Econometrics 1

Lecture 2: Basic functions of gretl

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼2613

E-mail: huangjp@szu.edu.cn

Tel: (0755) 2695 0548

Website: https://huangjp.com

Importing data

Import user defined data

- GretI can import many data file types such as CSV, ACSII (.txt), Excel (.xls, .xlsx), Stata (.dta), Eviews, SPSS, SAS, etc.

- From function menu:

```
File > Open data > User file …
```

Cross-sectional data

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Time series data

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

Pooled cross sections

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

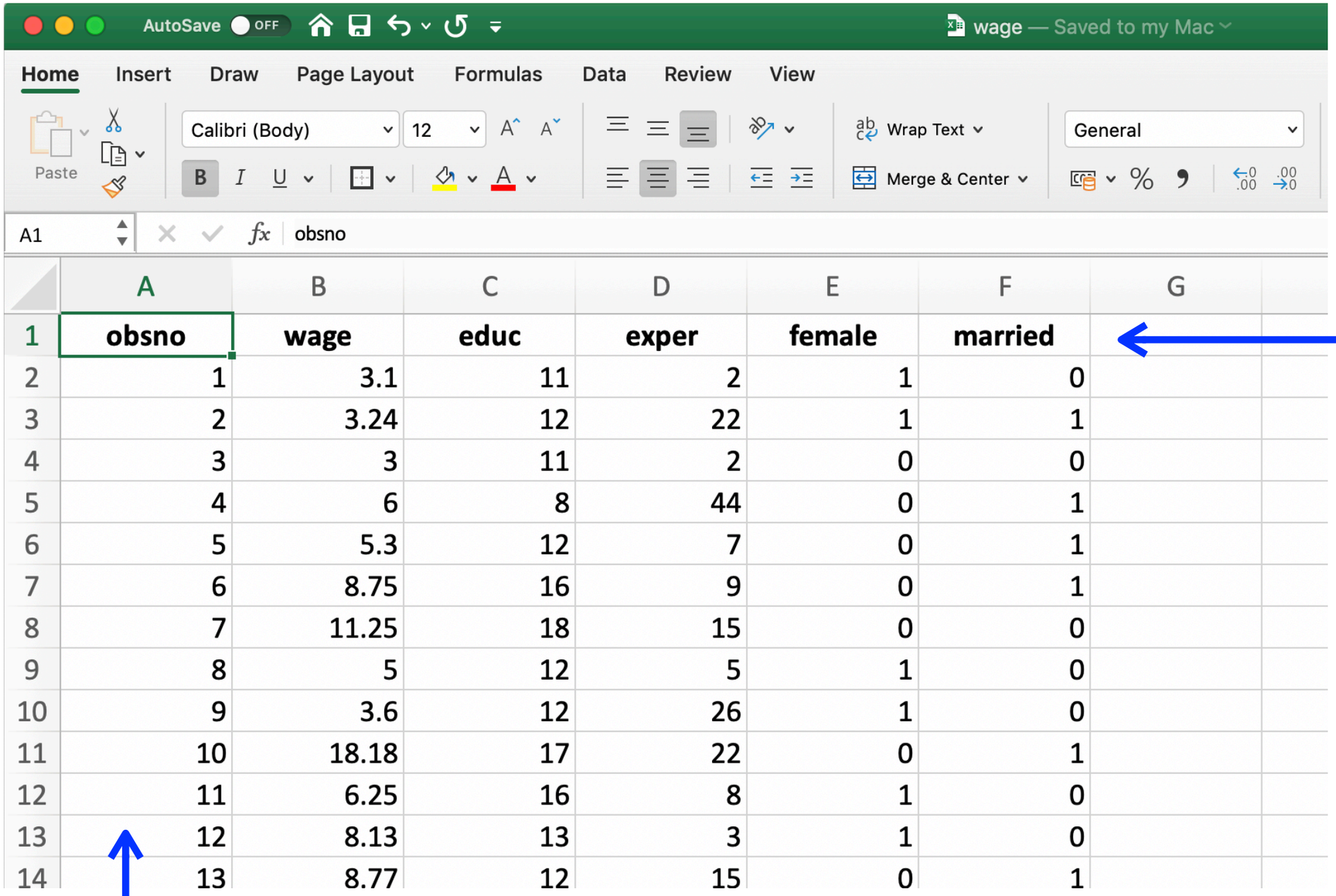
obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

Panel (longitudinal) data

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Creating a dataset with Excel



variable names

optional

Creating a dataset with Excel

- Save as am Excel file or CSV (.csv) file

Excel worksheet

G9

	A	B	C	D	E
1	Year	Make	Model	Description	Price
2	1997	Ford	E350	ac, abs, moon	3000.00
3	1999	Chevy	Venture "Extended Edition"		4900.00
4	1999	Chevy	Venture "Extended Edition, Very Large"		5000.00
5	1996	Jeep	Grand Cherokee	MUST SELL! air, moon roof, loaded	4799.00
6					

CSV file

csvexample.csv

```
Year,Make,Model,Description,Price
1997,Ford,E350,"ac, abs, moon",3000.00
1999,Chevy,"Venture ""Extended Edition""",,4900.00
1999,Chevy,"Venture ""Extended Edition, Very Large""",,5000.00
1996,Jeep,Grand Cherokee,"MUST SELL!
air, moon roof, loaded",4799.00
```

Can you see the problem?

Creating a dataset with Excel

- When you save your data as a CSV file, it is suggested to

    1. *clear format* before saving,
    2. save with unicode (UTF-8) encoding,
    3. open the saved file with any text editor to check

Date formats

- Annual data: 4-digits years

  1997

- Quarterly data: 4-digits years + a separator + 1-digit quarter

  1997.1, 2002:3, 1947Q1

- Monthly data: 4-digits years + a period or a colon + 2-digits month

  1997.01, 2002:10

Time series or panel data

- You need to (or will be asked to) specify data type

  `Data > Dataset structure …`

- For panel data, you can choose

- stacked time series,

  (this is default)

H15

	A	B	C	D	E
1	**state**	**year**	**pop**	**income**	**tax**
2	AL	1985	3973000	46015000	32.50000
3	AL	1995	4262731	83903300	40.50000
4	AR	1985	2327000	26210700	37.00000
5	AR	1995	2480121	45995500	55.50000
6	AZ	1985	3184000	43956900	31.00000
7	AZ	1995	4306908	88870500	65.33333

- or stacked cross sections

H9

	A	B	C	D	E
1	**state**	**year**	**pop**	**income**	**tax**
2	AL	1985	3973000	46015000	32.50000
3	AR	1985	2327000	26210700	37.00000
4	AZ	1985	3184000	43956900	31.00000
5	CA	1985	26444000	447103000	26.00000
6	CO	1985	3209000	49466700	31.00000

Arranging your data

Sub-sampling a data set

- “Setting” a sample — indicating the starting and ending position of the current sample range.

  `Sample > Set range …`

- “Restricting” a sample — selecting observations using logical (boolean) criterion.

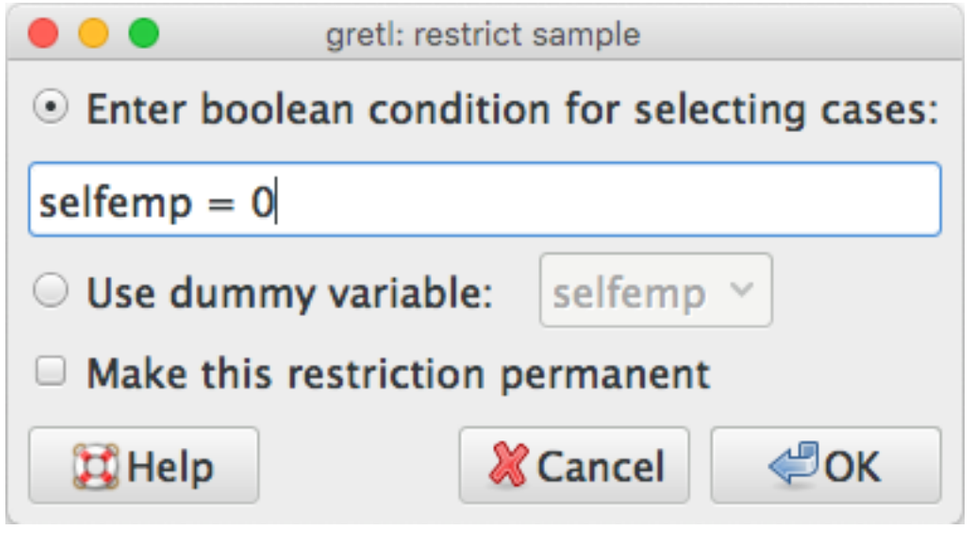
  `Sample > Restrict, based on criterion …`

- Random sampling

Restricting a sample

- Import the built-in dataset “green12_1” (Micro income and expenditure data).

- Select a sub-sample that only contains non self-employed individuals (`selfemp` = 0)



Define new variables

- Add the “log” or “square” of a variable

  `Add > Logs of selected variables …`
  `Add > Squares of selected variables …`

- Use “green12_1”, select `income` and `expend`, and add logs of the two variables.

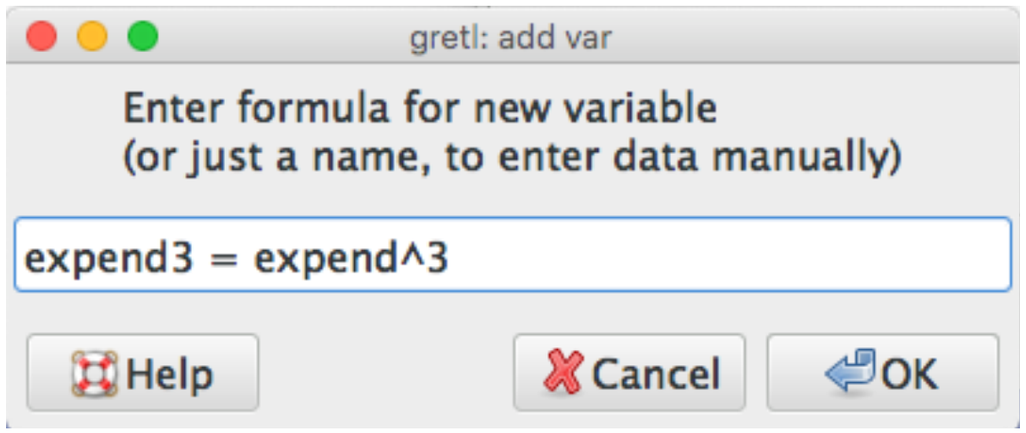
4	income	income/10000
5	expend	average monthly credit card expenditure
6	ownrent	own-rent: individual owns (1) or rents (0) home
7	selfemp	self-employed (yes = 1)
8	l_income	= log of income
9	l_expend	= log of expend

Define new variables

- You can also define a new variable based on other variables using

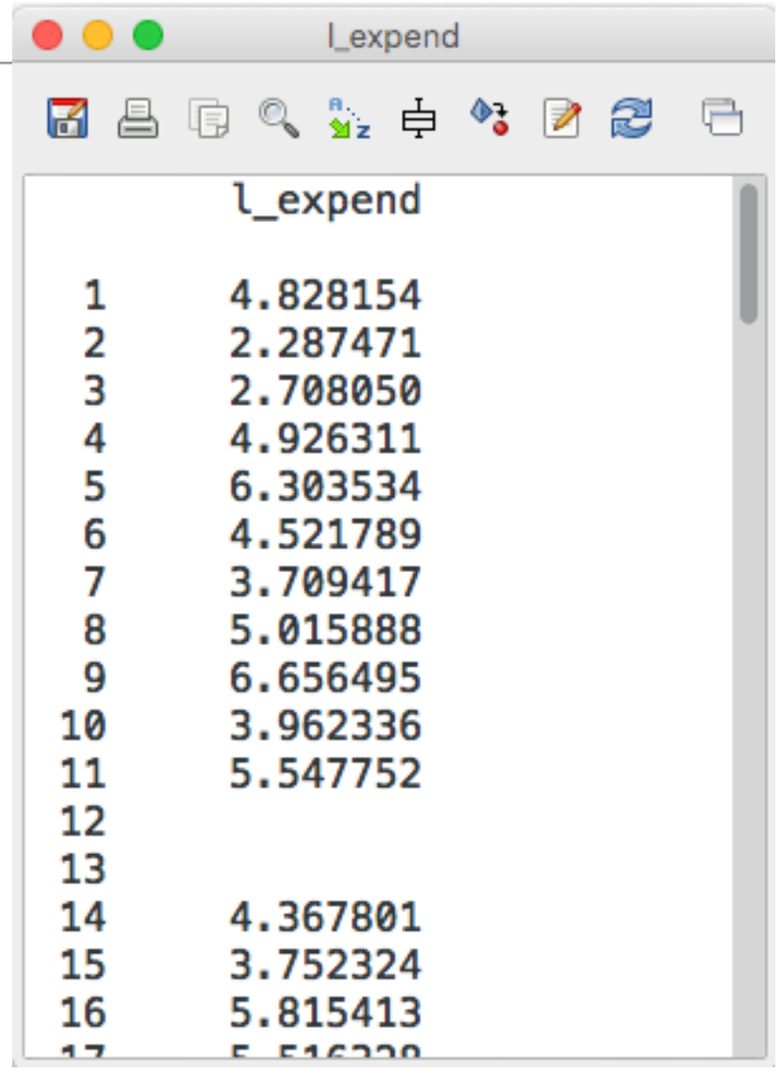
  `Add > Define new variable …`

- Use “green12_1”, define a variable as the third power of `expend`



Missing values

- When importing CSV data gretl accepts several common representations of missing values including –999, the string NA (in upper or lower case), a single dot, or simply a blank cell.



- You can other values as a missing value using

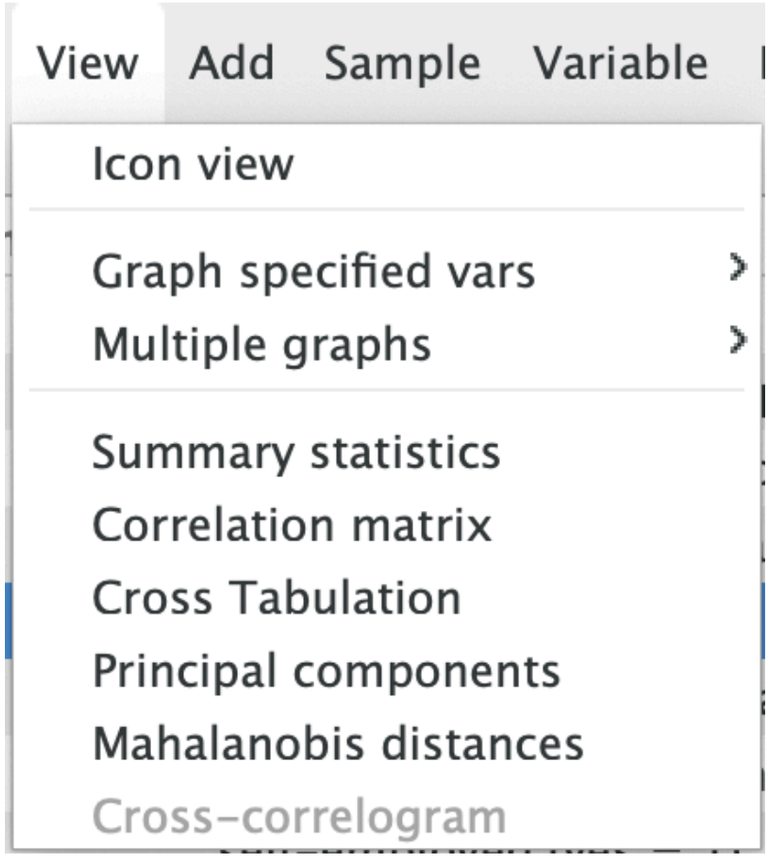
  `Variable > Set missing value code…`

- You can drop observations with missing values

Understanding your data

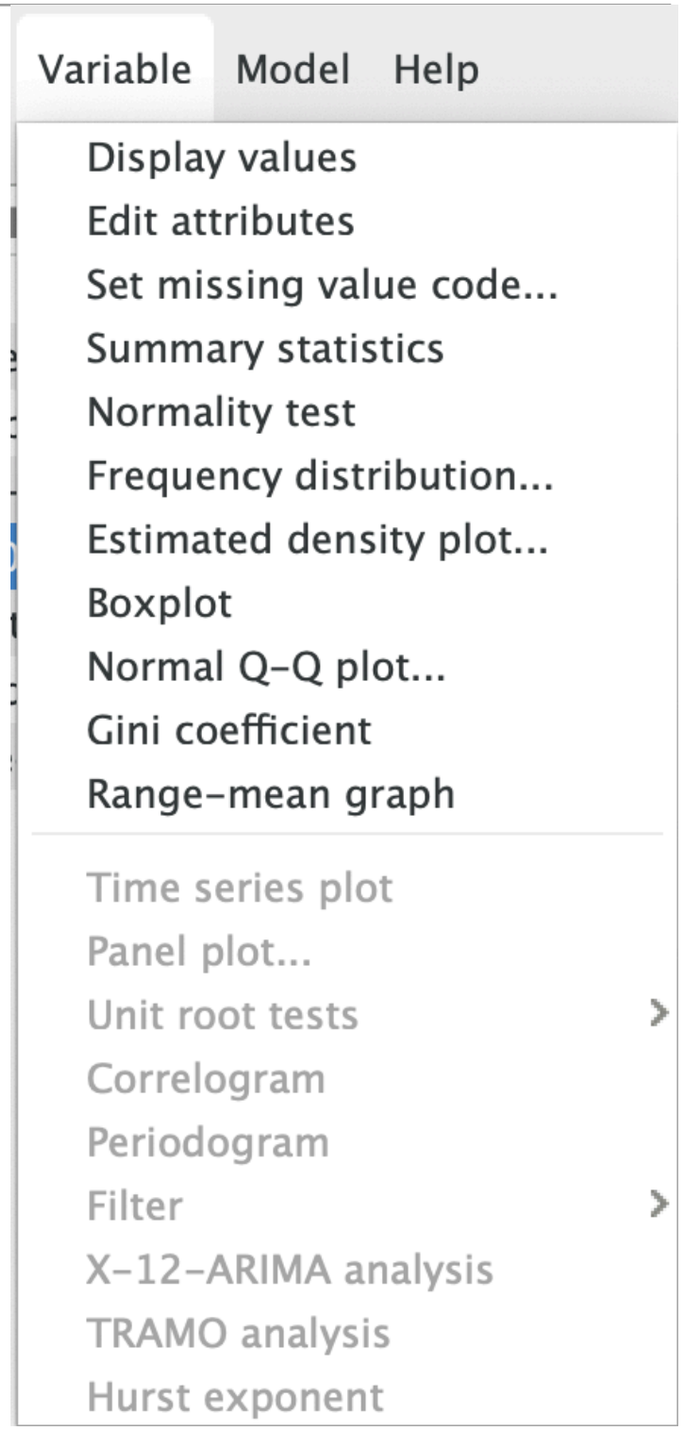
Descriptive analysis for many variables

- The “View” menu
- Summary statistics
- Correlation matrix
- Counting numbers — cross tabulation (categorical data only)
- Graphs — scatter plot, box plot, Q-Q plot, etc.



Descriptive analysis for single variable

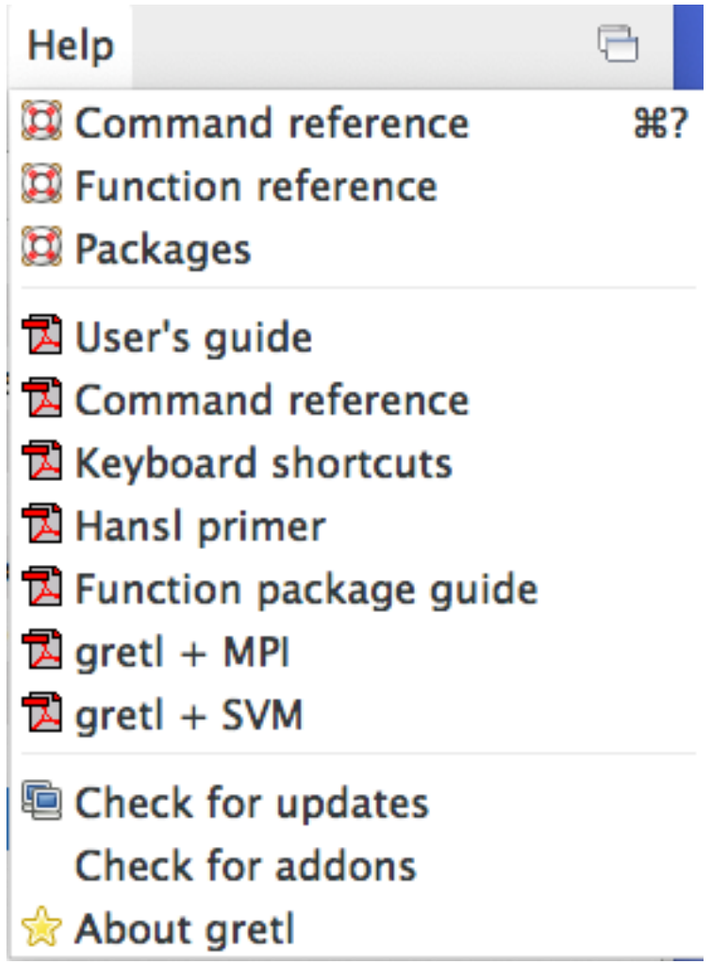
- The “Variable” menu
- Summary statistics
- Histogram — frequency distribution
- Normality check — normality test, normal Q-Q plot
- Density estimates — estimated density plot



Improving your skill

Self learning

- Make use of the `Help` menu
  - *User's guide* gives you a comprehensive introduction to gretl
  - When work in script mode you may need the *command ref.*, *function ref.*, and *Hansl primer* (Hansl is the programming language used in gretl).



The golden rule

- Trial and error!

Data sources

- Macro data
  - Statistical offices, central banks.
  - International and regional organizations:
    IMF, World Bank, OECD, WTO, U.N. Stats Division, EU, NAFTA, Asian Development Bank, etc.
- Micro data
  - US: Census Bureau, PSID, etc.
  - Survey data maintained by universities:
    IPUMS, CFPS（中国家庭追踪调查）, CHARLS（中国健康与养老追踪调查）, CHFS（中国家庭金融调查）, etc.
  - Useful links:
    北京大学开放研究数据平台 https://opendata.pku.edu.cn/
    中国人民大学中国国家调查数据库 http://www.cnsda.org/index.php

In-class practice

- Visit http://data.stats.gov.cn/index.htm

- Collect the following annual data from 1990 to 2018, and save them into a single CSV file.
  - Year, GDP (nominal), CPI (1990=100), Population, Energy consumption, Income per capita (nominal).

    If there are multiple choices, choose the most appropriate entry.

    You may need to transform the original data into the required format.

    We allow missing values.

- Import your data into gretl.