A short guide to

commands

DR. MOHAMED ELSHERIF

Stata language

The standard Stata syntax

[by:] [command] [variablename(s)] [if] [in] [, options]

Variablename(s): can be one variable: [varname], or multiple variables [varlist] based on the command

If: to run the analysis only on a subset of the data

In: to run the analysis in selected of a range of observations (rows) in the dataset

Options: used to modify/add to the command

Stata language, helpful commands

by [varlist]

Prefix to run the commands separately in the categories of the specified variables. It comes after sorting the data using that variable.

by [varlist]: [command] [variablename(s)] [if] [in] [, options]

by Sex: sum Age

sort [varlist]

Sorts the observations in the dataset from low to high based on the variables listed.

sort Sex

bysort [varlist]

It performs sorting and running the commands separately in one step (sort+by)

bysort Sex: sum Age

Stata language, helpful commands

[if] condition

to run the analysis only on a subset of the data based on a specific condition

Operators that can be used with if:

==	Equal	!=	Not equals
>	Greater than	&	And
<	Less than	|	Or
>=	Greater than or equal to	!	Not
<=	Less than or equal to	~	Not

```
tab Sex if Age == 40
tab Sex if Age < 40
```

Stata language, helpful commands

display

To use Stata as a calculator.

display 2+3

help [command/text]

To get information about that command, or search for the text

help display

Exploring data

describe [varlist]

It gives name of the dataset, number of variables and observations, list of all variables with the names, labels, and storage type.

describe

describe Age Drug

inspect [varlist]

It shows the number of missing and non-missing values and the number of unique values.

inspect

inspect Age Drug

Exploring data

codebook [varlist]

It shows the name, label, type, range, number of observations and number of missing values of each variable in the dataset.

codebook

codebook Age Drug

codebook [varlist] , compact

This option shows the number of observations (obs), number of different values (unique), mean, lowest value (min), highest value (max) and the variable label (label).

codebook , compact

codebook Age Drug , compact

Exploring data, listing observations

list [varlist] if

It shows a list of cases with data of all variables mentioned in [varlist] that meet the condition

list Height Weight Marital if Age == 35

list Height Weight Marital

list if Age == 35

Summarizing data, numeric variables

summarize [varlist]

It shows the number of observations, mean, SD and range.

sum

sum Age Height Weight

summarize [varlist] , detail

It gives extra details as percentiles, largest 4 values, smallest 4 values, and others.

sum , detail

sum Age Height Weight , detail

mean [varlist]

it gives the mean , standard error, and 95% CI of the mean.

mean Age

Summarizing data, numeric variables

tabstat [varlist] , statistics (name of required statistics)

It is used to get specified summary statistics.

tabstat Age Height Weight , s(mean sd median iqr)

The option col(s) can be used for the summary statistics to be on columns instead of rows

tabstat Age Height Weight , s(mean sd median iqr) col(s)

Some summary statistics that can be obtained from this command:

mean	mean	range	range = max – min
sd	standard deviation	median	median
max	maximum	iqr	interquartile range = p75 - p25
min	minimum	cv	coefficient of variation (sd/mean)

Summarizing data, numeric variables in groups

bysort [varlist] : summarize [varlist]

It shows the summary statistics for each category of the variables listed after by or bysort.

bysort Sex: sum Age Height Weight

bysort [varlist] : tabstat [varlist] , statistics ( )

It shows the summary statistics for each category of the variables listed after by or bysort.

bysort Sex: tabstat Age Height Weight , s(mean sd median iqr) col(s)

mean [varlist] , over ([varlist])

It gives the mean , standard error, and 95% CI of the mean in each of the specified categories in the option over().

mean Age , over (Sex)

Summarizing data, categorical variables

tabulate [varname]

It shows the numbers (frequency) and percentages (relative frequency) for a specific variable

tab Sex

tab1 [varlist]

It shows the numbers (frequency) and percentages (relative frequency) for multiple variables

tab1 Sex Edu Marital Excercise

bysort [varname]: tabulate [varname]

To get numbers and percentages for a specific variable in each of the categories of another variable

bysort Edu: tab Sex

Summarizing data, categorical variables (two way)

tabulate [varname1] [varname2]

It gives a two-way table for the TOW specified variables

tab Sex Edu

tabulate [varname1] [varname2] , row

It gives a two-way table for the TOW specified variables with percentages per rows

tab Sex Edu , row

tabulate [varname1] [varname2] , column

It gives a two-way table for the TOW specified variables with percentages per columns

tab Sex Edu , col

Testing for normality of distribution

Shapiro-Wilk test for normality

swilk [varlist]

It is used for normality testing of specified variables

swilk Age Height

bysort [varlist] : swilk [varlist]

It is used for normality testing of specified variables in different categories

bysort Sex : swilk Age Height

Testing for equality of variance

F test for equality of variance (homogeneity of variance)

sdtest [varname] , by ([groupingvariable])

It is used for testing equality of variance between groups for a specified variable

sdtest Age , by (Sex)

Levene's test for equality of variance (homogeneity of variance)

robvar [varname] , by ([groupingvariable])

It is used for testing equality of variance between groups for a specified variable

robvar Age , by (Sex)

Bartlett's test for equality of variance

oneway [varname] [groupingvariable]

It is used for testing equality of variance, and it is part of the oneway command output

oneway Age Sex

One-sample t-test

ttest [varname] =[value]

To test if the mean of a variable is equal to a specified value (for a single group)

ttest Age =35

Paired samples t test

ttest [varname1] = [varname2]

To test if the mean of the two variables is equal (difference is equal to zero). Data is presented in pairs

ttest Exampre = Examafter

Independent samples t-test

With equality of variance

ttest [varname] , by ([groupingvariable])

To compare the mean of two groups (testing if the mean in the two groups is equal)

Before running the test, we need to check for the normality of distribution and equality of variance.

ttest Age , by (Sex)

Without equality of variance

ttest [varname] , by ([groupingvariable]) unequal

If there is no equality of variance, the option unequal is added

ttest Age , by (Sex) unequal

One-way ANOVA

With equality of variance

oneway [varname] [groupingvariable]

To compare the mean of more than two groups (testing if the mean in the groups is equal)

Before running the test, we need to check for the normality of distribution and equality of variance. Equality of variance is part of the output “Bartlett's test”

oneway Age Edu

To add table for summary statistics

oneway [varname] [groupingvariable] , tab

Option tab is used to get summary statistics as mean and sd per group

oneway Age Edu , tab

One-way ANOVA

To get the result for Bonferroni post hoc test

oneway [varname] [groupingvariable] , bonferroni

To get the result for Bonferroni post hoc test

oneway Age Edu , bonferroni

If there is no equality of variance, Welch ANOVA test is used

findit wtest

wtest [varname] [groupingvariable]

First, we need to install the wtest command using “findit wtest”

wtest Age Edu

Correlation

Pearson's correlation

pwcorr [varlist]

To get the Pearson's correlation coefficient between two numeric variables.

pwcorr Age Height

If more than two variables are used, a correlation matrix is produced

pwcorr Age Height Weight

To add the p-value and number of observations

pwcorr [varlist] , sig obs

Option sig is used to get p-value and option obs is used to get the number of observations (pairs)

pwcorr Age Height Weight , sig obs

Correlation

Confidence interval for Pearson's correlation coefficient

findit corrci

corrci [varlist]

To get the Pearson's correlation coefficient with the 95% CI. The command corrci needs to be installed first using "findit corrci"

corrci Age Height Weight

Spearman's correlation

spearman [varlist]

To get the Spearman's correlation coefficient between two variables.

spearman Age Height Weight

Mann Whitney test = Two-sample Wilcoxon rank-sum

ranksum [varname] , by ([groupingvariable])

It is the non-parametric equivalent of independent samples t-test to compare two groups

ranksum Age , by (Sex)

Wilcoxon signed-rank test

signrank [varname1] = [varname2]

It is the non-parametric equivalent of paired samples t-test to compare two variables (paired data)

signrank Exampre = Examafter

Kruskal Wallis test

Kruskal Wallis test

kwallis [varname] , by ([groupingvariable])

It is the non-parametric equivalent of one-way ANOVA to compare more than two groups

kwallis Age , by (Edu)

Chi-square test

Chi-square test

tabulate [varname1] [varname2] , chi2

It is the non-parametric test to study the association between two categorical variables

tab Sex Edu , chi2

Chi-square test, with percentages per row

tabulate [varname1] [varname2] , chi2 row

tab Sex Edu , chi2 row

Chi-square test, with percentages per column

tabulate [varname1] [varname2] , chi2 col

tab Sex Edu , chi2 col

Chi-square test

Chi-square test, with expected values

tabulate [varname1] [varname2] , chi2 expect

tab Sex Edu , chi2 expect

Chi-square test, with other measures of association

tabulate [varname1] [varname2] , all

tab Sex Edu , all

Fisher's exact test

tabulate [varname1] [varname2] , exact

tab Sex Edu , exact

Linear regression

Linear regression

regress [dependent var] [independent varlist]

regress Weight Age Height

Linear regression using categorical independent variable

regress [dependent var] i.[independent variable]

We add i. before the independent categorical variable

the first category is taken by default as reference category

regress Weight Age Height i.Edu

Logistic regression

Logistic regression

logit [dependent var] [independent varlist]

logit StatsTraining Age Height Weight

Logistic regression reporting OR

logit [dependent var] [independent varlist] , or

logistic [dependent var] [independent varlist]

To report the odds ratio, we use the option , or

Another option is to use the command logistic

logit StatsTraining Age Height Weight , or

logistic StatsTraining Age Height Weight