Stata Workshop: Statistical Analysis Guide

Carlos Zambrana

February 25, 2025

Contents

1 Introduction

This guide provides a step-by-step explanation of the Stata commands used in the workshop.

2 Setting Up the Environment

2.1 Stata Version

```
version 15
```

You don't need to run this command today, but it ensured that Stata runs in compatibility mode for version 15 while I was preparing this material.

2.2 Setting the Working Directory

```
cd "C:\Users\Carlos Zambrana\OD\Work\KDSC_Stata_Workshops"
```

You should change this to your own directory to save exported files later.

3 Loading and Exploring a Dataset

3.1 Viewing Available Example Datasets

Stata comes with built-in example datasets. The WB data is very good but I could not find datasets with categorical variables (I didn't look very hard). So, we'll be using some of the example datasets you will see in some of Stata's help pages.

```
help dta_examples
```

3.2 Loading the Auto Dataset

```
sysuse auto, clear
```

Clears the memory and loads the auto.dta dataset.

3.3 Describing the Dataset

```
describe
```

Provides metadata about the dataset.

4 Summary Statistics

4.1 Basic Summary Statistics

```
summarize *
```

Displays summary statistics for all **numeric** variables.

4.2 Detailed Summary for a Specific Variable

```
summarize price, det
```

Provides detailed summary statistics for price.

5 Tabulations

5.1 One-way Frequency Tables

```
tab foreign
```

Generates a frequency table for foreign.

5.2 Two-way Frequency Tables

```
tab rep78 foreign
```

Generates a frequency table for rep78 and foreign.

5.3 Chi-Square Test for Independence

```
tab rep78 foreign, chi2
```

Tests the association between repair records and car origin.

5.4 Using tab to create dummy variables

```
tab rep78, gen(rep)
```

Creates dummy variables for each value of rep78. Useful for creating dummies quickly.

5.5 Apart from tabulate you can use table for more complicated analyses

```
table rep78 foreign, c(mean price median weight max turn min mpg)
```

Gives the requested summary statistics for each value of rep78 and foreign.

5.6 We can add a third variable to the tabulation

```
table rep78 foreign hiturn, c(mean price median mpg) f(%9.0f)
```

Gives the requested summary statistics for each value of rep78, foreign and hiturn, and shows you how to format the resulting statistics.

6 Regression Analysis

6.1 Basic Regression

```
reg price foreign
```

Runs a simple regression of price on car origin.

```
reg price foreign mpg, vce(robust)
```

Adds mpg as a covariate with robust standard errors.

6.2 Save estimates

```
reg price foreign mpg, vce(robust)
est save spec1

reg price i.foreign mpg i.turn, vce(robust)
est save spec2

est restore spec1
est replay
```

We can use `estimates save` to save a regression's estimates so we can use them later without having to run the regression again. We can also use `estimates replay` to replay the estimates.

6.3 Test hypotheses about coefficients

```
reg price mpg i.turn
test (_b[mpg]=200)
```

This tests whether the coefficient for mpg is equal to 200.

```
test (_b[mpg]=_b[33.turn])
```

This tests whether the coefficient for mpg is equal to the coefficient of the dummy for turn=33.

We can also use `suest` (seemingly unrelated estimation) to test hypotheses about coefficients from different regressions. We will run two regressions, one for domestic cars and another for foreign cars. When we do this, the omitted category for a variable may change because it could be missing on one of the subsets so, first, we need to fix the omitted category of turn to something that shows up in both regressions. In this case, it's 33.

```
fvset base 33 turn
```

Now we run the regressions, then use `suest` to combine the estimation results, and then we test whether the coefficient for mpg is the same in both.

```
reg price mpg i.turn if foreign==0
est store domestic
reg price mpg i.turn if foreign==1
est store foreign

suest domestic foreign

test ([domestic_mean = foreign_mean ]: mpg)
```

6.4 Generating Predictions

```
predict yhat, xb
```

Creates predicted values for the dependent variable based on the estimated regression coefficients.

```
predict rhat, residuals
```

Generates residuals (the difference between observed and predicted values).

6.5 Predictive Margins

```
reg price c.mpg##i.foreign, vce(robust)
margins, dydx(mpg)
```

Calculates predictive marginal effects of mpg on price. That is, it is the predicted change in price given a one-unit change in mpg, evaluated at the mean of mpg. We can also ask for the estimated effect at different values of mpg:

```
margins, dydx(mpg) at(mpg=(10(10)40)) plot
```

6.6 Contrasts and Margins Analysis

```
fvset base 33 turn
```

Sets category 33 as the reference level for turn.

```
reg price c.mpg##i.turn, vce(robust)
margins turn
```

This simply shows the predicted price for each value of turn

```
reg price c.mpg##i.turn, vce(robust)
margins r.turn
margins a.turn
margins ar.turn
```

Performs contrasts, analyzing differences from the reference category, then with respect to the next category, and then with respect to the previous category.

6.7 Exporting Results: outreg2

```
ssc install outreg2
reg price mpg, vce(robust)
outreg2 using myfile.xls, excel replace ctitle("Baseline")
```

Exports regression results to an Excel file.

```
reg price mpg i.rep78 trunk weight length , vce(robust)
outreg2 using myfile.xls, excel ctitle("Added more vars") append
```

Appends results to the same Excel file instead of replacing it.

`outreg2` has many options for deciding how to output your results. Below is a slightly more complicated case: we drop `foreign` from the list of variables we want, set the number of decimals to 3, and we asked for t-statistics below the coefficients, instead of standard errors.

```
outreg2 using myfile.xls, label ///
ctitle("A more complicated example") excel drop(i.foreign) dec(3) ///
title("Outreg2 example") stats(coef tstat) append
```

7 Panel Data

7.1 Setting Up Panel Data

```
webuse nlswork, clear
xtset idcode year
```

Loads and sets up a panel dataset. The first argument to `xtset` is the individual ID (the variable that uniquely identifies each unit), and the second is the time variable.

7.2 Fixed-Effects Regression

```
xtreg ln_wage union age i.race collgrad i.occ_code ttl_exp tenure i.year, fe vce(robust)
```

Runs a fixed-effects regression.

8 Instrumental Variables

8.1 Creating Simulated Data

```
clear
set obs 1000
gen z = rnormal(5,3)
gen conf = rnormal(12,3)
gen x1 = 2 + 3z + 5conf + rnormal()
gen x2 = rnormal(2,4)
gen y = 10 + 3x1 + 5x2 + 2*conf + rnormal()
```

Generates synthetic data where `z` is an instrument for `x1`, `conf` represents unobserved confounders, and `x2` is some other variable. Note that `z` affects `y` only through `x1`, and that both `x1` and `y` are functions of `conf`.

```
ivregress 2sls y x2 (x1 = z)
```

Performs a two-stage least squares regression where x1 is instrumented by z. Note that in parenthesis you have to specify the endogenous variable before the equals sign, and the exogenous variables to the right of the equals sign.