An introduction to Stata in 2 pages. Commands that you actually type into Stata are represented in monospace font. x and y refer to variables in your data. The treatment here is intended to be extremely brief, in order to create a kind of “cheat sheet” that can be presented in 2 pages. More documentation on any command is available in the printed or PDF Stata manuals, or by typing help command.

1 Get Acquainted With Your Data

lookfor allows you to find variables that contain a specified keyword. This is especially useful in large data sets with many variables. Often abbreviated keywords are the most helpful. e.g. to find a poverty variable, type lookfor pov.

describe tells you about the contents of a specific variable. E.g. describe x y. describe, short will tell you very basic things about your data, including the number of observations in the data set, and the size of your data file.

2 Process Your Data

recode x (oldvalue = newvalue), generate(z) will recode a variable into a new variable, often a good idea.

recode _all (-99/-1 = .) will recode all negative numbers from -99 to -1 to missing for all variables in your data. recode x (7/9 = .) changes 7 through 9 to be missing for x. Indeed, recode will change specific values in your data to anything you want, not just missing values.

It is often convenient to rename your variables so that the variables have more intuitively understandable names e.g. rename x depression.

You can create new variables out of old variables using generate newvar = expression e.g. generate newvar = oldvar1 + oldvar2.[alpha]

[alpha] alpha oldvar1 oldvar2 will calculate Cronbach’s alpha from this scale.

It is sometimes useful to sort your data. sort x will sort your data by the values of x.
3 Descriptive Statistics

summarize gives you basic descriptive statistics for a variable, such as the mean (average). Especially useful for continuous variables. E.g. summarize x y or summarize x y, detail.

     tabulate gives you a frequency distribution for your variable. Especially useful for categorical variables. e.g. tabulate x.

4 Bivariate Statistics\footnote{oneway continuous.var categorical.var, tabulate gives you an oneway ANOVA of a continuous variable over a categorical factor.}

Tabulating two categorical variables together gives you a cross-tabulation of those variables, e.g. tabulate x, row col chi2

     pwcorr x y, sig gives you the pairwise correlation of two continuous variables.

5 Multivariate Statistics\footnote{Stata’s factor variables allow one to generate interactions and indicator variables on the fly, but again are beyond the purview of this guide.}

regress y x regresses y on x.\footnote{After running many multivariate models estat summarize will give you simple descriptive statistics for the specific sample used in that particular analysis.}

     regress y x z regresses y on x and z.\footnote{Other regression commands follow a very similar format: command y x z but are beyond the purview of this 2 page guide.}

     regress y x i.z regresses y on x and z, treating x as continuous and z as a set of categorical indicator variables.

6 Graphing\footnote{For all graphs, options after a “,” will be helpful in titling your graph e.g. twoway lfit y x, title(“.”) xtitle(“.”) ytitle(“.”)}

histogram x will give you a nice display of one variable. histogram x, by(y) may be useful for comparing the distributions of two variables over the categories of y.

     histogram x, percent will scale the y-axis more intuitively in terms of percentages. histogram x, discrete\footnote{The percent and discrete options can be combined.} gives a nicer display for categorical variables.

     twoway scatter y x gives you a twoway scatterplot of your data. twoway lfit y x will give you a linear fit graph. The two syntaxes may be combined e.g. twoway (scatter y x)(lfit y x).

     graph bar x, over(y) is useful for creating a bar graph of a continuous or categorical variable graphed across the categories of a categorical variable.

7 by:

In many cases you may want to look at the results of some calculation for x, or x and y over a third variable z. In such cases the by: syntax will be especially useful. For example to look at the correlation of x and y over different values of z.

     sort z by z: pwcorr x y, sig