2 A brief description of Stata

Stata is a statistical package for managing, analyzing, and graphing data.

Stata is available for a variety of platforms. Stata may be used either as a point-and-click application or as a command-driven package.

Stata's GUI provides an easy interface for those new to Stata and for experienced Stata users who wish to execute a command that they seldom use.

The command language provides a fast way to communicate with Stata and to communicate more complex ideas.

Here is an extract of a Stata session using the GUI:

(Throughout the Stata manuals, we will refer to various datasets. These datasets are all available from https://www.stata-press.com/data/r19/. For easy access to them within Stata, type webuse dataset_name, or select File > Example datasets... and click on Stata 19 manual datasets.)

. webuse lbw (Hosmer & Lemeshow data)

We select **Data > Describe data > Summary statistics** and choose to summarize variables low, age, and smoke, whose names we obtained from the Variables window. We click on **OK**.

😑 sum	marize -	Summary s	tatistics		-			×
Main	by/if/in	Weights						
Variabl	es: (leave	empty for	all variab	les)				
low a	ge smoke	2					\sim	
Examp	les:	yr*	all vari	ables startin	ig with "yr"			
	3	xyz-abc	all vari	ables betwe	en xyz and ab	C		
	andard d isplay add o display se variabl 5 🗣 S	ditional sta ; just calcu le's display	late mear format ne every N	l variables (set 0 for none	=)		
? C				ОК	Cancel		Subn	nit

. sı	ummarize lo	w age smoke				
	Variable	Obs	Mean	Std. dev.	Min	Max
	low	189	.3121693	.4646093	0	1
	age	189	23.2381	5.298678	14	45
	smoke	189	.3915344	.4893898	0	1

Stata shows us the command that we could have typed in command mode—summarize low age smoke—before displaying the results of our request.

Next we fit a logistic regression model of low on age and smoke. We select Statistics > Binary outcomes > Logistic regression, fill in the fields, and click on OK.

	tic - Lo	ogistic regress	ion, reportin	g odds ratios	i	_	
Model	by/if	/in Weights	SE/Robust	Reporting	Maximization		
Depen	dent va	ariable: I	ndependent	variables:			
low		~	age smoke				~
		[Suppress of	constant terr	n		
Optio	ons						
Offse	et varia	ble:					
		~					
	etain n	erfect predicto	y variables				
	traints		i vanabies				
Cons	traints					~	Manage
						· ·	managem
? C					OK	Cancel	Submit
logistic	low	age smoke					
gistic r		•				Number of ob	os = 18
5	-0					LR chi2(2)	= 7.4
m likoli	hood	= -113.638	15			Prob > chi2 Pseudo R2	
3 IIKEII.	поод	113.030	10			rseuu0 KZ	- 0.031
1	ow	Odds ratio	Std. e	rr.	z P> z	[95% conf.	interval

low	Odds ratio	Std. err.	Z	P> z	[95% conf.	interval]
age	.9514394	.0304194	-1.56	0.119	.8936482	1.012968
smoke	1.997405	.642777	2.15	0.032	1.063027	3.753081
_cons	1.062798	.8048781	0.08	0.936	.2408901	4.689025

Note: _cons estimates baseline odds.

Here is an extract of a Stata session using the command language:

. use https://www (1978 automobile	-	ss.com/data/	r19/auto		
. summarize mpg w	eight				
Variable	Obs	Mean	Std. dev.	Min	Max
mpg	74	21.2973	5.785503	12	41
weight	74	3019.459	777.1936	1760	4840

The user typed summarize mpg weight and Stata responded with a table of summary statistics. Other commands would produce different results:

- . generate gp100m = 100/mpg
- . label var gp100m "Gallons per 100 miles"
- . format gp100m %5.2f

weight

- correlate gp100m weight
- (obs=74)

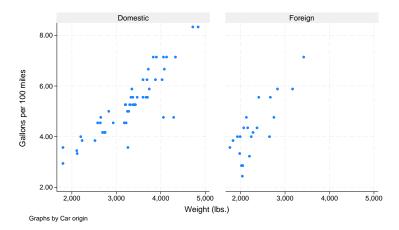
gp100m weight gp100m 1.0000 0.8544

1.0000

. regress gp100m weight gear_ratio

Source	SS	df	MS	Number of obs	=	74
Model	87.4543721	2	43.7271861	F(2, 71) Prob > F	=	96.65 0.0000
Residual	32.1218886	71	.452420967		=	0.7314
Total	119.576261	73	1.63803097	Adj R-squared Root MSE	=	0.7238 .67262
gp100m	Coefficient	Std. err.	t	P> t [95% co	onf.	interval]
weight gear_ratio _cons	.0014769 .1566091 .0878243	.0001556 .2651131 1.198434	0.59	0.000 .001166 0.557372013 0.942 -2.30178	15	.0017872 .6852297 2.477435

. scatter gp100m weight, by(foreign)



The user-interface model is type a little, get a little, etc., so that the user is always in control.

Stata's model for a dataset is that of a table—the rows are the observations and the columns are the variables:

	mpg	weight	gp100m
1.	22	2,930	4.55
2.	17	3,350	5.88
3.	22	2,640	4.55
4.	20	3,250	5.00
5.	15	4,080	6.67
6.	18	3,670	5.56
7.	26	2,230	3.85
8.	20	3,280	5.00
9.	16	3,880	6.25
10.	19	3,400	5.26

. list mpg weight gp100m in 1/10

Observations are numbered; variables are named.

Stata is fast. That speed is due partly to careful programming, and partly because Stata keeps the data in memory. Stata's file model is that of a word processor: a dataset may exist on disk, but the dataset in memory is a copy. Datasets are loaded into memory, where they are worked on, analyzed, changed, and then perhaps stored back on disk.

Working on a copy of the data in memory makes Stata safe for interactive use. The only way to harm the permanent copy of your data on disk is if you explicitly save over it.

Having the data in memory means that the dataset size is limited by the amount of computer memory. Stata stores the data in memory in an efficient format—you will be surprised how much data can fit. Nevertheless, if you work with extremely large datasets, you may run into memory constraints. You will want to learn how to store your data as efficiently as possible; see [D] compress.

Stata, Stata Press, and Mata are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. StataNow and NetCourseNow are trademarks of StataCorp LLC. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2025 StataCorp LLC, College Station, TX, USA. All rights reserved.



For suggested citations, see the FAQ on citing Stata documentation.