

Data Analysis with Stata 14.1 Cheat Sheet

For more info see Stata's reference manual (stata.com)
Results are stored as either **r**-class or **e**-class. See **Programming Cheat Sheet**
Summarize Data — Examples use auto.dta (sysuse auto, clear) unless otherwise noted

univar price mpg, **boxplot** — calculate univariate summary, with box-and-whiskers plot
stem mpg — return stem-and-leaf display of mpg
summarize price mpg, **detail** — frequently used commands are highlighted in yellow
ci mean mpg price, **level**(99) — for Stata 13: **ci** mpg price, **level**(99)
compute standard errors and confidence intervals
correlate mpg price — return correlation or covariance matrix
pwcorr price mpg weight, **star**(0.05) — return all pairwise correlation coefficients with sig. levels
mean price mpg — estimates of means, including standard errors
proportion rep78 foreign — estimates of proportions, including standard errors for categories identified in varlist
ratio — estimates of ratio, including standard errors
total price — estimates of totals, including standard errors

Statistical Tests

tabulate foreign rep78, **chi2 exact expected** — tabulate foreign and repair record and return chi² and Fisher's exact statistic alongside the expected values
ttest mpg, **by**(foreign) — estimate t test on equality of means for mpg by foreign
prtest foreign == 0.5 — one-sample test of proportions
ksmirnov mpg, **by**(foreign) **exact** — Kolmogorov-Smirnov equality-of-distributions test
ranksum mpg, **by**(foreign) **exact** — equality tests on unmatched data (independent samples)
anova systolic drug, **webuse** systolic, **clear** — analysis of variance and covariance
pwmean mpg, **over**(rep78) **pveffects** **mcompare**(tukey) — estimate pairwise comparisons of means with equal variances include multiple comparison adjustment

Estimation with Categorical & Factor Variables

CONTINUOUS VARIABLES	OPERATOR	DESCRIPTION	EXAMPLE
measure something	i.	specify indicators	regress price i.rep78
CATEGORICAL VARIABLES	ib.	specify base indicator	regress price ib(3).rep78
	fvset	command to change base	fvset base frequent rep78
INDICATOR VARIABLES	c.	treat variable as continuous	regress price i.foreign#c.mpg i.foreign
	o.	omit a variable or indicator	regress price io(2).rep78
	#	specify interactions	regress price mpg c.mpg#c.mpg
T F denote whether something is true or false	##	specify factorial interactions	regress price c.mpg###c.mpg

Declare Data

By declaring data type, you enable Stata to apply data munging and analysis functions specific to certain data types

TIME SERIES

tsset time, **yearly** — declare sunspot data to be yearly time series
tsreport — report time series aspects of a dataset
generate lag_spot = L1.spot — create a new variable of annual lags of sun spots
tsline spot — plot time series of sunspots
arima spot, **ar**(1/2) — estimate an auto-regressive model with 2 lags
TIME SERIES OPERATORS
L. lag x_{t-1} L2. 2-period lag x_{t-2}
F. lead x_{t+1} F2. 2-period lead x_{t+2}
D. difference $x_t - x_{t-1}$ D2. difference of difference $x_t - x_{t-1} - (x_{t-1} - x_{t-2})$
S. seasonal difference $x_t - x_{t-12}$ S2. lag-2 (seasonal difference) $x_t - x_{t-2}$
USEFUL ADD-INS
tscollapse — compact time series into means, sums and end-of-period values
carryforward — carry non-missing values forward from one obs. to the next
tsspell — identify spells or runs in time series

SURVIVAL ANALYSIS

stset studytime, **failure**(died) — declare survey design for a dataset
stsum — summarize survival-time data
stcox drug age — estimate a cox proportional hazard model

PANEL / LONGITUDINAL

xtset id year — declare national longitudinal data to be a panel
xtdescribe — report panel aspects of a dataset
xtsum hours — summarize hours worked, decomposing standard deviation into between and within components
xtline ln_wage if id <= 22, **tlab**el(#3) — plot panel data as a line plot
xtreg ln_w c.age##c.age ttl_exp, **fe** **vce**(robust) — estimate a fixed-effects model with robust standard errors
SURVEY DATA
svyset psuid [pweight = finalwgt], **strata**(stratid) — declare survey design for a dataset
svydescribe — report survey data details
svy: mean age, **over**(sex) — estimate a population mean for each subpopulation
svy, subpop(rural): mean age — estimate a population mean for rural areas
svy: tabulate sex heartatk — report two-way table with tests of independence
svy: reg zinc c.age##c.age female weight rural — estimate a regression using survey weights

1 Estimate Models

stores results as **e**-class

regress price mpg weight, **robust** — estimate ordinary least squares (OLS) model on mpg weight and foreign, apply robust standard errors
regress price mpg weight if foreign == 0, **cluster**(rep78) — regress price only on domestic cars, cluster standard errors
rreg price mpg weight, **genwt**(reg_wt) — estimate robust regression to eliminate outliers
probit foreign turn price, **vce**(robust) — estimate probit regression with robust standard errors
logit foreign headroom mpg, **or** — estimate logistic regression and report odds ratios
bootstrap, reps(100): **regress** mpg /* — estimate regression with bootstrapping
jackknife r(mean), **double:** **sum** mpg — jackknife standard error of sample mean

ADDITIONAL MODELS

pca	built-in Stata command	principal components analysis
factor		factor analysis
poisson	nbreg	count outcomes
tobit		censored data
ivregress	ivreg2	instrumental variables
diff	user-written	difference-in-difference
rd	ssc install ivreg2	regression discontinuity
xtabond	xtabond2	dynamic panel estimator
psmatch2		propensity score matching
synth		synthetic control analysis
oaxaca		Blinder-Oaxaca decomposition

2 Diagnostics

not appropriate after robust cluster()

estat **hettest** — test for heteroskedasticity
ovtest — test for omitted variable bias
vif — report variance inflation factor
dfbeta(length) — calculate measure of influence
rvfplot, yline(0) — plot residuals against fitted values
avplots — plot all partial-regression leverage plots in one graph

3 Postestimation

commands that use a fitted model

regress price headroom length — Used in all postestimation examples
display _b[length] **display** _se[length] — return coefficient estimate or standard error for mpg from most recent regression model
margins, dydx(length) — returns e-class information when post option is used
margins, eyex(length) — return the estimated marginal effect for mpg
predict yhat if **e**(sample) — create predictions for sample on which model was fit
predict double resid, **residuals** — calculate residuals based on last fit model
test mpg = 0 — test linear hypotheses that mpg estimate equals zero
lincom headroom - length — test linear combination of estimates (headroom = length)