L2. Initial Looks at Data

The goal of this seminar is to learn how to use STATA by engaging in an actual research project. Let’s test Blau and Duncan’s occupational model. Before Blau and Duncan’s stratification process was put together, it was generally thought amongst social scientists that the roles of education and parents’ occupation were mutually exclusive. However, the central discovery of The American Occupational Structure proved these two elements – education and parents’ occupation – are heavily interrelated, as education is a channel through which parental occupation relates to occupational attainment (Blau and Duncan 1967). Furthermore, the Blau and Duncan model has been established as an empirical generalization that is robust in time and across countries ( Hout and DiPrete 2006).

You do not have to understand this immediately. We will go in depth further as we progress through the project. For now, we want to find in ESS the variables that are in the study. We need 3 variables. 1) Ideally, occupation would be our dependent variable. However, for now, we are going to look at household income. 2) Education is going to be our mediating variable. 3) Parent’s occupation is going to be our key independent variable. We can also add in a couple variables that will act as controls, let’s say age and gender.

In ESS we can find:

- occf14b 🡪 Father’s occupation when the respondent was 14.

- eduyrs 🡪 Years of full-time education completed.

- hinctnta 🡪 Household income (For now, let’s look at income. Later we’ll go for occupation).

- gndr 🡪 Gender

- agea 🡪 Age of respondent, calculated.

L2.1 Basic inspection of data

There are many options to start describing the data in STATA. The main point of these commands is to gain an understanding of unknown data. Let’s go through them:

Browse and editare useful commands to browse through the entire dataset or a specific variable. If we use the editcommand, we can browse and change the values that we want:

 browse [varlist] [if] [in] [, nolabel]

 edit [varlist] [if] [in] [, nolabel]

(I do not recommend editing data this way as it is difficult to document)

Examples:

 browse

 browse occf14b

Describe produces a summary of all the variables in the dataset in memory or of the data stored in a Stata-format dataset. If we introduce a variable after

Inspect. The inspect command provides a quick summary of the number of negative, zero, and positive values; the number of integers and nonintegers; the number of unique values; and the number of missing; and it produces a small histogram. Its purpose is not analytical but is to allow you to quickly gain familiarity with unknown data.

Codebook. The command codebook examines the variable names, labels, and data to produce a “codebook” describing the dataset. If we add a variable after the command, we get a codebook entry for that specific variable.

Syntax for codebook:

codebook [varlist] [if] [in] [, options]

Examples:

codebook

codebook occf14b

codebook occf14b, compact

Try inspecting ESS data using the commands above. Try to figure out what type are the variables we have singled out. Are they categorical? Ordinal? Dichotomous? Continuous?

L2.2 Descriptive statistics.

The next step in our project is to begin to dive in the analytical side of STATA. Once we’ve gained an understanding of the nature of our data, our aim is to look at the distribution of our variables of interest and how they relate to each other.

One of the most useful commands in STATA is summarize:

The command summarize calculates and displays a variety of univariate summary statistics. If no varlist is specified, summary statistics are calculated for all the variables in the dataset. From this simple summary we can learn quite a bit about the data.

Syntax for summarize:

summarize [varlist] [if] [in] [weight] [, options]

Examples:

 summarize

 sum agea, detail

 sum agea gndr occf14b

Let’s take a look at the output of the summarize command.



Would you say this is an old sample?

One of the most useful features of this command is the detail option. Let’s take a look at the output:



**Tabulate:**

The command tabulate or (tab for short) produces one-way tables of frequencies.

Note that you can only specify one variable to run a one-way frequency on with the tabulate option. If you want to specify multiple variables for a one-way frequency, use the tab1 command. You can also create two-way tables using the tab command.

Syntax for tabulate:

tabulate varname [if] [in] [weight] [, tabulate\_options]

Example:

 tabulate gndr

 tabulate occf14b gndr

 tabulate occf14b agea

What happens when we introduce agea? We will delve deeper into this when we look at bivariate relationships.

Tab1:

The tab1 command is a convenience tool for producing one-way tables for each of multiple variables specified.

Syntax for tab1:

tab1 varlist [if] [in] [weight] [, tab1\_options]

Example:

 tab1 occf14b gndr agea

Tabulate and Tab1 have some important options: let’s use the help function to figure out what options we have to introduce.

L2.2.1 Variable description with graphics

**Histograms:**

Histograms plot the frequencies of groups or intervals of the continuous variable. Grouping the continuous variable generally requires that you choose an origin and the width of the intervals. When dealing with variables with just a few categories, you do not need to make this choice. Instead, you can just plot the frequency of each category with the histogram command and the discrete option. Histograms are especially useful in experimental scenarios where you can compare treatment and control groups. From the histogram we can also infer whether the distribution of our variables is not normal, if it is skewed, kurtosis is high, or whether there is a lot of dispersion.

bin(#) and width(#) are alternatives specifying how the data are aggregated into bins; bin() by specifying the number of bins (from which the width can be derived) and width() by specifying the bin width (from which the number of bins can be derived).

discrete specifies that varname is discrete and that you want each unique value of varname to have its own bin (bar of histogram).

Syntax

histogram varname [if] [in] [weight] [, [continuous\_opts | discrete\_opts] options]

Examples:

histogram eduyrs

histogram eduyrs, bin(5)
histogram eduyrs, width(10)

histogram occf14b, discrete

histogram eduyrs,normal

hist eduyrs, normal

(bin=46, start=0, width=1.173913)



In this example we can see the variable years of education is normally distributed. Skewness does not seem to be a problem, hence; we can move on with the following tests to see whether the variable needs any transformation.

**Boxplots**

Syntax

graph box varname [if] [in] [weight] [, options]

Examples:

graph box eduyrs

Boxplots can be seen as graphical representations of the “five number summary” with some enhancements. Boxplots in general are composed of a “box”, two “whiskers”, two “fences” and some marker symbols. The box is the rectangle in the middle of the graph.

The lower border of the box is the first quartile; the upper border is the third quartile. The line in the middle of the box is the median, but sometimes this falls on a quartile, so you can put in a marker (like in the example below). The height of the box therefore shows the quartile range.

The whiskers are the two vertical lines below and above the box, which are terminated by thin horizontal lines called the fences. The upper fence is the highest value of the distribution that is smaller than or equal to the third quartile plus 1.5 times the interquartile range. The lower fence is the lowest value of the distribution that is greater than or equal to the first quartile minus 1.5 times the interquartile range.

Observations below the lower fence or above the upper fence are regarded potentially influential data points and are plotted with single plot symbols.



**References**

Blau, Peter M. and Otis Dudley Duncan. 1967. *The American Occupational Structure*. New York: Free Press.

Hout, Michael, and Thomas A. DiPrete. “What We Have Learned: RC28’s Contributions to Knowledge about Social Stratification.” *Research in Social Stratification and Mobility* 24, no. 1 (st 2006): 1–20. <https://doi.org/10.1016/j.rssm.2005.10.001>.