command `pwcorr` (pair-wise correlation), which displays all **pairwise correlation coefficients**. It does a pairwise deletion of missing data (a pair of data points is deleted from the calculation of the correlation only if one –or both– of the data points in that pair is missing).

t-Test on Mean Difference

Hypothesis testing is a **method of statistical inference** used in determining whether a certain hypothesis should be rejected or not.

The `ttest` command performs **t-tests** for one sample, two samples and paired observations.

Stata calculates the t-statistic and its **p-value** under the assumption that the sample comes from an approximately normal distribution.

If the p-value of the test is **small** (0.05 is often used as threshold), there is evidence that the mean is different from the hypothesized value. Otherwise ($p > 0.05$), the null hypothesis cannot be rejected.

t-Test on Mean Difference

Single sample t-test. Independent group t-test

A **single sample t-test** assesses the null hypothesis that the population mean is equal to the pre-selected value:

```
ttest earnings==25
```

An **independent group t-test** is designed to compare means of the same variable between two groups. In other words, it tests whether the difference of the means is 0.

This test assumes that the variances of the two populations are the same. To allow for unequal variances, add the `unequal` option:

```
ttest earnings, by(female) [unequal]
```

Paired t-test. Cross tabulation and corresponding tests

A **paired t-test** is used to test whether the means of two variables are statistically different from each other:

```
ttest sm==sf
```

To create a **cross table of two variables**, the `tabulate` command is the alternative way to compute a cross table of the specified variables. For example:

```
tabulate female divorce
tabulate female divorce, cell | row | column
```

To determine whether two variables are **independent**, use the **Pearson's chi-squared test**. The null hypothesis is that of independence: `tabulate female divorce, chi2`

OLS Regression

Suppose that you want to study the impact of an extra year of schooling on wages (earnings).

You can perform **regression analysis** using the `regress` command.

Stata stores the values of **coefficient estimates** in `_b[]` (or in `_coeff[]`) and **standard errors** in `_se[]` until you run another regression. You can state the coefficients you want to retrieve inside brackets:

```
display _b[_cons] & display _b[s]
display _se[_cons] & display _se[s]
```

OLS Regression

Display/save selected regression results

You can save the estimated values using the `scalar` command:

```
scalar b0 = _b[_cons]
scalar list _all
```

Stata saves selected results temporarily in the `e()` function. These results change every time a new `regress` command is used:

```
ereturn list
```

To display (but not save) current results, call the `e()` function:

```
display e(N) | e(mss) | e(df_m) | e(rss) |
e(df_r) | e(r2) | e(F)
```

OLS Regression

Display/save selected regression results

You can display the OLS **coefficient vector** and the **variance-covariance matrix** of the coefficient estimates from the most recent regress command:

```
matrix list e(b)
matrix list e(V)
```

To save the OLS coefficient vector and variance-covariance matrix of the OLS coefficients, give them names:

```
matrix beta = e(b)
matrix VarCov = e(V)
```

Use the `predict` command to calculate the **predicted values** of the dependent variable given by the last regression equation and name them. For example: `predict yhat`.

To calculate the OLS **residuals**, use the `predict` command with the `residuals` option and name them. For example: `predict uhat, residuals`.

Exercise

Using the *womenwage.dta* dataset (Stata's online resources):

1. Test whether the average wage is different in rural and urban areas.
2. Run a regression of log wage on schooling, age, age squared and `nev_married`.
3. Save the coefficient and standard error of school as scalars.
4. Interpret the coefficient of schooling.
5. Test the hypothesis that schooling has no impact on wages with (1) the t-statistic, (2) the p-value, and (3) the 95% confidence interval.
6. Display the number of observations.
7. Compute the R^2 from the total sum of squares and the sum of squared residuals.