Data Analysis with Stata 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** — ssc install univar
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
calculate a variety of univariate summary statistics

ci mean mpg price, **level**(99) — for Stata 13: **ci** mpg price, **level** (99)
compute standard errors and confidence intervals

r

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

e

ratio price/mpg
estimates of ratio, including standard errors

total price
estimates of totals, including standard errors

Declare data

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

Time series

webuse sunspot, clear

tsset time, **yearly**
declare sunspot data to be yearly time series

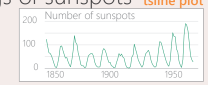
r

tsreport
report time-series aspects of a dataset

generate lag_spot = L1.spot
create a new variable of annual lags of sunspots

tsline spot
plot time series of sunspots

tsline plot



e

arima spot, **ar**(1/2)
fit an autoregressive 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 $x_t - x_{t-1}$	S2.	lag-2 (seasonal difference) $x_t - x_{t-2}$

Useful add-ins

tscollap compact time series into means, sums, and end-of-period values
carryforward carry nonmissing values forward from one obs. to the next
tsspell identify spells or runs in time series

Survival analysis

webuse drugtr, clear

r

stset studytime, **failure**(died)
declare survey design for a dataset

stsum
summarize survival-time data

e

stcox drug age
fit a Cox proportional hazards model

Panel / longitudinal

webuse nlswork, clear

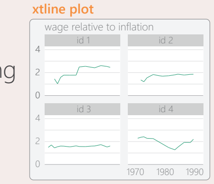
xtset id year
declare national longitudinal data to be a panel

r

xtdescribe
report panel aspects of a dataset

xtsum hours
summarize hours worked, decomposing standard deviation into between and within components

xtline plot



xtline ln_wage if id <= 22, **tlabel**(#3)
plot panel data as a line plot

e

xtreg ln_w c.age##c.age ttl_exp, **fe vce**(robust)
fit a fixed-effects model with robust standard errors

Survey data

webuse nhanes2b, clear

svyset psuid [**pweight** = finalwgt], **strata**(stratid)
declare survey design for a dataset

r

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

e

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

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

r

prtest foreign == 0.5
one-sample test of proportions

ksmirnov mpg, **by**(foreign) **exact**
Kolmogorov–Smirnov equality-of-distributions test

ranksum mpg, **by**(foreign)
equality tests on unmatched data (independent samples)

anova systolic drug — **webuse** systolic, clear
analysis of variance and covariance

e

pwmean mpg, **over**(rep78) **pveffects mcompare**(tukey)
estimate pairwise comparisons of means with equal variances include multiple comparison adjustment

1 Fit models

stores results as e-class

regress price mpg weight, **vce**(robust)
fit ordinary least-squares (OLS) model on mpg, weight, and foreign, apply robust standard errors

regress price mpg weight **if** foreign == 0, **vce**(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 /* */ weight gear foreign
estimate regression with bootstrapping

jackknife r(mean): sum mpg
jackknife standard error of sample mean

Additional models

built-in Stata command: pca · user-written: ivreg2 (ssc install ivreg2)

pca	principal components analysis
factor	factor analysis
poisson • **nbreg**	count outcomes
tobit	censored data
ivregress **ivreg2**	instrumental variables
didregress	difference-in-difference
rd	regression discontinuity
xtabond **xtdpdsys**	dynamic panel estimator
teffects psmatch	propensity score matching
synth	synthetic control analysis
oaxaca	Blinder–Oaxaca decomposition

2 Diagnostics

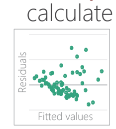
some are inappropriate with robust SEs

r

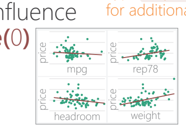
estat hettest test for heteroskedasticity
ovtest test for omitted-variable bias
vif report variance inflation factor

dfbeta(length)
calculate measure of influence

Type help regress postestimation plots for additional diagnostic plots



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 length from most recent regression model

margins, dydx(length) — returns e-class information when post option is used
return the estimated marginal effect for length

r

margins, eyex(length)
return the estimated elasticity for length

predict yhat if **e(sample)**
create predictions for sample on which model was fit

predict double resid, **residuals**
calculate residuals based on last fitted model

test headroom = 0
test linear hypotheses that headroom estimate equals zero

r

lincom headroom - length
estimate linear combination (headroom - length)

Estimation with categorical & factor variables

more details at https://www.stata.com/manuals/u26.pdf

Continuous variables — measure something

Categorical variables — identify a group to which an observation belongs

Indicator variables — T F — denote whether something is true or false

Operator	Description	Example	
i.	specify indicators	regress price **i.**rep78	specify rep78 variable to be an indicator variable
ib.	specify base indicator	regress price **ib**(3)**.**rep78	set the third category of rep78 to be the base category
fvset	command to change base	fvset base frequent rep78	set the base to most frequently occurring category for rep78
c.	treat variable as continuous	regress price i.foreign#**c.**mpg i.foreign	treat mpg as a continuous variable and specify an interaction between foreign and mpg
o.	omit a variable or indicator	regress price **io**(2)**.**rep78	set rep78 as an indicator; omit observations with rep78 == 2
#	specify interactions	regress price mpg c.mpg**#**c.mpg	create a squared mpg term to be used in regression
##	specify factorial interactions	regress price c.mpg**##**c.mpg	create all possible interactions with mpg (mpg and mpg²)

Tim Essam (tessam@usaid.gov) • Laura Hughes (lhughes@usaid.gov)
follow us @StataRGIS and @flaneuseks

inspired by RStudio's awesome Cheat Sheets (rstudio.com/resources/cheatsheets)

geocenter.github.io/StataTraining

Disclaimer: we are not affiliated with Stata. But we like it.

updated May 2021