

Introduction to Stata for students

Chris Cheng

Senior Econometrician

StataCorp LLC

Stata webinar

[materials here](#)

Introduction

Why Stata
Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets
Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

- 1 Introduction
- 2 Interface navigation
- 3 Getting data into Stata
- 4 Data manipulation and management
- 5 Data visualization
- 6 Statistical analysis
- 7 Learning resources

Introduction

Why Stata
Who uses Stata

**Interface
navigation****Getting data
into Stata****Data
manipulation
and
management**

Exploring datasets
Exploring variables

**Data
visualization****Statistical
analysis**

Data analysis

**Learning
resources**

Introduction

Introduction

Why Stata

Who uses Stata

Interface
navigation

Getting data
into Stata

Data
manipulation
and
management
Exploring datasets
Exploring variables

Data
visualization

Statistical
analysis
Data analysis

Learning
resources

“Stata delivers everything you need for reproducible data analysis—powerful statistics, visualization, data manipulation, and automated reporting—all in one intuitive platform.”

One tool. One workflow. From raw data to final report.

Who uses Stata?

Introduction

Why Stata

Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets

Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

Economics and business

- Economics
- Econometrics
- Finance
- Accounting
- Marketing
- Management

Public policy and social sciences

- Public policy
- Political science
- Sociology
- Demography
- Education

Health and life sciences

- Biostatistics
- Epidemiology
- Public health
- Medicine
- Pharmaceuticals

Science, engineering, and data science

- Statistics
- Data science
- Engineering
- Environmental science

Government, NGOs, and industry

- Government agencies
- Central banks
- Consulting
- Research organizations

Introduction

[Why Stata](#)[Who uses Stata](#)**Interface
navigation**[Getting data
into Stata](#)[Data
manipulation
and
management](#)[Exploring datasets](#)[Exploring variables](#)[Data
visualization](#)[Statistical
analysis](#)[Data analysis](#)[Learning
resources](#)

Interface navigation

Stata's Interface — GUI

- Start Stata from the Desktop, the Start menu, or the Applications folder (Mac)

The screenshot shows the Stata GUI with several annotations pointing to different parts of the interface:

- Past commands appear here:** Points to the History window on the left, which lists previously entered commands.
- Current dataset name appears here:** Points to the top of the main window, which displays "C:\Stata\ado\base\1\auto.dta".
- Results are displayed here:** Points to the main command window, which shows the output of the `regress` command, including summary statistics and coefficient estimates.
- Variable list appears here:** Points to the Variables window on the right, which lists the variables in the current dataset.
- Data properties appear here:** Points to the Properties window on the right, which shows details about the current dataset, such as the number of observations and variables.
- Current working directory appears here:** Points to the bottom left of the main window, which displays "C:\Users\Stata_User\Documents".
- Commands are typed here:** Points to the Command window at the bottom, where new commands are entered.
- Current log status appears here:** Points to the "log on (smcl)" status indicator at the bottom right of the Command window.
- Command log status appears here:** Points to the "cmdlog on" status indicator at the bottom right of the Command window.

The main command window displays the following output:

```

. generate gp100m = 100/ mpg
. regress gp100m weight

Source      SS           df          MS              Number of obs   =      74
             87.2964969          1      87.2964969          F(1, 72)         =    194.71
Residual    32.2797639          72     .448330054          Prob > F          =    0.0000
Total      119.576261          73     1.63803997          R-squared         =    0.7300
                                           Adj R-squared     =    0.7263
                                           Root MSE         =    .66957

gp100m      Coefficient   Std. err.      t    P>|t|     [95% conf. interval]
weight      .001407       .0001008       13.95   0.000       .001206   .0016081
_._cons     .7707669       .3142571       2.45    0.017       .1443069   1.397227

. predict yhat
(option xb assumed; fitted values)

. cd C:\Users\Stata_User\Documents
C:\Users\Stata_User\Documents

```

The Variables window on the right lists the following variables:

Name	Label
make	Make and model
price	Price
mpg	Mileage (mpg)
rep78	Repair record 1978
headroom	Headroom (in.)
trunk	Trunk space (cu. ft.)
weight	Weight (lbs.)
length	Length (in.)
turn	Turn circle (ft.)
displacement	Displacement (cu. in.)
gear_ratio	Gear ratio

The Properties window on the right shows the following data properties:

Property	Value
Frame	default
Filename	auto.dta
Label	1978 automobile data
Notes	1 note
Variables	14
Observations	74
Size	3.8KB
Memory	64MB
Sorted by	foreign

Stata's Interface — Data Editor

Introduction

Why Stata
Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets
Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

The diagram illustrates the Stata Data Editor interface. At the top is a menu bar with 'File', 'Edit', 'View', 'Data', and 'Tools'. Below it is a toolbar with icons for opening files, saving, printing, and searching. The main area displays a data table. A callout box labeled 'observation number' points to the number '4' in the first column. Another callout box labeled 'variable name' points to the text 'price[4]' in the second column. A third callout box labeled 'value' points to the number '4816' in the third column. The data table has columns: observation number, make, price, mpg, rep78, and headroom.

	make	price	mpg	rep78	headroom
1	AMC Concord	4,099	22	3	2.5
2	AMC Pacer	4,749	17	3	3.0
3	AMC Spirit	3,799	22	.	3.0
4	Buick Century	4,816	20	3	4.5
5	Buick Electra	7,827	15	4	4.0

Stata's Interface — Dialog Box

Introduction

Why Stata

Who uses Stata

Interface
navigation

Getting data
into Stata

Data
manipulation
and
management

Exploring datasets

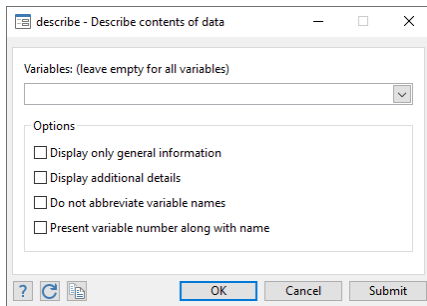
Exploring variables

Data
visualization

Statistical
analysis

Data analysis

Learning
resources



- **OK:** execute command and close dialog box
- **Cancel:** cancel options and close dialog box
- **Submit:** execute command and keep dialog box open

Stata's Interface — Do-file Editor

Introduction

Why Stata

Who uses Stata

Interface
navigation

Getting data
into Stata

Data
manipulation
and
management

Exploring datasets

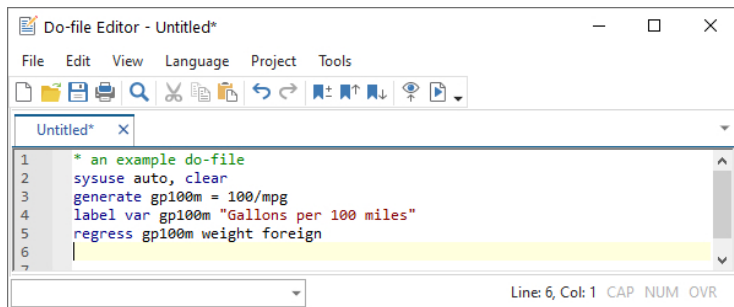
Exploring variables

Data
visualization

Statistical
analysis

Data analysis

Learning
resources



The screenshot shows the Stata Do-file Editor window titled "Do-file Editor - Untitled*". The window has a menu bar with "File", "Edit", "View", "Language", "Project", and "Tools". Below the menu bar is a toolbar with icons for opening files, saving, printing, searching, cutting, pasting, undo, redo, and other editing functions. The main text area contains a sample do-file script with line numbers 1 through 7 on the left. The script is as follows:

```
1 * an example do-file
2 sysuse auto, clear
3 generate gp100m = 100/mpg
4 label var gp100m "Gallons per 100 miles"
5 regress gp100m weight foreign
6
7
```

The line numbers 1 through 7 are displayed in a light gray column on the left. The script text is color-coded: asterisks are green, keywords like sysuse, generate, label, and regress are blue, and variable names and strings are black. The line containing "regress gp100m weight foreign" is highlighted in yellow. At the bottom of the window, there is a status bar showing "Line: 6, Col: 1" and three tabs labeled "CAP", "NUM", and "OVR".

Navigating Stata: Menu vs. Command vs. Do-file Editor

Introduction

Why Stata
Who uses Stata

Interface navigation

Getting data
into Stata

Data
manipulation
and
management
Exploring datasets
Exploring variables

Data
visualization

Statistical
analysis
Data analysis

Learning
resources

- **Menu and dialog boxes:** Ideal for beginners; enable complex analyses without memorizing syntax.
- **Command window:** Faster once familiar; ideal for interactive exploration and quick checks.
- **Do-file Editor:** Essential for reproducibility, automation, and documenting your work.

Learning bridge: Dialog boxes display generated syntax in the Results window, helping you learn Stata commands as you click.

Introduction

[Why Stata](#)[Who uses Stata](#)Interface
navigation**Getting data
into Stata**Data
manipulation
and
management[Exploring datasets](#)[Exploring variables](#)Data
visualizationStatistical
analysis[Data analysis](#)Learning
resources

Getting data into Stata

Getting data into Stata

Introduction

Why Stata

Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets

Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

- Stata's native data format is `.dta`.
- Data from many common sources can be imported directly into Stata, including:
 - Excel files (`.xls`, `.xlsx`)
 - Text files (`.csv`, `.txt`)
 - Other statistical software (e.g., SAS, SPSS)
 - Geographic data (shapefiles)
 - Databases via ODBC/JDBC connections
- Once imported, all data are stored and analyzed in a single, consistent format.

Opening and importing data

Introduction

Why Stata

Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets

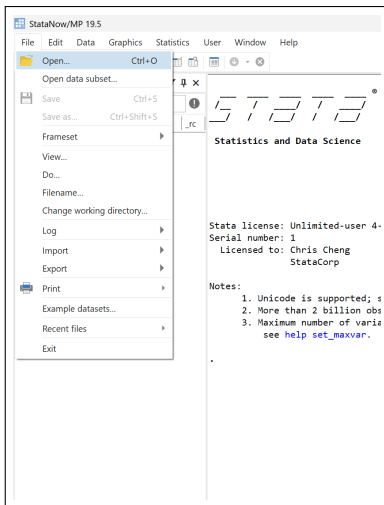
Exploring variables

Data visualization

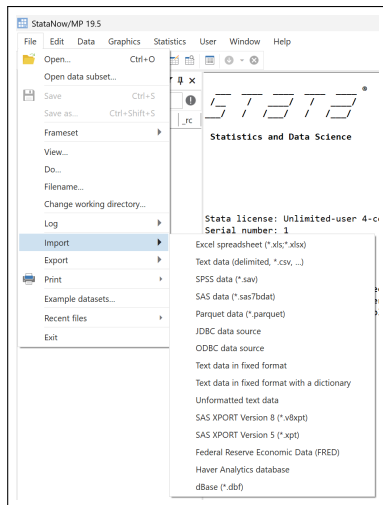
Statistical analysis

Data analysis

Learning resources



Opening .dta files



Importing other data files

Getting data into Stata: The big picture

Introduction

Why Stata

Who uses Stata

Interface
navigation

Getting data
into Stata

Data
manipulation
and
management

Exploring datasets
Exploring variables

Data
visualization

Statistical
analysis
Data analysis

Learning
resources

General syntax ⇒ `command <path/to/file>, options`

Built-in example data ⇒ `sysuse auto`

User Stata data files ⇒ `use <datafile.dta>`

Excel ⇒ `import excel <file.xlsx>`

Text ⇒ `import delimited <file.csv>`

SPSS ⇒ `import spss <file.sav>`

SAS ⇒ `import sas <file.sas7bdat>`

FRED data ⇒ `import fred <series_list>`

Shapefiles ⇒ `spshape2dta <shapefile>`

... ⇒ ...

Once imported (and saved), all data live in a .dta file, ready for analysis.

Introduction

Why Stata
Who uses Stata

Interface
navigationGetting data
into StataData
manipulation
and
management

Exploring datasets
Exploring variables

Data
visualizationStatistical
analysis

Data analysis

Learning
resources

Data manipulation and management

A quick peek at your data

Introduction

Why Stata

Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets

Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

- Let's load the dataset `marketcamp.dta` into our Stata session.
- Open the Data Editor to explore your variables and their values:



Browse mode: view only



Edit mode: make changes

- To view the structure of a dataset, use the `describe` command or navigate to **Data > Describe data > Describe data in memory or in a file.**
`. describe`

Introduction

Why Stata
Who uses Stata

Interface navigation

Getting data into Stata

Data
manipulation
and
management
Exploring datasets
Exploring variables

Data visualization

Statistical
analysis
Data analysis

Learning
resources

- **Cross-sectional (default)** — no declaration needed
- **Time-series** — `tsset time`
- **Panel/longitudinal** — `xtset panel time`
- **Survival/event-time** — `stset etime, failure(died)`
- **Complex survey** — `svyset sunit1, strata(groups)`

For all types except cross-sectional, declare the structure first; analyses will then handle time, panels, or survey designs automatically.

Understanding variable contents

Introduction

Why Stata

Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets

Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

- After understanding how a dataset is structured, the next step is examining individual variables.
- The `codebook` command summarizes variable values, labels, missing values, and storage types.
- Type `codebook` in the Command window or navigate to **Data > Describe data > Describe data contents (codebook)**.
 - . `codebook`

Variable types in Stata

Introduction

Why Stata

Who uses Stata

Interface
navigation

Getting data
into Stata

Data
manipulation
and
management

Exploring datasets

Exploring variables

Data
visualization

Statistical
analysis

Data analysis

Learning
resources

- **Numeric** (black in Data Editor)
 - **Continuous:** numeric values measured on a continuous scale (e.g., income, age, temperature)
 - **Categorical:** numeric codes representing discrete categories (e.g., 0/1 for gender, 1/2/3 for education levels)
- **String** (red in Data Editor)
 - Text values (e.g., names, addresses)
 - Cannot be used directly in most statistical models
 - Convert to numeric values using `encode` (categorical) or `destring` (numeric values stored as text)
 - `encode marital_status, generate(marital)`

Introduction

Why Stata
Who uses Stata

Interface
navigationGetting data
into StataData
manipulation
and
management

Exploring datasets
Exploring variables

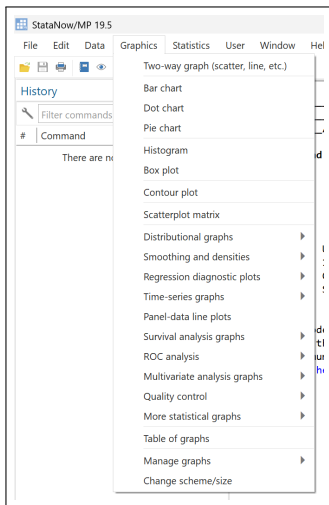
**Data
visualization**Statistical
analysis

Data analysis

Learning
resources

Data visualization

Stata data and visualization tools



Data visualization and graphics tools

★ Univariate graphs

- † Histograms
- † Bar charts
- † Box plots

★ Bivariate graphs

- † Scatterplots
- † Fitted regression lines
- † Line plots

Visualization in Stata

Introduction

Why Stata
Who uses Stata

Interface navigation

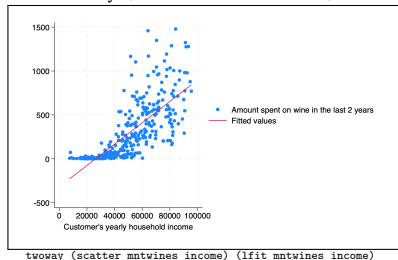
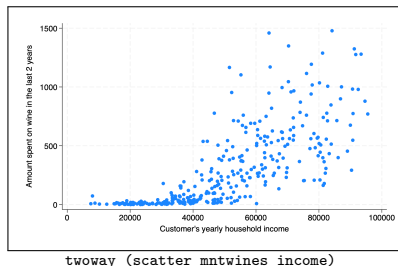
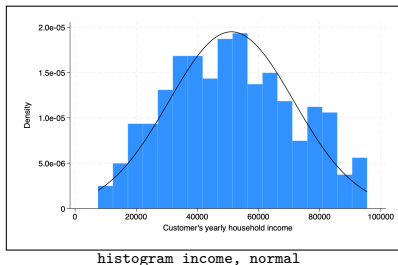
Getting data into Stata

Data manipulation and management
Exploring datasets
Exploring variables

Data visualization

Statistical analysis
Data analysis

Learning resources



Introduction

Why Stata
Who uses Stata

**Interface
navigation****Getting data
into Stata****Data
manipulation
and
management**

Exploring datasets
Exploring variables

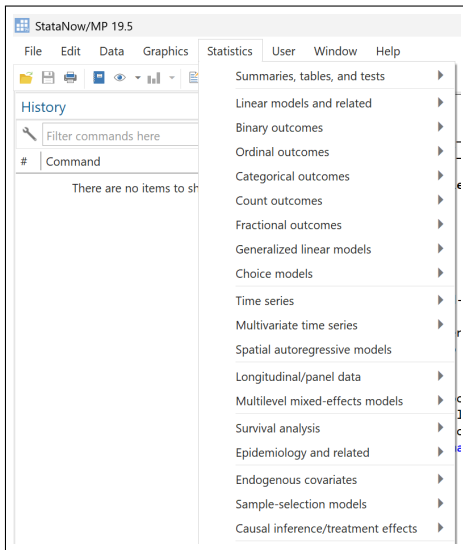
**Data
visualization****Statistical
analysis**

Data analysis

**Learning
resources**

Statistical analysis

Statistical tools in Stata



Partial list of statistical tools

- ★ Linear models
- ★ Multilevel mixed-effects
- ★ Binary, count, and limited outcomes
- ★ Choice models
- ★ Bayesian analysis
- ★ Time series
- ★ Panel/longitudinal data
- ★ Survival analysis
- ★ Meta-analysis
- ★ Power, precision, and sample size
- ★ Causal inference/Treatment effects
- ★ Multiple imputation
- ★ SEM
- ★ Latent class analysis
- ★ Tests, predictions, and effects
- ★ ANOVA/MANOVA
- ★ Survey methods
- ★ Nonparametric methods
- ★ Finite mixture models
- ★ Python integration
- ★ And more

Introduction

Why Stata

Who uses Stata

Interface
navigation

Getting data
into Stata

Data
manipulation
and
management

Exploring datasets
Exploring variables

Data
visualization

Statistical
analysis

Data analysis

Learning
resources

Introduction

Why Stata

Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets

Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

- Descriptive statistics summarize and organize characteristics of a sample dataset without making inferences about a broader population.

- For continuous variables, examine the mean, standard deviation, and overall distribution.

```
. summarize income kidhome mntwines
```

Variable	Obs	Mean	Std. dev.	Min	Max
income	328	51142.63	20467.57	7500	95529
kidhome	331	.4199396	.5298038	0	2
mntwines	331	303.0665	336.2635	0	1478

- For discrete or categorical variables, examine frequencies and relative proportions.

```
. tabulate marital_status
```

Customer's marital status	Freq.	Percent	Cum.
Divorced	45	13.60	13.60
Married	166	50.15	63.75
Single	102	30.82	94.56
Widow	18	5.44	100.00
Total	331	100.00	

Introduction

Why Stata

Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets

Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

- Inferential statistics use sample data to test hypotheses, estimate parameters, and draw conclusions about a population.

Introduction

Why Stata

Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets

Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

- Test whether wine consumption differs between consumers who filed a complaint and those who did not:

```
. ttest mntwines, by(complain)
```

Introduction

Why Stata

Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets

Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

Applied data analysis in Stata follows three core steps:

- Exploring and understanding the structure of the data
- Summarizing key variables using descriptive statistics and visualization
- **Estimating statistical models to answer substantive research questions**

Now, we turn to step three: estimating statistical models.

Regression as the primary tool

Introduction

Why Stata

Who uses Stata

Interface
navigation

Getting data
into Stata

Data
manipulation
and
management
Exploring datasets
Exploring variables

Data
visualization

Statistical
analysis
Data analysis

Learning
resources

Regression analysis is the foundation for empirical work in Stata.

- Quantifies relationships between outcome variables and covariates
- Provides a unified framework for hypothesis testing and inference
- Extends directly to advanced methods (e.g., Difference-in-Differences, fixed effects, IV)

Applied work in Stata centers on choosing an appropriate regression model and interpreting estimates correctly.

Regression examples

Introduction

Why Stata

Who uses Stata

Interface
navigationGetting data
into StataData
manipulation
and
managementExploring datasets
Exploring variablesData
visualizationStatistical
analysis

Data analysis

Learning
resources

Continuous outcome

The 'regress' dialog box is titled 'regress - Linear regression'. It features a tabbed interface with 'Model' selected. The 'Dependent variable' is 'mntwines' and the 'Independent variables' are 'income kidhome recency i.marital'. Under 'Treatment of constant', the options are: ☐ Suppress constant term, ☐ Has user-supplied constant, and ☐ Total SS with constant (advanced). The bottom has buttons for '?', a circular arrow icon, a document icon, 'OK', 'Cancel', and 'Submit'.

Linear regression: understanding wine expenditures

Binary outcome

The 'logistic' dialog box is titled 'logistic - Logistic regression, reporting odds ratios'. It features a tabbed interface with 'Model' selected. The 'Dependent variable' is 'complain' and the 'Independent variables' are 'income recency mntwines numdealspurchases i.marital'. There is an unchecked checkbox for 'Suppress constant term'. Under 'Options', there is an 'Offset variable' dropdown menu, an unchecked checkbox for 'Retain perfect predictor variables', and a 'Constraints' dropdown menu with a 'Manage...' button. The bottom has buttons for '?', a circular arrow icon, a document icon, 'OK', 'Cancel', and 'Submit'.

Logistic regression: understanding consumer complaints

A typical regression output in Stata

Dependent variable Independent variables

```
. regress mntwines income kidhome recency i.marital
```

Model fit: R^2 , F statistic, sample size

Source	SS	df	MS
Model	20974525.3	6	3495754.22
Residual	16235475.1	321	50577.8042
Total	37210000.5	327	113792.05

Number of obs	=	328
F(6, 321)	=	69.12
Prob > F	=	0.0000
R-squared	=	0.5637
Adj R-squared	=	0.5555
Root MSE	=	224.9

mntwines	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
income	.0108467	.000703	15.43	0.000	.0094635	.0122298
kidhome	-.98.38089	27.17554	-3.62	0.000	-151.8456	-44.91623
recency	.4344202	.424241	1.02	0.307	-.4002238	1.269064
marital						
Married	-36.28351	38.28923	-0.95	0.344	-111.6131	39.04602
Single	-24.59594	40.62607	-0.61	0.545	-104.5229	55.33105
Widow	-22.45544	64.38409	-0.35	0.727	-149.1235	104.2126
_cons	-205.0048	59.91705	-3.42	0.001	-322.8845	-87.12506

Estimated effects

Test statistics, p -values, confidence intervals

Introduction

Why Stata

Who uses Stata

Interface
navigation

Getting data
into Stata

Data
manipulation
and
management
Exploring datasets
Exploring variables

Data
visualization

Statistical
analysis
Data analysis

Learning
resources

Introduction

Why Stata

Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

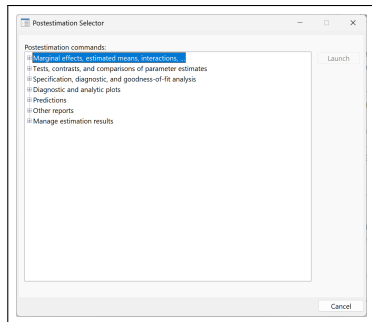
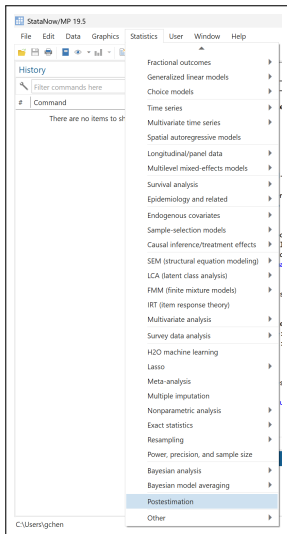
Exploring datasets
Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources



- ★ Diagnostic plots
- ★ Heteroskedasticity tests
- ★ Predictions
- ★ Parameter tests
- ★ Marginal effects
- ★ And more

Postestimation tools

Diagnostic plots — examples

Introduction

Why Stata

Who uses Stata

Interface
navigation

Getting data
into Stata

Data
manipulation
and
management

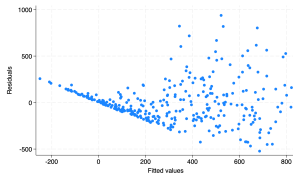
Exploring datasets
Exploring variables

Data
visualization

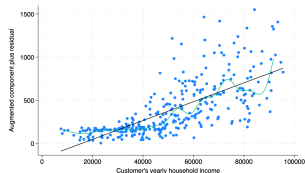
Statistical
analysis

Data analysis

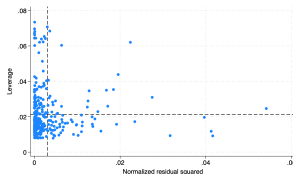
Learning
resources



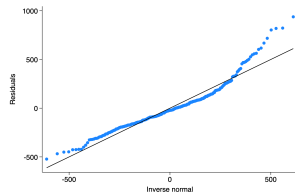
Residuals vs. fitted values plot



Augmented component-plus-residual plot



Leverage vs. squared residuals plot



Normal Q-Q plot

Introduction

[Why Stata](#)[Who uses Stata](#)Interface
navigationGetting data
into StataData
manipulation
and
management[Exploring datasets](#)[Exploring variables](#)Data
visualizationStatistical
analysis[Data analysis](#)**Learning
resources**

Learning resources

Introduction

Why Stata
Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

Exploring datasets
Exploring variables

Data visualization

Statistical analysis

Data analysis

Learning resources

- **Finding assistance (`help` vs. `search`):**
 - `help <command>` — view official command documentation.
 - `search <keywords>` — search local topics, books, FAQs, and web resources.
 - *Note:* If `help` fails to find a command, it runs `search` automatically (e.g., `help data import`).
- **Understanding the viewer window (`help logistic`):**
 - Displays concise syntax, option lists, and worked examples.
 - Quick-access links at the top provide direct shortcuts to:
 - **Dialog:** opens the corresponding GUI interface
 - **Also see / Jump to:** opens related topics, sections, or YouTube demos

Introduction

Why Stata

Who uses Stata

Interface
navigationGetting data
into StataData
manipulation
and
management

Exploring datasets

Exploring variables

Data
visualizationStatistical
analysis

Data analysis

Learning
resources

- Click the PDF link at the top of any help viewer to access the full manuals:
 - **Remarks and examples:** detailed context and empirical applications.
 - **Methods and formulas:** exact mathematical specifications.
- **Tips for manual navigation:**
 - Access the full PDF manual index via **Help > PDF documentation**.
 - Enable **Previous/Next** buttons in Adobe Reader for seamless navigation.
 - Free online access available at stata.com/features/documentation.

Introduction

- Why Stata
- Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

- Exploring datasets
- Exploring variables

Data visualization

Statistical analysis

- Data analysis

Learning resources

Using Stata Effectively: The Fundamentals

October 6–9, 2026

<https://instats.org/seminar/using-stata-effectively-the-fundamentals>

Introduction

Why Stata

Who uses Stata

Interface

navigation

Getting data

into Stata

Data

manipulation

and

management

Exploring datasets

Exploring variables

Data

visualization

Statistical

analysis

Data analysis

Learning

resources

- **Stata Press books**

<https://www.stata-press.com/>

- **Stata documentation**

<https://www.stata.com/features/documentation/>

- **Free webinars**

<https://www.stata.com/training/webinar/>

- **Stata YouTube channel**

<https://www.youtube.com/user/statacorp>

- **Web training**

<https://www.stata.com/training/web-training/>

- **Tech support**

<https://www.stata.com/support/tech-support/contact/>

Introduction

- Why Stata
- Who uses Stata

Interface navigation

Getting data into Stata

Data manipulation and management

- Exploring datasets
- Exploring variables

Data visualization

Statistical analysis

- Data analysis

Learning resources

Thank you!