



**Carleton**  
UNIVERSITY

**Canada's Capital University**

# **An Introduction to Stata**

**January 19, 2022**

**Tahreen Zahra**

**Data Services, MacOdrum Library**



# Overview

- **What is Stata?**
- **Why use Stata? Why not?**
- **First steps in Stata**
- **Where to get help**
- **Some data exercises**



# What is Stata?

- **Powerful data analysis software package for:**
  - | Statistical analysis
  - | Data management and manipulation
  - | Data visualization
- **Written for researchers**



# Why use Stata?

- **Simple user interface**
- **Compact command syntax**
- **Publication quality tables and graphics**
- **Reproducible research**
- **Comprehensive help manual**
- **Large online community**
- **Lots of different functions that are used for specific analysis**



## Some disadvantages

- **Not Open Source – Costly**
- **Limited to one dataset in memory at a time**
- **Community smaller than R**
  - | Fewer user written extensions



- **Students can get access to Stata from Information Technology Services (ITS)**

| <https://carleton.ca/its/all-services/computers/site-licensed-software/stata/>

- **Different flavours of Stata are BE, SE and MP**

| BE ≤ SE ≤ MP, regarding size of dataset allowed, number of processors used, and cost

	Max. no. of variables	Max. no. of independent variables	Max. no. of observations
<b>Stata/MP</b>	120,000	65,532	20 billion*
<b>Stata/SE</b>	32,767	10,998	2.14 billion
<b>Stata/BE</b>	2,048	798	2.14 billion

\*The maximum number of observations is limited by the amount of available RAM on your system.



ABOUT ITS +

SERVICE ALERTS

GET STARTED +

SERVICES +

SECURITY +

HELP AND SUPPORT

FINANCE &  
ADMINISTRATION

ONLINE HELP CENTRE

SUBMIT A HELP DESK  
TICKET

**Description:**

**Stata** is a powerful statistical software that enables users to analyze, manage, and produce graphical visualizations of data. It is primarily used by researchers in the fields of economics, biomedicine, and political science to examine data patterns.

**OS Supported:**

Win, Mac, Linux, Solaris

**Available to:**

Faculty, Staff and Students on/off campus use

**Cost:**

No Charge

**How to Get:**

To download STATA SE 16, please go to:

STATA SE 16

**NEW:** To download STATA SE 17, please go to:

STATA SE 17

**You must login with your MC1 credentials to gain access to either website.**

This media is not available at **Hardware Services**. If you would like to access it through VDI (Teaching/Learning) or through RCDC (Research), then please email the **ITS Service Desk** with computer name (pcab123-xx) to request the service you prefer.

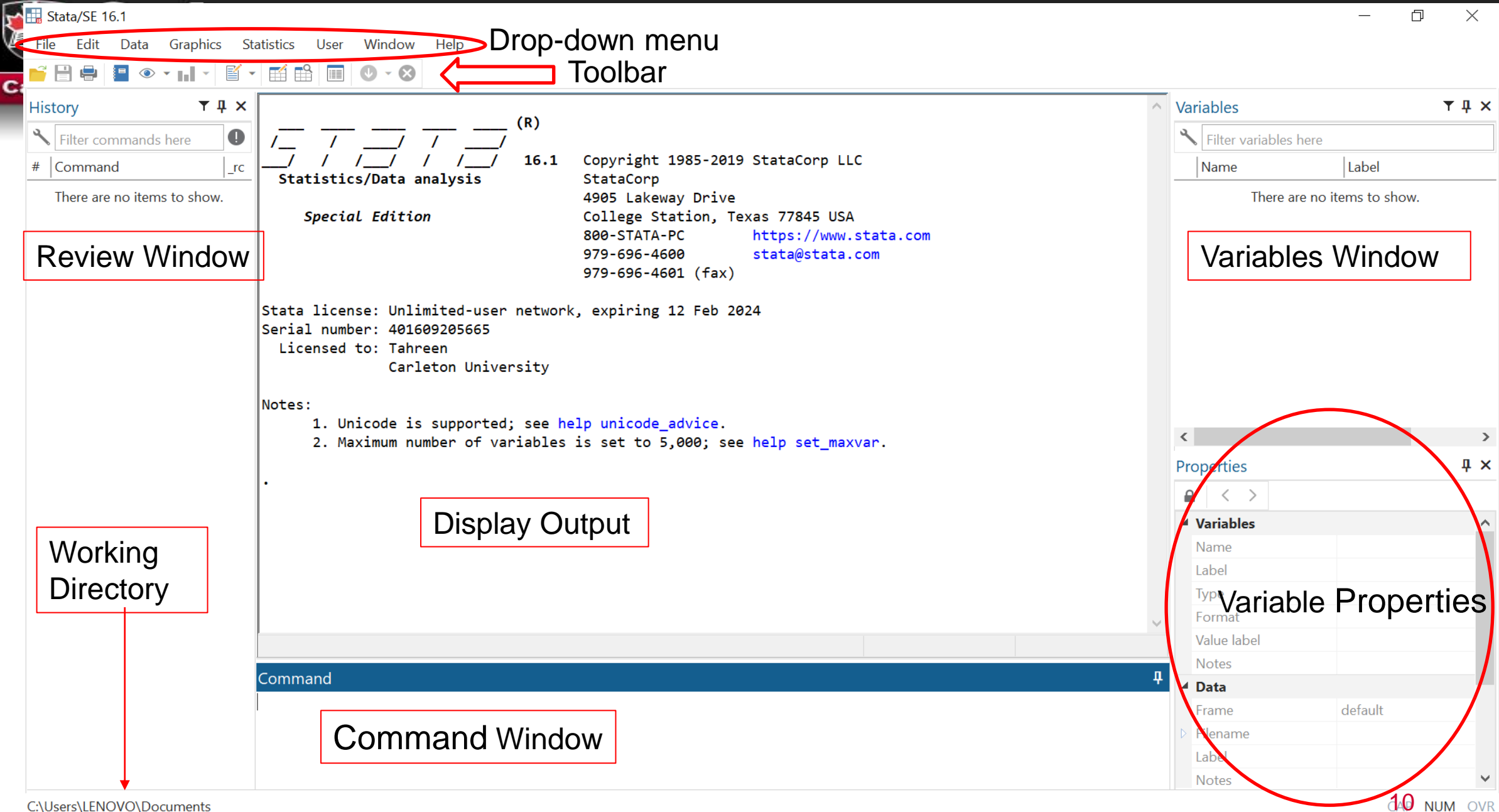


- **Google**
  - | Unknown commands: E.g. How to join datasets?
- **Stata help manual**
  - | Known commands: help regress
- **Stata forum:**  
<https://www.statalist.org/forums/>
- **Statistics consultant at MacOdrum!**
- **Books**
- **Stata command sheets**





# Navigating Stata's interface



File Edit Data Graphics Statistics User Window Help

Drop-down menu

File Edit Data Graphics Statistics User Window Help

Toolbar

History  
Filter commands here  
# Command | \_rc  
There are no items to show.

Review Window

Statistics/Data analysis 16.1  
Copyright 1985-2019 StataCorp LLC  
StataCorp  
4905 Lakeway Drive  
College Station, Texas 77845 USA  
800-STATA-PC <https://www.stata.com>  
979-696-4600 [stata@stata.com](mailto:stata@stata.com)  
979-696-4601 (fax)  
  
Stata license: Unlimited-user network, expiring 12 Feb 2024  
Serial number: 401609205665  
Licensed to: Tahreen  
Carleton University  
  
Notes:  
1. Unicode is supported; see [help unicode\\_advice](#).  
2. Maximum number of variables is set to 5,000; see [help set\\_maxvar](#).

Display Output

Working Directory

Command Window

Variables  
Filter variables here  
Name | Label  
There are no items to show.

Variables Window

Properties  
Variables  
Name | Label  
Type | Format  
Value label | Notes  
Data  
Frame | default  
Filename | Label | Notes

Variable Properties



- **Display Output:** after running a command, output or error will display
- **Command Window:** where you can enter commands directly
- **Variables Window:** variables in the dataset will be listed with their labels in the order they appear in the dataset
- **Variable Properties:**
  - | Variable section - what properties a selected variable has
  - | Data section – lists information about entire dataset
- **Review Window:** lists previously issued commands; Unsuccessful will appear in **red**
- **Working Directory:** address of working directory; can change with command `cd`
- **Drop down menu**
- **Toolbar**



# Stata drop-down menu

The screenshot shows the Stata SE 16.1 interface. The main window displays the Command window with the following text:

```
1 use "G:\Carleton M...
2 use "G:\Carleton M...
```

The Graphics menu is open, showing options such as Twoway graph (scatter, line, etc.), Bar chart, Dot chart, Pie chart, Histogram, Box plot, Contour plot, Scatterplot matrix, Distributional graphs, Smoothing and densities, Regression diagnostic plots, Time-series graphs, Panel-data line plots, Survival analysis graphs, ROC analysis, Multivariate analysis graphs, Quality control, More statistical graphs, Table of graphs, Manage graphs, and Change scheme/size.

The Variables window on the right lists the following variables:

Name	Label
REC_NUM	Order of record in file
SURVYEAR	Survey year
SURVMNTH	Survey month
LFSSTAT	Labour force status
PROV	Province
CMA	3 largest CMAs
AGE_12	5 year age group of r
AGE_6	Age in 2 and 3 year g
SEX	Sex of respondent
MARSTAT	Marital status of respo


The Properties window shows the following details for the selected variable REC\_NUM:

Property	Value
Name	REC_NUM
Label	Order of record in file
Type	long
Format	%7.0f
Value label	
Notes	

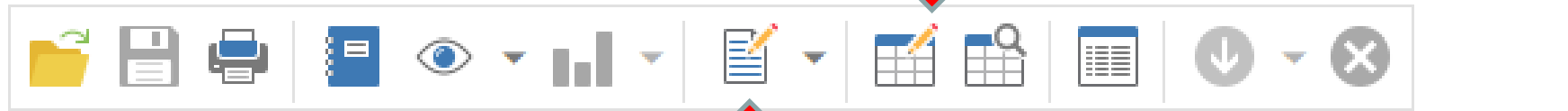
The Command window at the bottom is empty.



**Open Data Editor/ Viewer**

 Stata/SE 16.1

File   Edit   Data   Graphics   Statistics   User   Window   Help



**Open Do-file Editor**



Data Editor (Browse) - [LFS\_complete]

File Edit View Data Tools



	REC_NUM	SURVYEAR	SURVMNTH	LFSSTAT	PROV	CMA	AGE_12	AGE_6	SEX	MARSTAT	ED76to89	EDUC90	MJH	EVERWORK	FTPTLAST	
1	1	1995	1	Employe...	Newfoun...	Other C...	30 to 34	.	Male	Married	.	Some se...	Single	.	.	Pr
2	2	1995	1	Unemplo...	Quibec	Other C...	30 to 34	.	Male	Married	.	Some po...	.	Yes,wit...	Full-ti...	Pr
3	3	1995	1	Employe...	Ontario	Toronto	25 to 29	27 to 29	Female	Married	.	0 to 8 ...	Single	.	.	Pr
4	4	1995	1	Employe...	Alberta	Other C...	35 to 39	.	Female	Married	.	Bachelo...	Single	.	.	Pr
5	5	1995	1	Not in ...	Quibec	Other C...	50 to 54	.	Female	Married	.	Some se...	.	Yes,> 1...	.	.
6	6	1995	1	Employe...	Ontario	Other C...	35 to 39	.	Female	Married	.	College...	Single	.	.	Pr
7	7	1995	1	Not in ...	Ontario	Other C...	20 to 24	20 to 21	Male	Living ...	.	Some po...	.	Yes,wit...	Part-ti...	Pr
8	8	1995	1	Employe...	Alberta	Other C...	30 to 34	.	Female	Married	.	College...	Single	.	.	Pr
9	9	1995	1	Employe...	British...	Other C...	45 to 49	.	Male	Married	.	Some se...	Single	.	.	Pr
10	10	1995	1	Employe...	Manitoba	Other C...	15 to 19	15 to 16	Male	Living ...	.	Some se...	Single	.	.	Pr
11	11	1995	1	Employe...	Ontario	Toronto	20 to 24	22 to 24	Male	Living ...	.	Bachelo...	Single	.	.	Pr
12	12	1995	1	Employe...	Alberta	Other C...	40 to 44	.	Female	Married	.	Bachelo...	Single	.	.	Pr
13	13	1995	1	Employe...	Alberta	Other C...	35 to 39	.	Female	Married	.	Grade 1...	Single	.	.	Pr
14	14	1995	1	Not in ...	Quibec	Montreal	70+	.	Male	Married	.	0 to 8 ...	.	Yes,> 1...	.	.
15	15	1995	1	Employe...	Newfoun...	Other C...	40 to 44	.	Male	Married	.	Bachelo...	Single	.	.	Pr
16	16	1995	1	Not in ...	Manitoba	Other C...	60 to 64	.	Female	Married	.	Some se...	.	Yes,wit...	Full-ti...	Pr
17	17	1995	1	Employe...	British...	Other C...	25 to 29	27 to 29	Female	Married	.	Some po...	Single	.	.	Pr
18	18	1995	1	Not in ...	Nova Sc...	Other C...	40 to 44	.	Female	Married	.	Some se...	.	Yes,> 1...	.	.
19	19	1995	1	Employe...	New Bru...	Other C...	40 to 44	.	Male	Married	.	College...	Single	.	.	Pr
20	20	1995	1	Employe...	Alberta	Other C...	15 to 19	17 to 19	Female	Living ...	.	Grade 1...	Single	.	.	Pr
21	21	1995	1	Not in ...	Ontario	Toronto	70+	.	Male	Widowed	.	Grade 1...	.	Yes,> 1...	.	.
22	22	1995	1	Employe...	Quibec	Other C...	40 to 44	.	Male	Married	.	Some se...	Single	.	.	Pr
23	23	1995	1	Employe...	Manitoba	Other C...	20 to 24	22 to 24	Male	Living ...	.	Some se...	Single	.	.	Pr
24	24	1995	1	Not in ...	Ontario	Toronto	45 to 49	.	Female	Married	.	Some se...	.	Yes,> 1...	.	.
25	25	1995	1	Not in ...	Ontario	Other C...	30 to 34	.	Female	Married	.	Grade 1...	.	Yes,> 1...	.	.
26	26	1995	1	Not in ...	Manitoba	Other C...	60 to 64	.	Male	Married	.	Grade 1...	.	Yes,> 1...	.	.
27	27	1995	1	Employe...	Ontario	Other C...	35 to 39	.	Female	Married	.	Grade 1...	Single	.	.	Pr

### Variables

Filter variables here

<input checked="" type="checkbox"/>	Name	Label
<input checked="" type="checkbox"/>	REC_NUM	Order of record in file
<input checked="" type="checkbox"/>	SURVYEAR	Survey year
<input checked="" type="checkbox"/>	SURVMNTH	Survey month
<input checked="" type="checkbox"/>	LFSSTAT	Labour force status
<input checked="" type="checkbox"/>	PROV	Province
<input checked="" type="checkbox"/>	CMA	3 largest CMAs
<input checked="" type="checkbox"/>	AGE_12	5 year age group of re:
<input checked="" type="checkbox"/>	AGE_6	Ace in 2 and 3 year arc

### Properties

Variables	
Name	REC_NUM
Label	Order of record in fi
Type	long
Format	%7.0f
Value label	
Notes	
Data	
Frame	default
Filename	LFS_complete.dta
Label	
Notes	
Variables	82



# Do-file editor

The screenshot shows a window titled "Do-file Editor - Untitled\*" with a menu bar (File, Edit, View, Language, Project, Tools) and a toolbar. The editor contains the following text:

```
1
2  *Comments can be used to annotate the do-file. They make the codes easier to
3  understand*
4  *Comments are in green and are not executed*
5
6  * They are usually enclosed with * or */ ahfhkahf /*
7
8  * you can open the do-file editor with the following command*
9  doedit
10
11  *Words in quotes are coloured in red*
12
13  use "G:\Carleton Masters\Microeconometrics\Data Exercise 3\LFS_complete.dta", clear
14
15
```



- **Commands can be stored in do-files**
  - | Reproducible codes
  - | Easy debugging and changes
- **It is always a good idea to use a do-file instead of typing in command window**
- **Long lines can be broken down into multiple using `///` at the end of each line**





# Getting data into Stata

Stata/SE 16.1

File Edit Data Graphics Statistics User Window Help

Open... Ctrl+O  
Open data subset...

Save Ctrl+S  
Save as... Ctrl+Shift+S

View...  
Do...  
Filename...  
Change working directory...  
Log

**Import**

- Excel spreadsheet (\*.xls;\*.xlsx)
- Text data (delimited, \*.csv, ...)
- SPSS data (\*.sav)
- SAS data (\*.sas7bdat)
- Text data in fixed format
- Text data in fixed format with a dictionary
- Unformatted text data
- SAS XPORT Version 8 (\*.v8xpt)
- SAS XPORT Version 5 (\*.xpt)
- Federal Reserve Economic Data (FRED)
- Haver Analytics database
- ODBC data source
- dBase (\*.dbf)

Command

C:\Users\LENOVO\Documents

import excel - Import Excel files

Excel file:  
G:\Carleton Masters\ECON5880\_Data\_Science\A2\Compiled-data.xlsx Browse...

Worksheet:  
Sheet1 A1:D119

Cell range:  
A1:D119

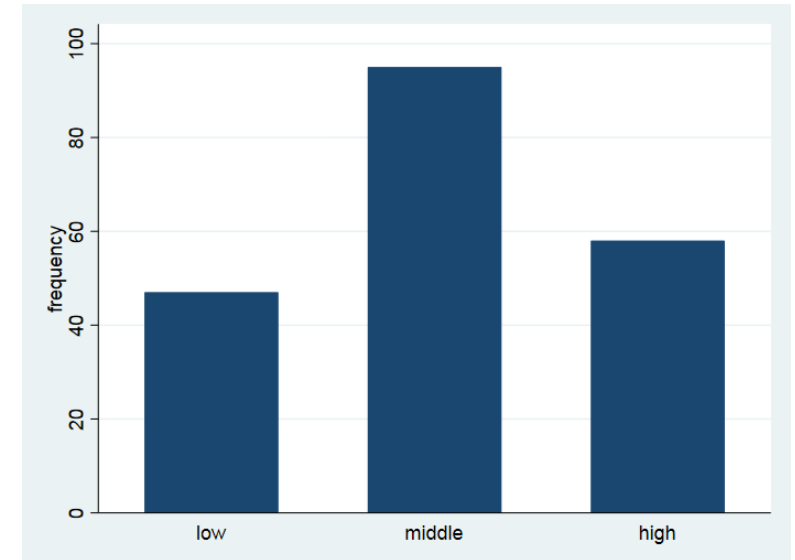
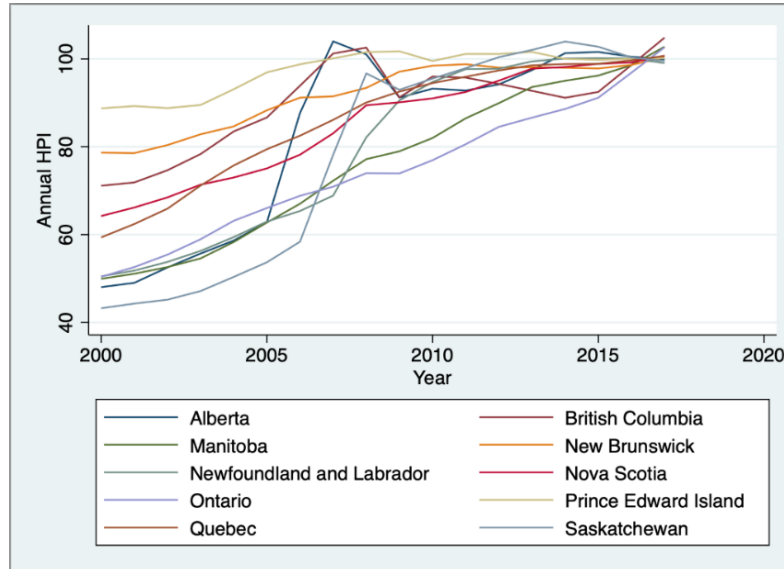
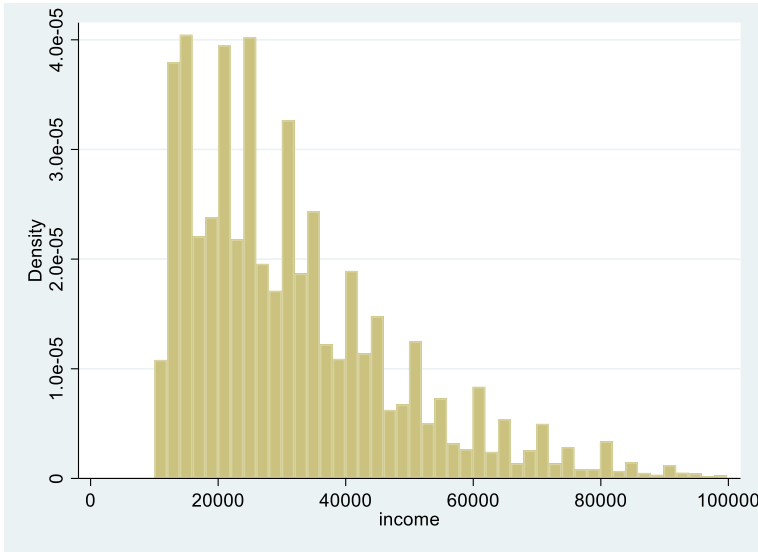
Import first row as variable names  
 Import all data as strings

Variable case:  
Preserve

Preview: (showing rows 2-51 of 119)

	Month	ActualpricesDollarsperbarre	ForecastPricesMarchbaseprofe	SP500
2	Jan-11	89.17	89.17	1282.62
3	Feb-11	88.58	88.58	1321.12
4	Mar-11	102.86	102.86	1304.49
5	Apr-11	109.53	109.53	1331.51
6	May-11	100.9	100.9	1338.31
7	Jun-11	96.26	96.26	1287.29
8	Jul-11	97.3	97.3	1325.18
9	Aug-11	86.33	86.33	1185.31

OK Cancel



tab female

female	Freq.	Percent	Cum.
0	35,374	72.23	72.23
1	13,598	27.77	100.00
<b>Total</b>	<b>48,972</b>	<b>100.00</b>	

summarize income

Variable	Obs	Mean	Std. Dev.	Min	Max
income	39,423	31058.64	16602.77	10001	99800



## Some graph commands

- **histogram income**
- **graph bar (count), over(ses)**
  - | Count-frequency; ses – categorical variable
- **twoway line y-var x-var**



## OLS Regression Model

	Monthly <sub>ns</sub>	Monthly <sub>ns</sub>	Monthly <sub>ns</sub>	Monthly <sub>ns</sub>
getsTreatment	43.60*** (1.484)	8.511*** (0.751)	7.762*** (0.764)	7.499** (0.755)
Observations	5040	5040	5040	5040
Neighbourhood FE		X	X	X
Year FE			X	X
Month FE				X

Standard errors in parentheses

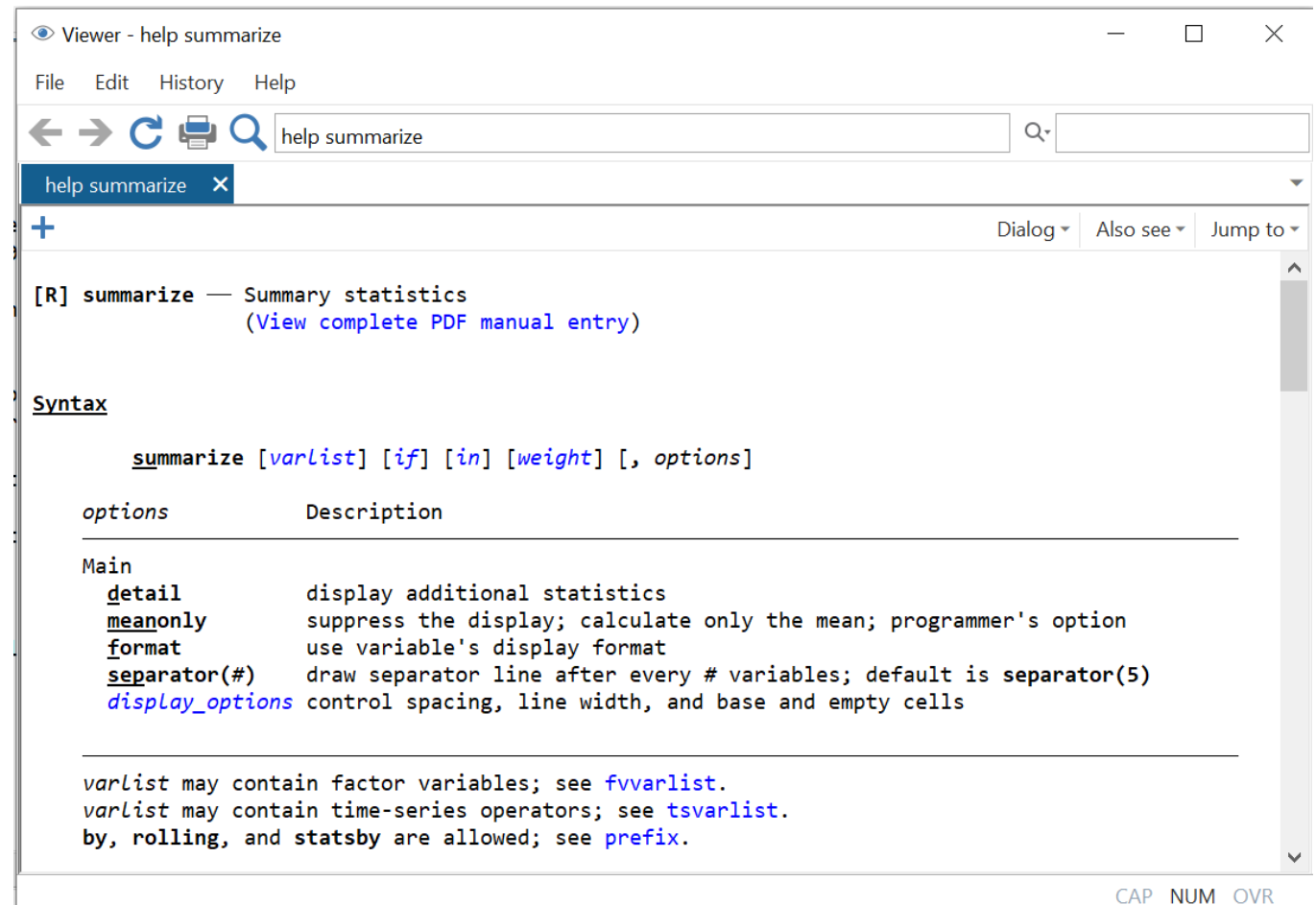
Standard errors are clustered at the neighbourhood level

\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$



# Help files

- Type *help command* in command window; e.g., `help summarize`



The screenshot shows a window titled "Viewer - help summarize" with a menu bar (File, Edit, History, Help) and a search bar containing "help summarize". The main content area displays the following text:

```
[R] summarize — Summary statistics
      (View complete PDF manual entry)
```

**Syntax**

```
summarize [varlist] [if] [in] [weight] [, options]
```

options	Description
<b>detail</b>	display additional statistics
<b>meanonly</b>	suppress the display; calculate only the mean; programmer's option
<b>format</b>	use variable's display format
<b>separator(#)</b>	draw separator line after every # variables; default is separator(5)
<b>display_options</b>	control spacing, line width, and base and empty cells

*varlist* may contain factor variables; see [fvvarlist](#).  
*varlist* may contain time-series operators; see [tsvarlist](#).  
*by*, *rolling*, and *statsby* are allowed; see [prefix](#).

CAP NUM OVR



# Managing data



# Knowing your data

- Browse dataset in data viewer with command browse
- Other commands:
  - | List – view few variables out of many; or a number of observations
  - | Codebook – inspect variable values
  - | Summarize – view the number of observations of a variable, mean, standard deviation...
  - | Tabulate – useful for categorical variables
- Missing values:
  - | Missing numeric values are represented as .
  - | Missing string values are represented by empty quotes “ ”



- **Generate – creates a variable**
  - | generate total\_cost = fixed\_cost + variable\_cost
  - | generate log\_cost = ln(total\_cost)
- **Encode – convert string variable into numeric variable**
  - | encode firm\_type
- **Label – give description to a variable**
- **Rename – change name of a variable**
  - | rename total\_cost t\_cost
- **Recode – change the values of a variable**
  - | recode female (1=0) (2=1)
- **Egen – extended variable generation functions**
  - | There are several functions!





## Some dataset operations

- **Order, sort and reshape variables in a dataset**
- **Keep or drop selected observations (rows) or variables (columns)**
- **Merge datasets – add more variables**
- **Append datasets – add more observations**

## ■ Linear regressions

| regress supply price capital

- *OLS regression of quantity supplied on prices and capital employed*

| regress supply price capital, robust

- *This OLS regression corrects for heteroskedastic error terms (uses White's standard error)*

## ■ T-tests

| ttest var

| test if means vary between two groups

| Assumes the variable is normally distributed

## ■ Logit regressions

| logit win experience talent practice

| used when outcome variable is binary

| analysis provides odds ratios



- **Stata's Mata environment**
  - | matrix programming language
- **Loops: to run a block of codes repeatedly**
  - | For loops – Runs for a number of times (10 times, 100 times, 1000 times...)
  - | While loops – Runs until a specific condition is met ( until object is at least 50 units)



**Thank you for being here!**

**For more information ...**

**<https://library.carleton.ca/services/statistical-consulting>**

**[dataservices@carleton.ca](mailto:dataservices@carleton.ca)**