

计量经济学

第四讲：gretl 入门之数据篇

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

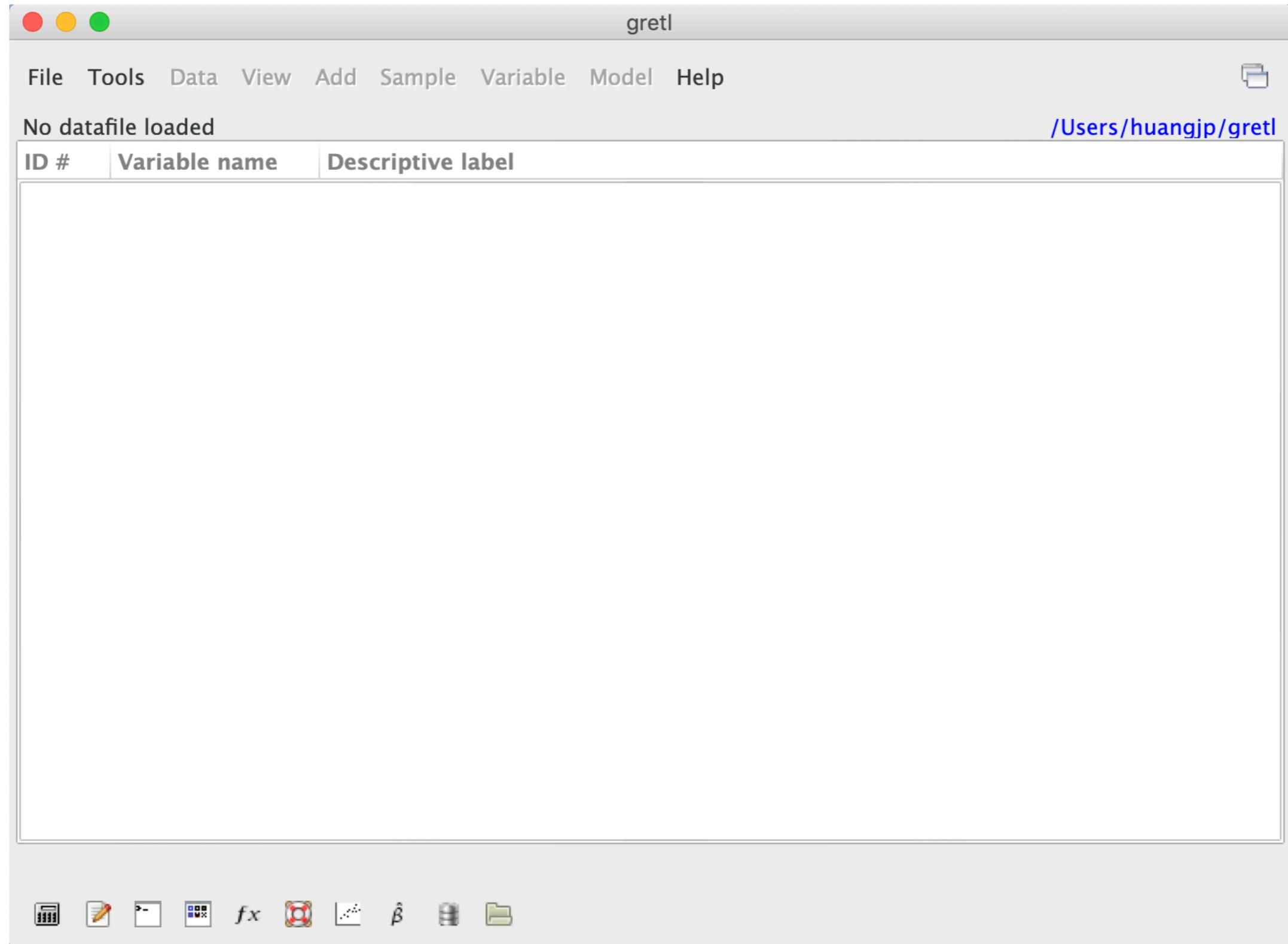
办公室 粤海校区汇文楼2613
E-mail huangjp@szu.edu.cn
Website <https://huangjp.com>

主要内容

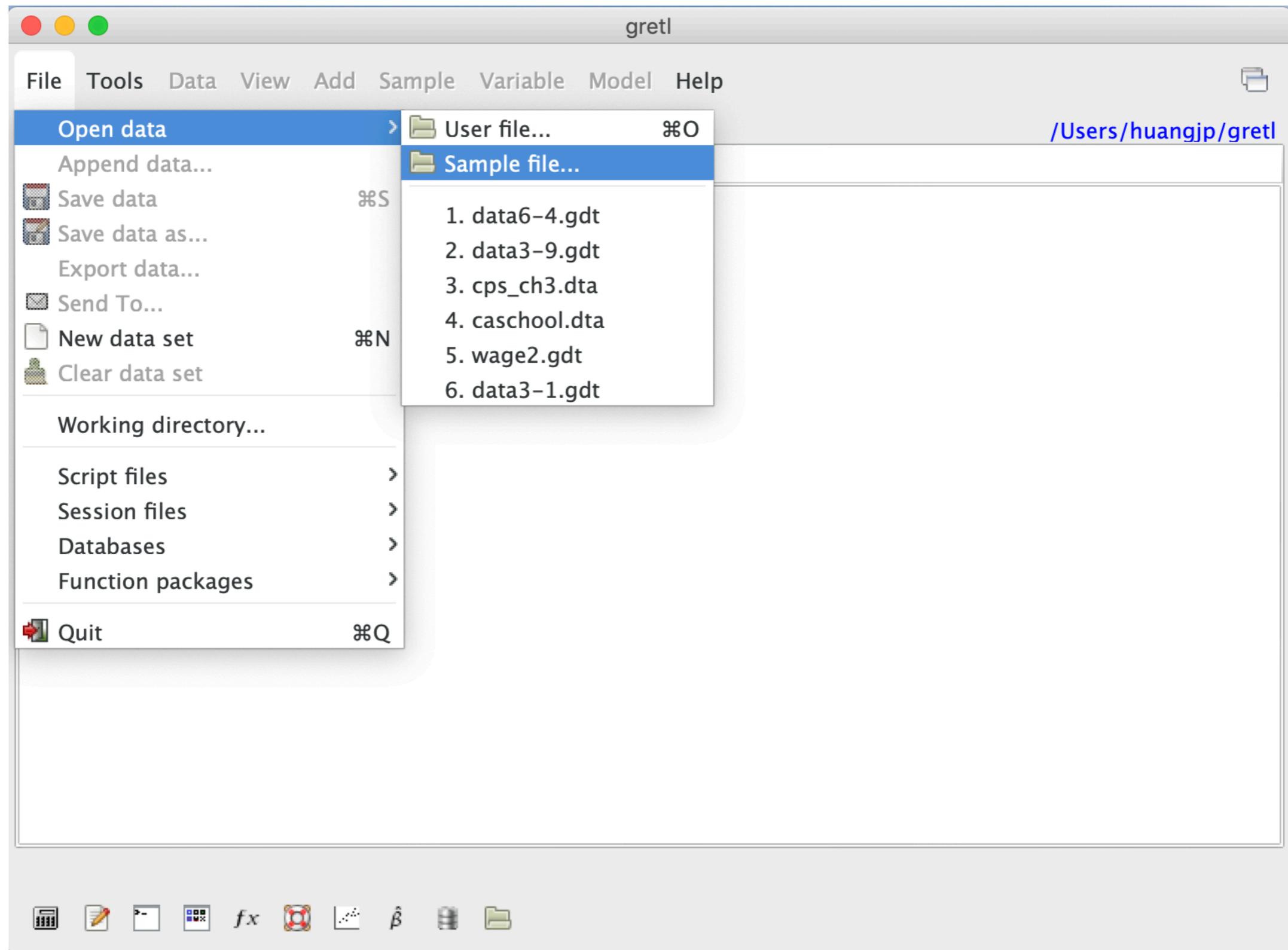
- 初次使用 gretl
- 导入外部数据
 - 数据分类
 - Excel 与 CSV 格式
- 整理数据
- 理解数据
- 如何自我提高

初次使用 gretl

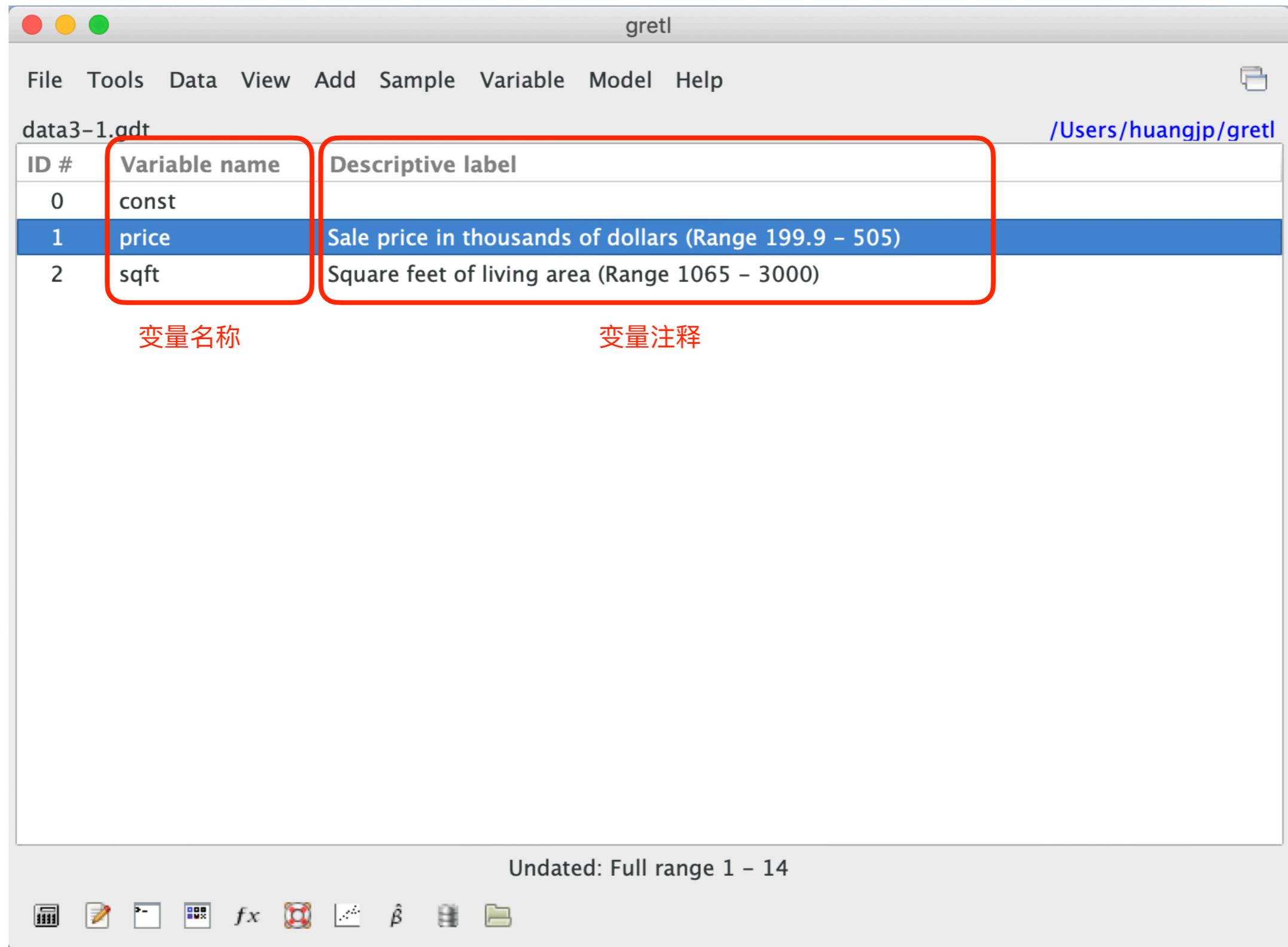
主程序窗



导入内置数据



导入数据后的主程序窗



导入外部数据

导入外部数据

- gretl 可以使用多种格式的数据，包括
 - CSV (comma separated values file, .csv)
 - ACSII (.txt)
 - Excel (.xls, .xlsx)
 - Stata (.dta)
 - Eviews, SPSS, SAS, etc.
- 导入数据路径

`File > Open data > User file ...`

数据的分类

- 截面数据 (cross-sectional data)
 - 针对不同个体在同一时期内收集到的数据。可识别个体 i 。
- 时间序列数据 (time series data)
 - 针对同一个体在多个时期内收集到的数据。可识别时间 t 。
- 面板数据 (panel data)
 - 多个个体分别在多个时期内观测到的数据。可识别 $i \times t$ 。
- 混合截面数据 (pooled cross sections)
 - 跨多个时期的截面数据，即在不同时期分别抽样。
 - 作用：1. 增加样本容量；2. 比较不同时期的变化。

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

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

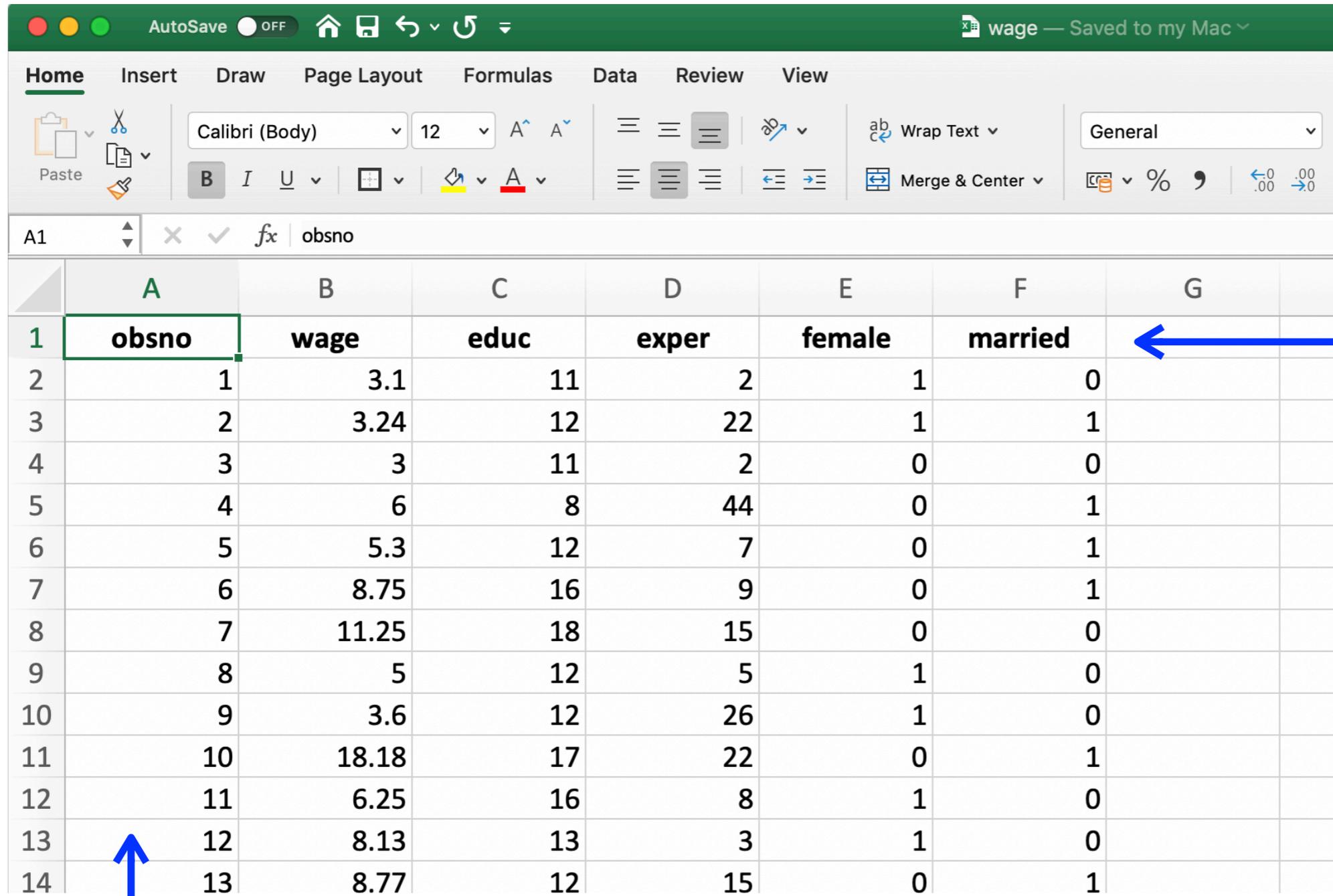
Pooled cross sections

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

利用 Excel 编辑数据

第一步：录入



The screenshot shows the Microsoft Excel interface with a data table. The table has 7 columns labeled A through G. Column A is labeled 'obsno', B is 'wage', C is 'educ', D is 'exper', E is 'female', and F is 'married'. The data rows are numbered 1 through 14. A blue arrow points to the 'obsno' column header, and another blue arrow points to the 'married' column header. The text 'optional' is written below the 'obsno' column, and 'variable names' is written to the right of the 'married' column.

	A	B	C	D	E	F	G
1	obsno	wage	educ	exper	female	married	
2	1	3.1	11	2	1	0	
3	2	3.24	12	22	1	1	
4	3	3	11	2	0	0	
5	4	6	8	44	0	1	
6	5	5.3	12	7	0	1	
7	6	8.75	16	9	0	1	
8	7	11.25	18	15	0	0	
9	8	5	12	5	1	0	
10	9	3.6	12	26	1	0	
11	10	18.18	17	22	0	1	
12	11	6.25	16	8	1	0	
13	12	8.13	13	3	1	0	
14	13	8.77	12	15	0	1	

variable names

optional

利用 Excel 编辑数据

第二步：导出

Excel worksheet

	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

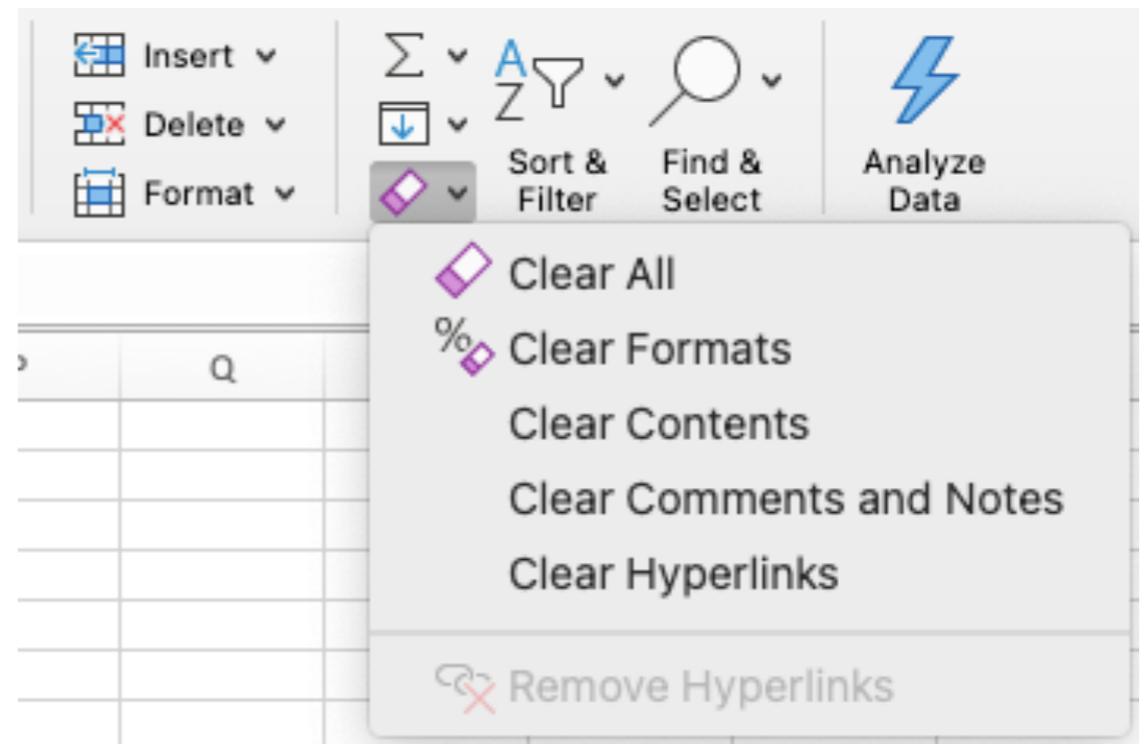
```
Year,Make,Model,Description,Price
1997,Ford,E350,"ac, abs, moon",3000.00
1999,Chevy,"Venture ""Extended Edition""",,4900.00
1999,Chevy,"Venture ""Extended Edition, Very Large""",,5000.00
1996,Jeep,Grand Cherokee,"MUST SELL!
air, moon roof, loaded",4799.00
```

在 Excel 单元格中强制换行会导致 CSV 文件出问题

利用 Excel 编辑数据

导出 CSV 文件时的建议

- 在导出前清除所有格式



- 保存为 unicode (UTF-8) 编码

File Format: CSV UTF-8 (Comma delimited) (.csv)

- 使用前用任意文本编辑软件打开 CSV 文件，检查是否有错误

用 CSV 格式保存数据时的注意事项

- 强烈建议使用英文输入。注意变换输入法，避免使用全角字符。
- 第一行应为变量名称
 - 以字母开始，只包含字母、数字、下划线。
- 数据排列方式
 - 每一列为一个变量，每一行为一个观测值，整体应为长方形。
 - 尽量使用数值数据。如原始数据中有字符类分类数据，将其转换为数字。gretl 也会在导入数据时自动将字符数据转换成数字。
- 时间数据格式
 - 年度数据：4位数字，如 1997
 - 季度数据：4位数字 + 分隔符 + 1位数字，如 1997.1, 2002:3, 1947Q1
 - 月度数据：4位数字 + 点或冒号 + 2位数字，如 1997.01, 2002:10

指定数据的种类

- 在导入 gretl 后需要指定数据种类（遵照提示或通过下面路径）

Data > Dataset structure ...

- 针对面板数据，有两种数据保存方式

- stacked time series (default)

	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

- stacked cross sections

	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

练习用数据

- wage.csv

截面数据。包含 wage, educ, exper, female, married 五个变量，其中 female 和 married 为虚拟变量。

包含变量 obsno，为个体识别编码。

- panel.csv

面板数据。包含 state, year, pop, income, tax 五个变量。其中 state 为个体名称，year 为年份。保存形式为 stacked cross sections。

整理数据

限定数据范围

Sample 菜单

- “Setting” — 通过指定数据起点和终点限定数据范围，多用于时间序列数据，其路径为

`Sample > Set range ...`

- “Restricting” — 通过设定逻辑条件限定数据，其路径为

`Sample > Restrict, based on criterion ...`

逻辑条件例：`income > 50000, year == 2010, etc.`

- 通过随机抽样选择样本的子集

定义新变量

Add 菜单

- 添加一个已有变量的对数项或二次项，其路径为

Add > Logs of selected variables ...

Add > Squares of selected variables ...

- 添加其他非线性项，其路径为

Add > Define new variable ...

定义式例： $\text{income3} = \text{income}^3$

新变量名称

新变量定义：income的三次方

缺失值

Missing values

- gretl 自动接受多种形式的缺失值，例如

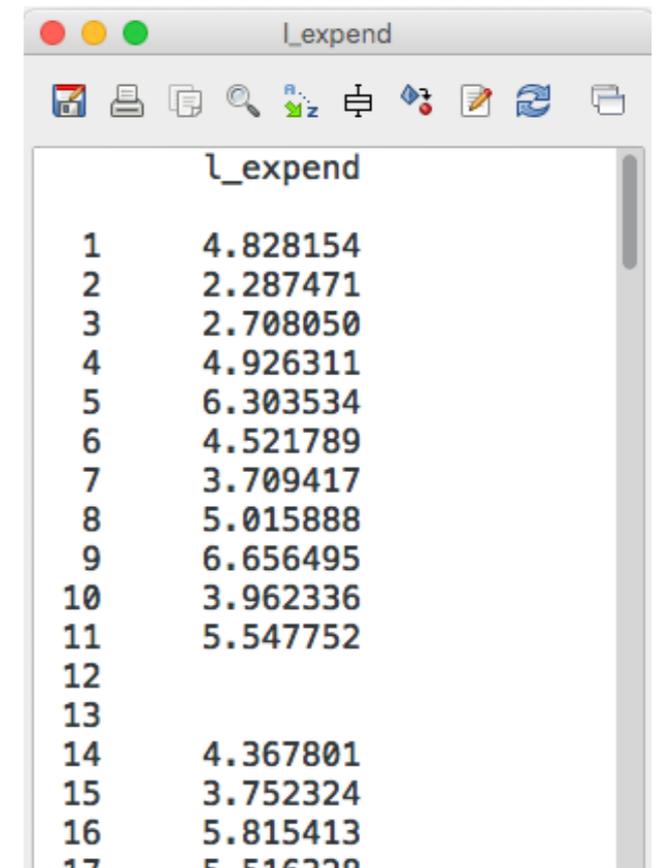
-999, NA, na, ., 或缺 (即CSV格式中两个逗号之间无其他字符)

- 也可手动指定某些字符为缺失值，其路径为

`Variable > Set missing value code...`

- 建议删除包含缺失值的观测值，其路径为

`Sample > Drop observations with missing values...`



The screenshot shows a window titled 'l_expend' with a list of 17 rows. The first column contains row numbers from 1 to 17, and the second column contains numerical values. The values are: 4.828154, 2.287471, 2.708050, 4.926311, 6.303534, 4.521789, 3.709417, 5.015888, 6.656495, 3.962336, 5.547752, 4.367801, 3.752324, 5.815413, and 5.516220.

	l_expend
1	4.828154
2	2.287471
3	2.708050
4	4.926311
5	6.303534
6	4.521789
7	3.709417
8	5.015888
9	6.656495
10	3.962336
11	5.547752
12	
13	
14	4.367801
15	3.752324
16	5.815413
17	5.516220

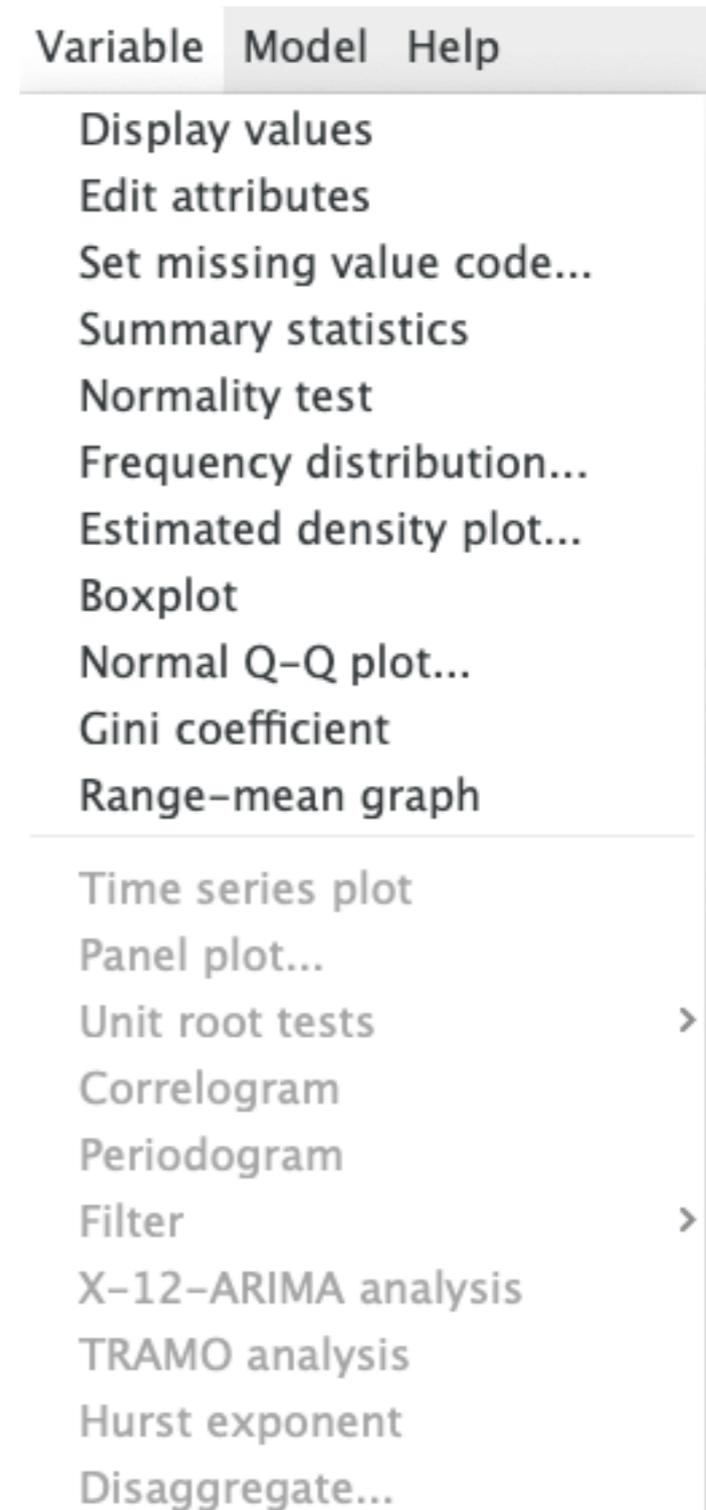
添加对数项也可以产生缺失值
(原变量去值为 0 时)

理解数据

单一变量的描述性统计

Variable 菜单

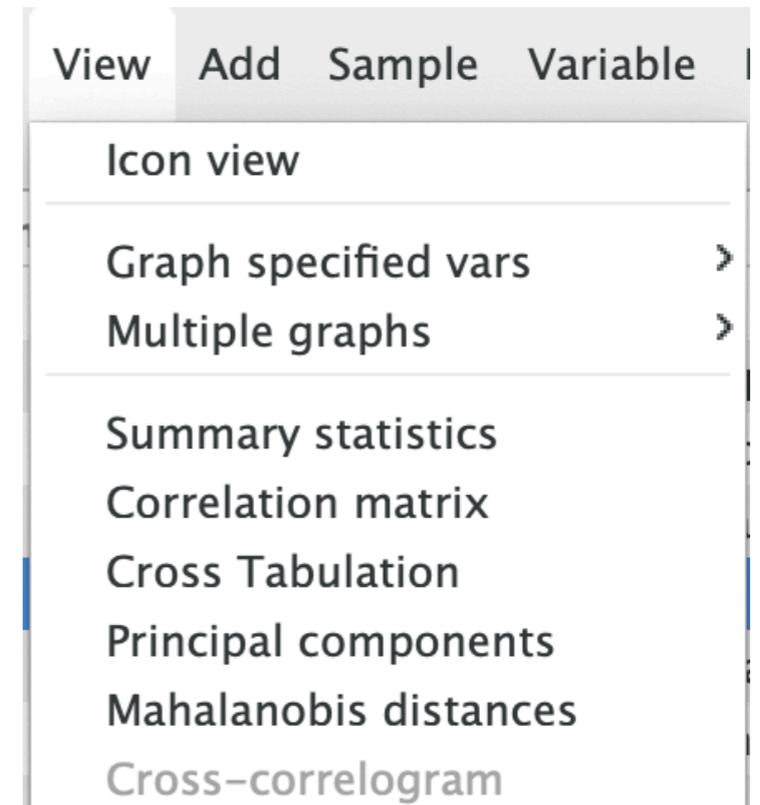
- 常用描述性统计量：
Summary statistics
- 直方图：
Frequency distribution...
- 正态性检验：
Normality test
Normal Q-Q plot...
- 密度估计：
Estimated density plot...



多个变量的描述性统计

View 菜单

- 常用描述性统计量：
`Summary statistics`
- 相关系数矩阵：
`Correlation matrix`
- 计数（分类数据）：
`Cross Tabulation`
- 画图：
`Graph specified vars >`
`Multiple graphs >`

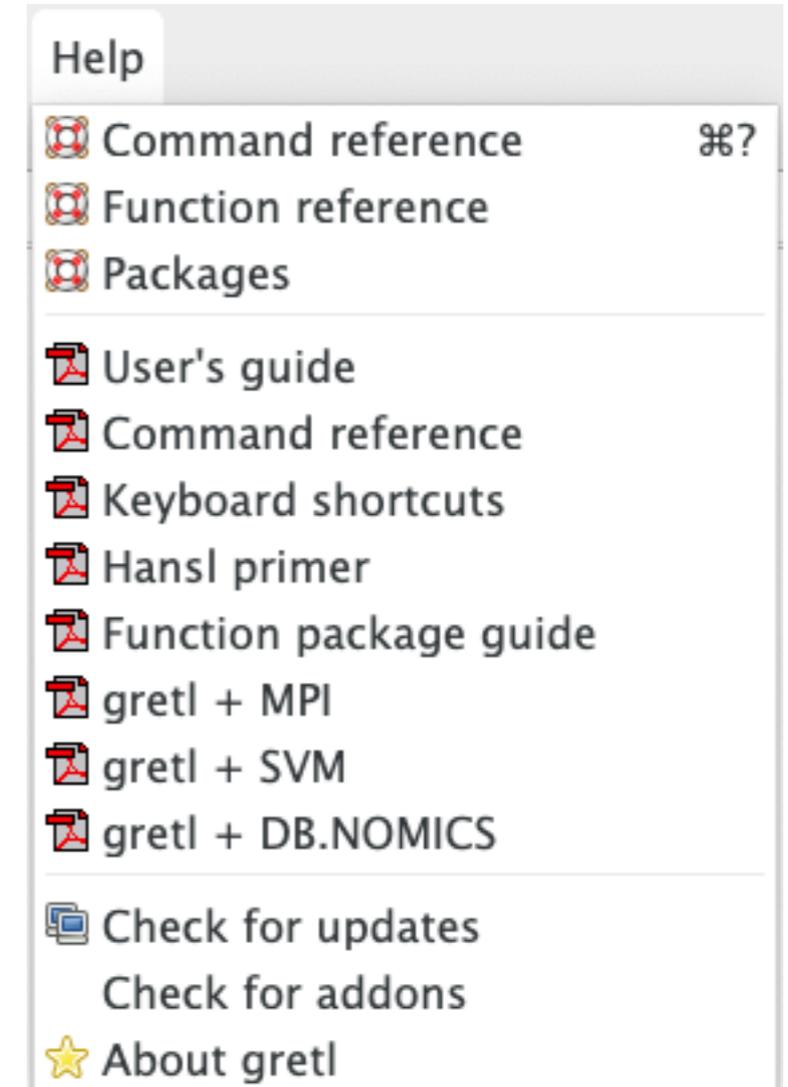


如何自我提高

利用官方资料学习

Help 菜单

- 对 gretl 的全面介绍：
User's guide
- 当你进阶到需要编程模式时：
Command reference
Function reference
Hansl primer
(Hansl 是 gretl 内嵌的编程语言)
- 不断尝试才能不断进步



数据源

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>

课后练习

课后练习（不需提交）

- 访问“国家数据”（国家统计局提供）网站：
<http://data.stats.gov.cn/index.htm>
- 找到下列变量在1990-2020年间的年度数据，并保存到同一个CSV文件中：
 - 年度、国内生产总值、居民消费价格指数（1990=100）、总人口、能源消耗总量、居民人均可支配收入
 - 注：当有多种变量可选择时，选择你认为最合适的。
有些数据可能需要整理或再计算。
允许出现缺失值。
- 将你的数据导入 gretl ，并了解其特征。