eman ta zabal zazu

Universidad del País Vasco | Euskal Herriko Unibertsitatea

CAMPUS OF INTERNATIONAL EXCELLENCE

Lesson 3

Data management in Gretl

Pilar González and Susan Orbe

Dpt. Applied Economics III (Econometrics and Statistics)

Data management in Gretl

Data are necessary to carry out any empirical econometric analysis. This lesson is going to deal with how to organise available data so that they can be used to estimate and validate econometric models with Gretl. The lesson is split into two parts:

A. Data file management in Gretl, including
   - How to create new data files in Gretl format.
   - What a Gretl session is.
   - How to modify and save Gretl data files.

B. Data analysis using Gretl, including
   - Descriptive data analysis: summary statistics (means, variances, ...), correlation matrices, normality tests, boxplots, ...
   - Graphics for data analysis (univariate and multivariate) to help understand the nature of the relationships between the variables of interest.

Learning objectives

- To generate Gretl data files loading the data from different sources.
- To modify data in Gretl.
- To define new variables using the data available in a Gretl data file.
- To simulate random samples.
- To review basic probability and inference concepts.
- To obtain the main descriptive statistics for one or more variables using Gretl.
- To graph data using both time series plots and scattergrams.

Contents

Contents

Gretl data files.

A. To create a new data file.

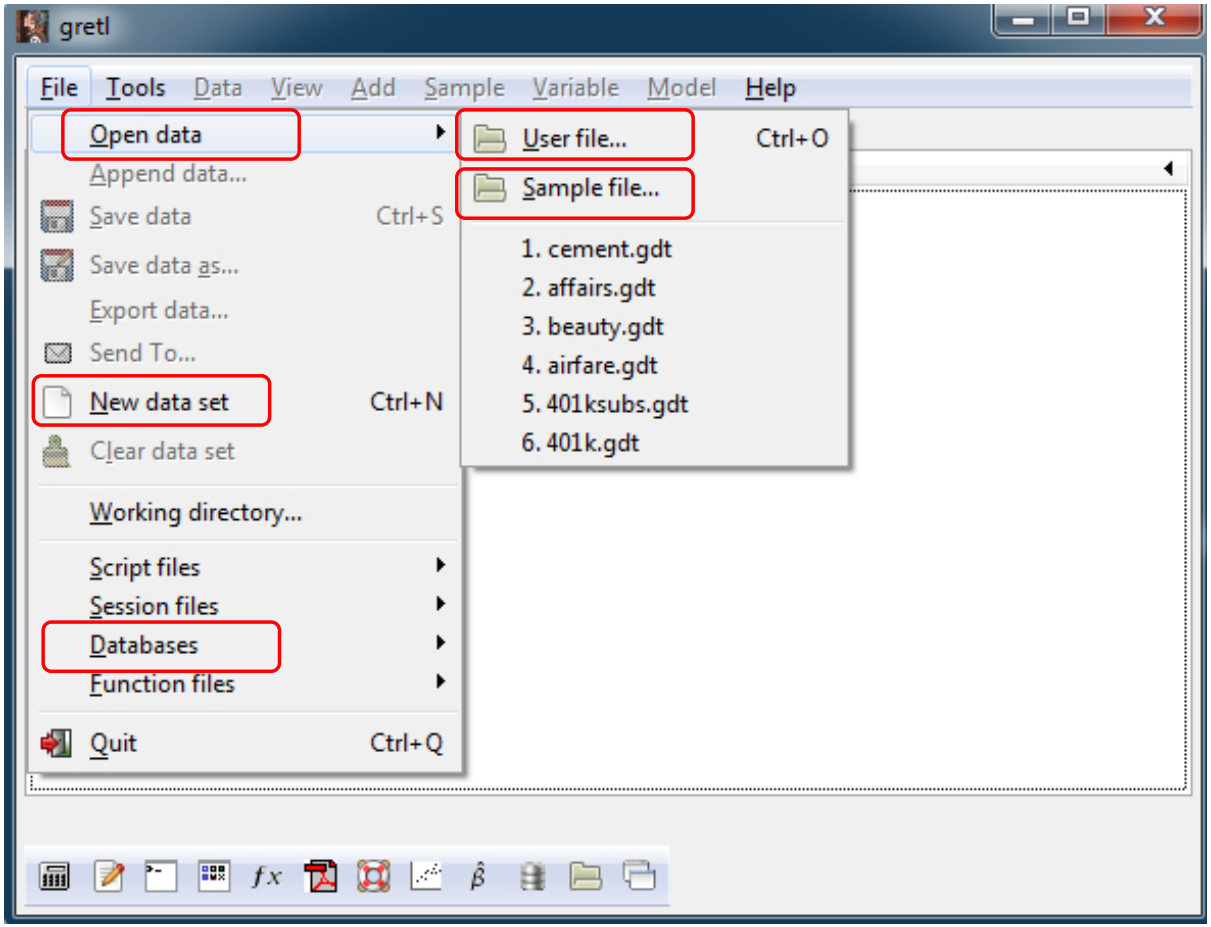
1. Entering data manually.
2. Selecting a data set from a database.

B. To load a existing data file.

1. User file:
   - Gretl file
   - Import data from other formats (Excel, Stata, EViews, CSV, ...)
2. Sample file.

Gretl data files.

The pulldown menu **File** includes all the commands related to creating and opening data files.



Gretl data files.

A. To create a new data file.

1. To enter data in Gretl manually, go up to the Menu Bar and click

   `File -> New data set`

   See **Example 3.1.1** for applications.

2. To select a time series data from a specific data base, click

   `File -> Databases`

Gretl data files.

B. To open a data file.

1. To open a user file, go up to the Menu Bar and click

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

   - You may import data in other formats, such as Excel, Stata, EViews, CSV, ...

     See **Example 3.1.2** for applications.

   - You may open a Gretl data file.

     See **Example 3.1.3** for applications.

2. To open a sample file, click

   `File -> Open data -> Sample file ...`

   See **Example 3.1.4** for applications.

Contents

What is a Gretl session?

Concept of session.

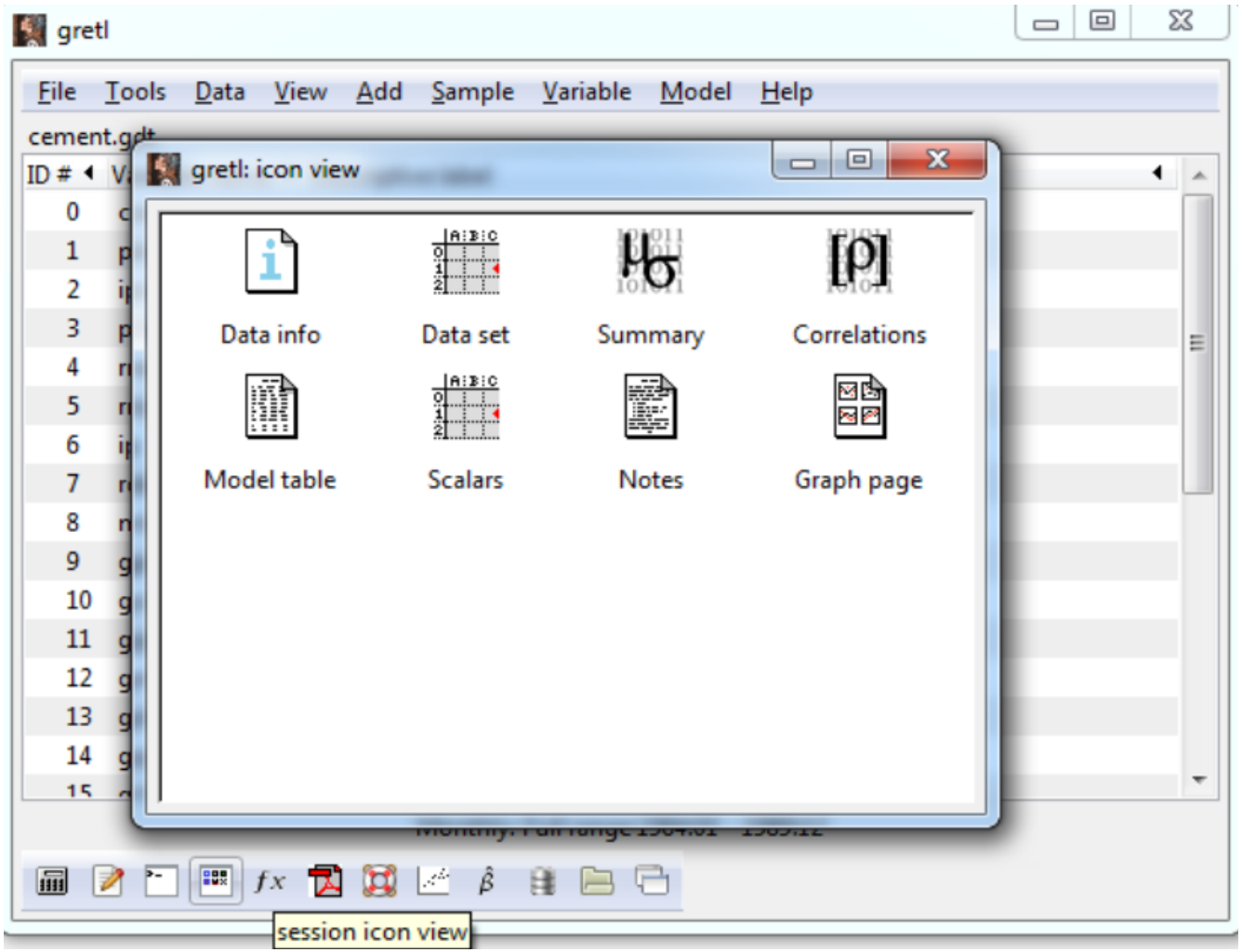
A **session** is a common space where it is possible to save data files, graphs, models and other information.

- Each object saved in a session is represented as an icon. If the session is saved, all the objects saved can be used in another working session.
- It is possible to retrieve the information saved double-clicking on any icon of the session.

Clicking on the fourth element of the Gretl's Toolbar, *session icon view*, you may see all the icons saved in the session.

What is a Gretl session?

Elements of a session.



What is a Gretl session?

Elements of the session.

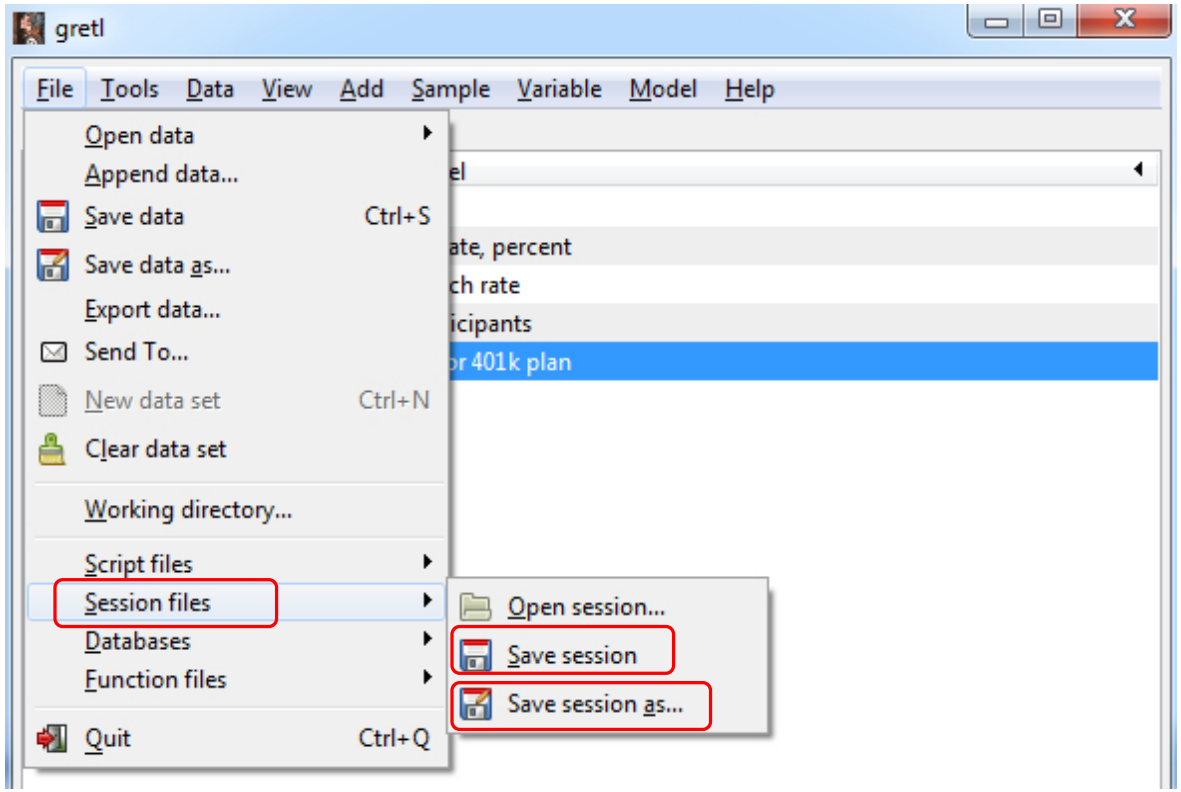
When a data set is opened, Gretl generates by default a session including the icons:

- **Data info** with information about the data set. If this icon is empty, you can fill it in and save it.
- **Data set** that allows you to edit the data.
- **Summary** that contains the main summary statistics for all the variables.
- **Correlations** that contains the correlation matrix among all the variables.
- **Model table** where it is possible to include the models estimated during the session.
- **Scalars** where it is possible to save the scalars calculated during the session.
- **Notes** where it is possible to write any comments on the work you are doing.
- **Graph page**, where it is possible to save the graphs made during the session.

What is a Gretl session?

Save a session.

If you want to save all the icons included in a Gretl session to use them for further work, DO NOT forget to save the session.

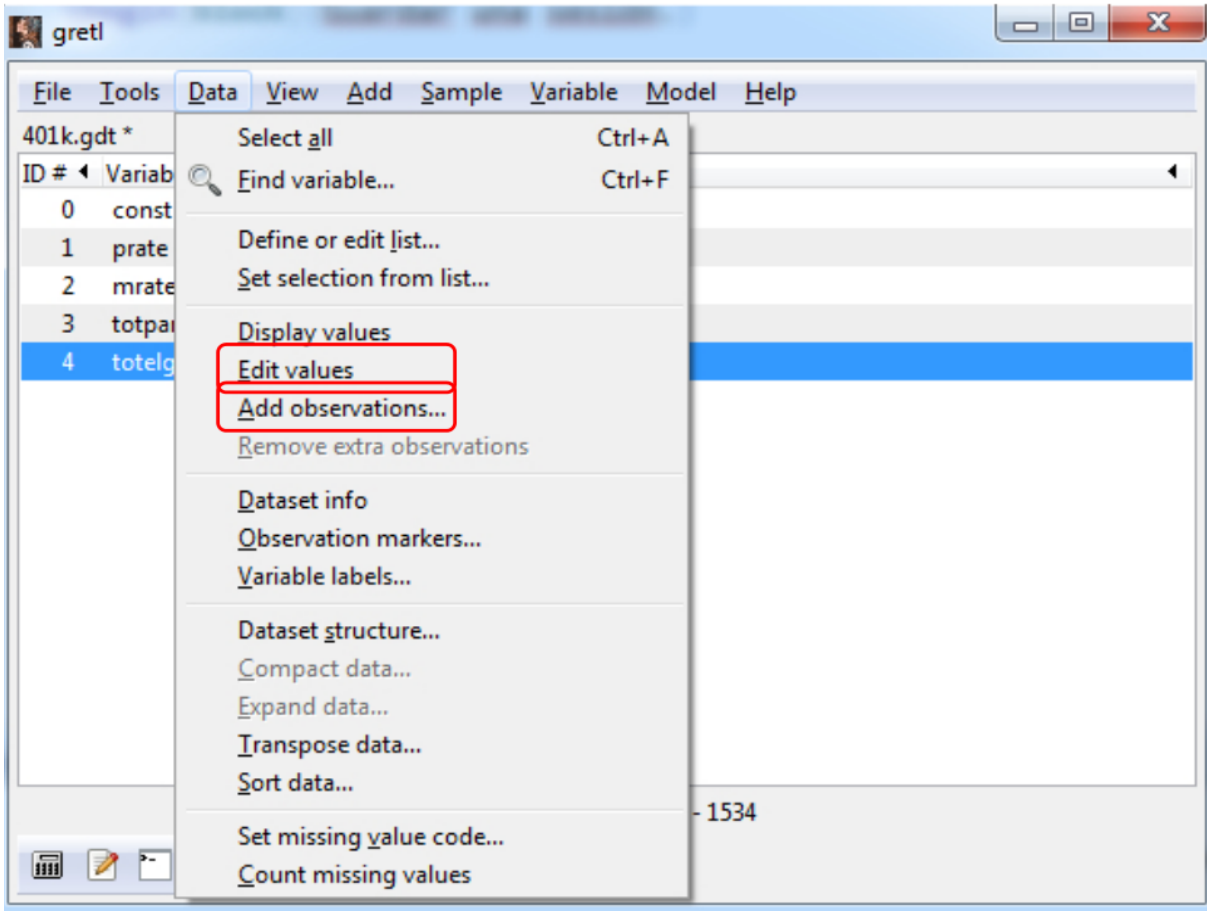


Contents

Modifying Gretl data files.

Add data.

The pulldown menu **Data** includes all the commands related to adding observations and/or variables to a Gretl data set.



Modifying Gretl data files.

Add data.

1. To enter new observations manually, go up to the Menu Bar and click

   `Data -> Add observations...`

   See **Example 3.2.1** for applications.

2. To enter new variables manually, click

   `Data -> Edit values`

   See **Example 3.2.1** for applications.

3. To add data from another file, for instance, from an Excel file, click

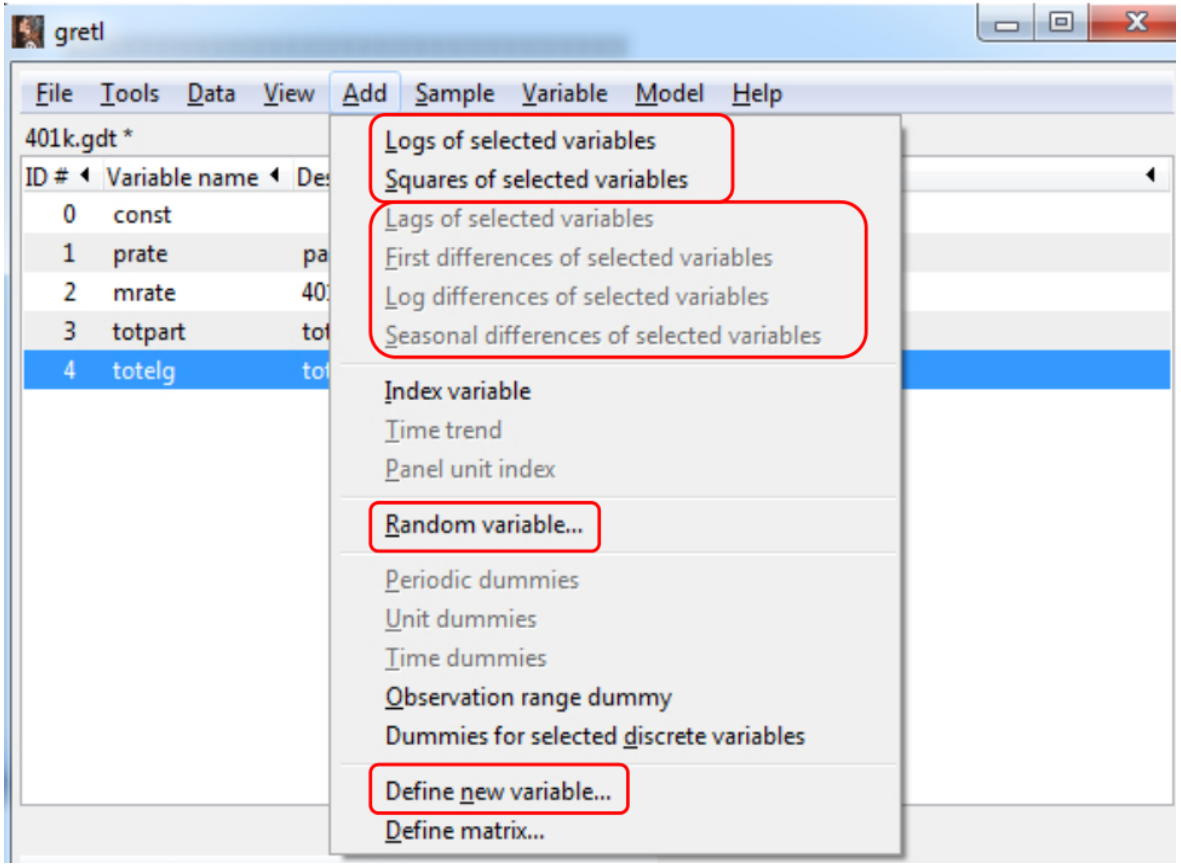
   `File -> Append data...`

   See **Example 3.2.2** for applications.

Modifying Gretl data files.

Define new variables.

All the commands related to defining new variables using the available data in the Gretl file may be found in the pulldown menu **Add**.



Modifying Gretl data files.

Define new variables.

1. Add new variables generated using the variables included in the data set.
   - Logarithms and squares of the variables in the file: $Y \rightarrow \ln Y, \quad Y^2$
   - For time series data:

Lags of a selected variable:	$Y_{t-1}, Y_{t-2}, \ldots$
First differences of a selected variable:	$\Delta Y = Y_t - Y_{t-1}$
Differences of the logarithms of a selected variable:	$\Delta \ln Y = \ln Y_t - \ln Y_{t-1}$
Seasonal differences of a selected variable: (only for quarterly or monthly time series data)	$\Delta_s Y = Y_t - Y_{t-s}$

See **Example 3.2.3** for applications.

Modifying Gretl data files.

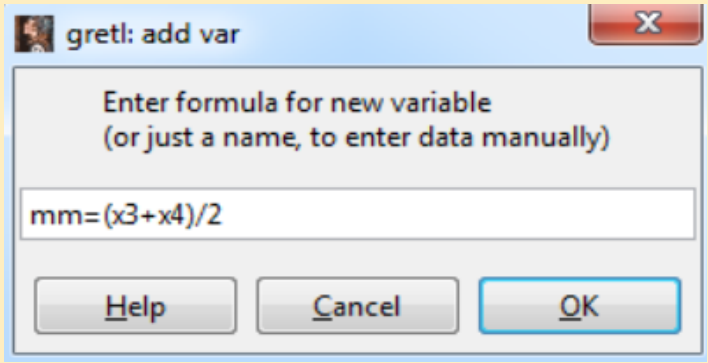
Define new variables.

1. Add new variables generated using the variables included in the data set.

   New variables can be generated as transformations (sums, products, ...) of the variables included in the data set. Go up to the Menu Bar, click

   `Add -> Define new variable...`

   and write the expression for the new variable in the dialog box. For instance, to obtain a new variable which is an average of variables $X3$ and $X4$ write



   See **Example 3.2.3** for applications.

Modifying Gretl data files.

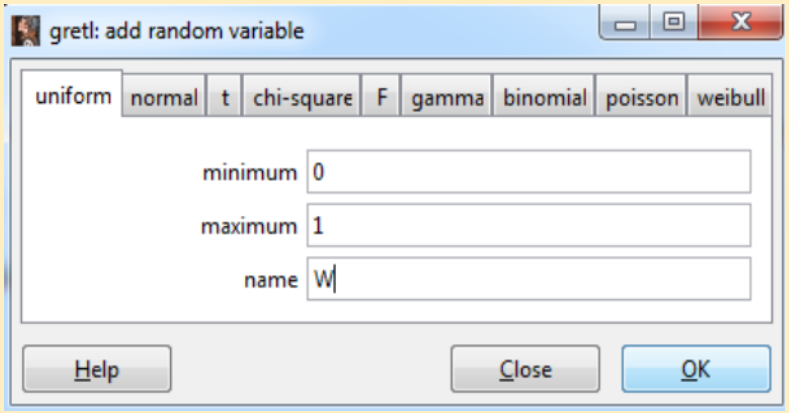
Define new variables.

2. Simulate variables.

   It is possible to generate a new variable by simulating a random sample from one of these distributions: uniform, normal, t, χ^2, F, gamma, binomial, Poisson y Weibull. Go up to the Menu Bar, click

   `Add -> Random variable...`

   and select the distribution, its parameters and a name for the new variable in the dialog box.



   See **Example 3.2.4** for applications.

Contents

Descriptive statistics.

Some basic statistics concepts.

Random variable versus observation.

In Econometrics, the economic variables are considered as random variables, because their actual values are not known until they are observed.
An observation is one of the possible outcomes of the random variable.

It is necessary to distinguish between variables whose values are not yet observed (random variables) and those whose values have already been observed (observations).

It is important to keep in mind always that an observation is one of the many possible values that a variable can take. Another draw from the random variable will usually result in a different value being observed.

Random variables can be:

1. Discrete, taking any of a specified finite or countable list of values. For instance, the number of kids.
2. Continuous, taking any numerical value in an interval or collection of intervals. For instance, consumption, income, production, ...

Descriptive statistics.

Some basic statistics concepts.

Probability distribution.

The probability distribution gives the relative frequency (or probability) with which each possible value of the random variable is observed.

The probability distribution function of a random variable X, or simply distribution function, is the mathematical function that assigns a probability to each measurable subset of the possible outcomes of a random variable:

$$F_X(x) = P(X \leq x)$$

The density probability function (or, simply, density function) represented as $f(x)$, is the derivative of the probability distribution function.

Descriptive statistics.

Some basic statistics concepts.

The probability function contains all the information about the random variable. But, its mathematical form is usually quite complicated and it is common to summarize this information by concentrating on some simple numerical characteristics of the probability functions referred as parameters. For instance, the mean and the variance.

Mean. The expected value of a random value X is the sum of the product of the probability of each event $((p(.), f(.))$ times the value of the event (x_i, x).

$$\mu = E(X) = \begin{cases} \text{Discrete variable} & \sum_{i=1}^{n} x_i p(x_i) \\ \text{Continuous variable} & \int_{-\infty}^{\infty} x f(x) \end{cases}$$

Variance is a measure of dispersion of a random variable X around the expected value $E(X)$.

$$V(X) = E[X - E(X)]^2 = \sigma_X^2$$

Descriptive statistics.

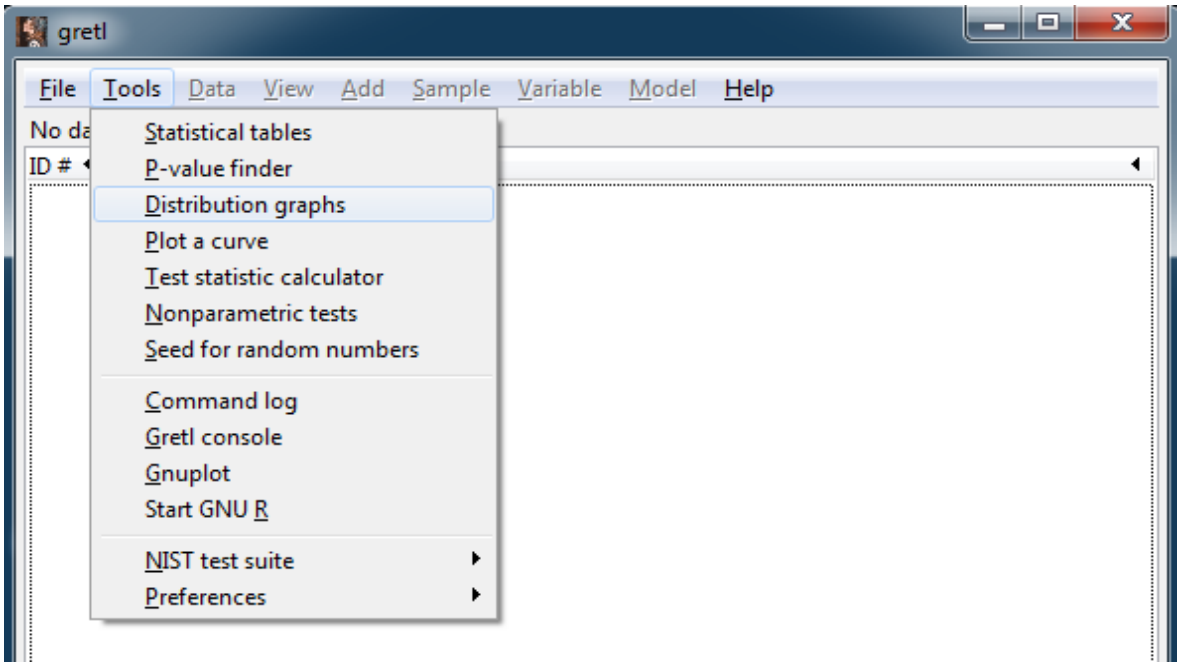
Some basic statistics concepts.

Distribution graphs in Gretl.

Go up to the Menu Bar and click

`Tools -> Distribution graphs`

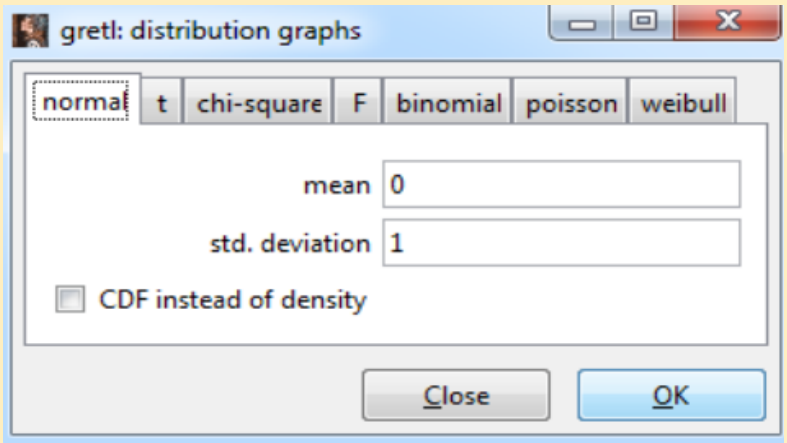
to represent the most common probability distributions.



Some basic statistics concepts.

Distribution graphs in Gretl.

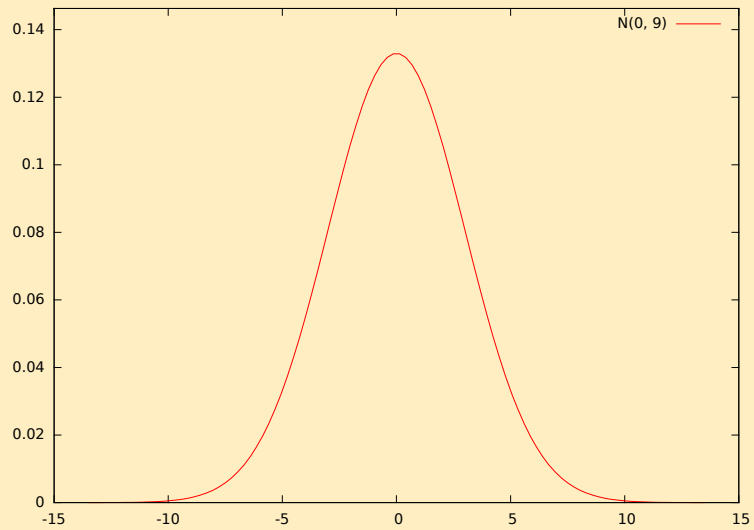
- It is necessary to select the distribution and provide its parameters.
- The density function is represented by default, but it is possible to graph the probability density function for the normal distribution exclusively by marking CDF instead of density.

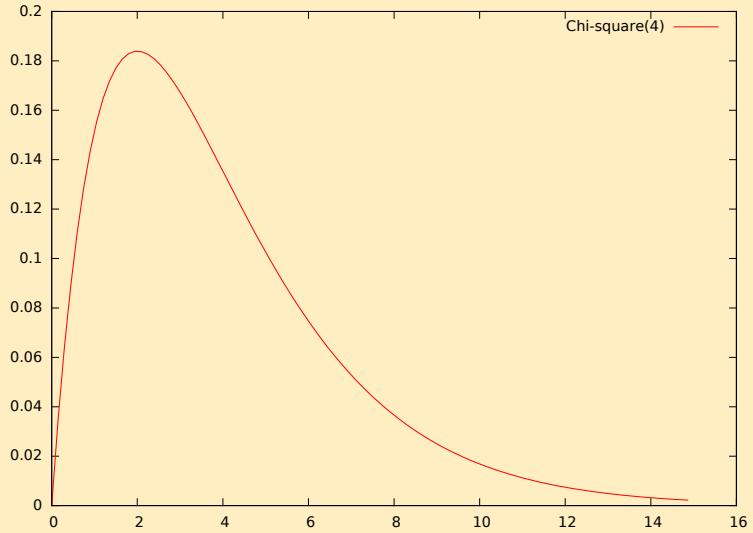


Some basic statistics concepts.

Distribution graphs in Gretl.

The left-hand side graph shows the density function of a normal random variable with mean 0 and variance 9. The right-hand side graph shows the density function of a χ^2 distribution with four degrees of freedom.





Descriptive statistics.

Some basic statistics concepts.

Consider two random variables X and Y that follow a joint distribution $f(X, Y)$.

Covariance. It is a measure of the linear association between two variables. If the covariance is positive, there is a direct linear relationship between the variables, and if the covariance is negative, there is an inverse linear relationship between the variables. If there is no linear relationship, the covariance is zero.

$$cov(X, Y) = E[X - E(X)][Y - E(Y)] = \sigma_{XY}$$

Correlation. It is a measure of the linear association between two variables. It takes values between -1 and 1.

$$corr(X, Y) = \frac{cov(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \rho_{XY}$$

Descriptive statistics.

Some basic statistics concepts.

If the distribution function of a random variable is known, it is possible to obtain the mean and the variance. But, usually, both the values of X and its distribution function are unknown.

Population: set of individuals relevant to the study.

Sample: subset of the population.

Characteristics of the distribution function. Population parameters of interest:

mean, variance, covariance, correlation.

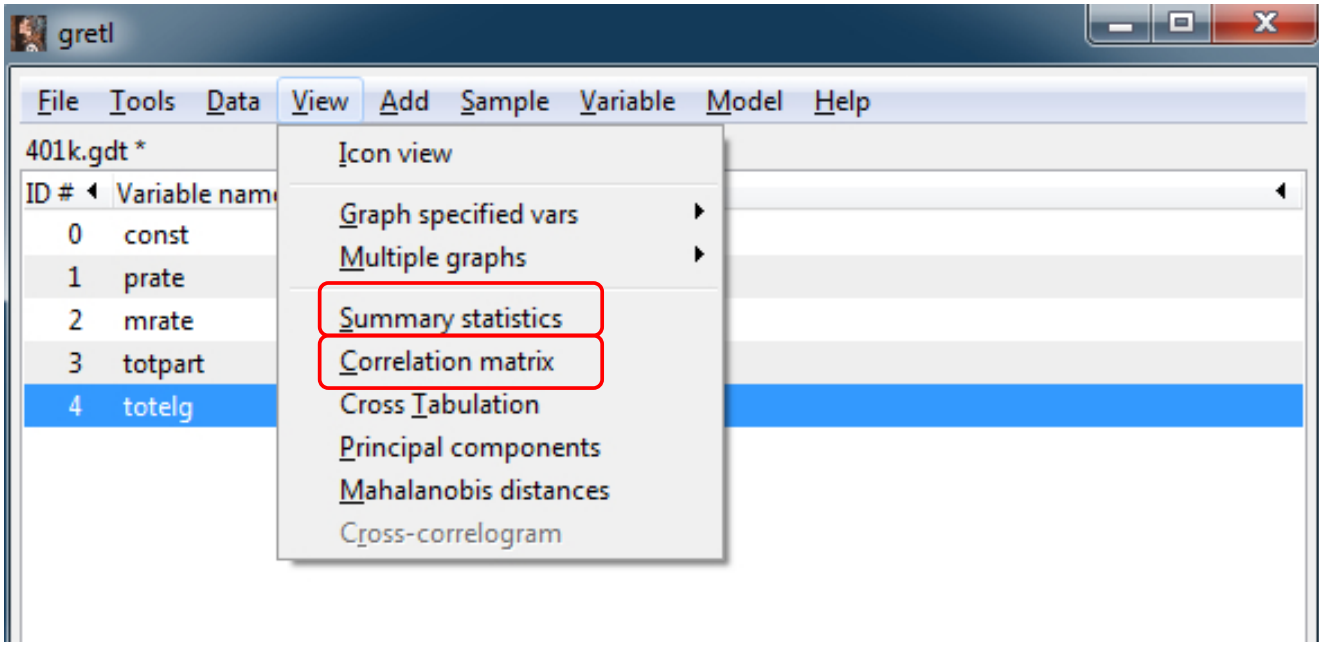
Estimation of the population parameters using the sample:

sample mean, sample variance, sample covariance, sample correlation.

Descriptive statistics.

Summary statistics in Gretl.

The pulldown menu **View** includes all the commands necessary to estimate the main parameters of the distribution function.



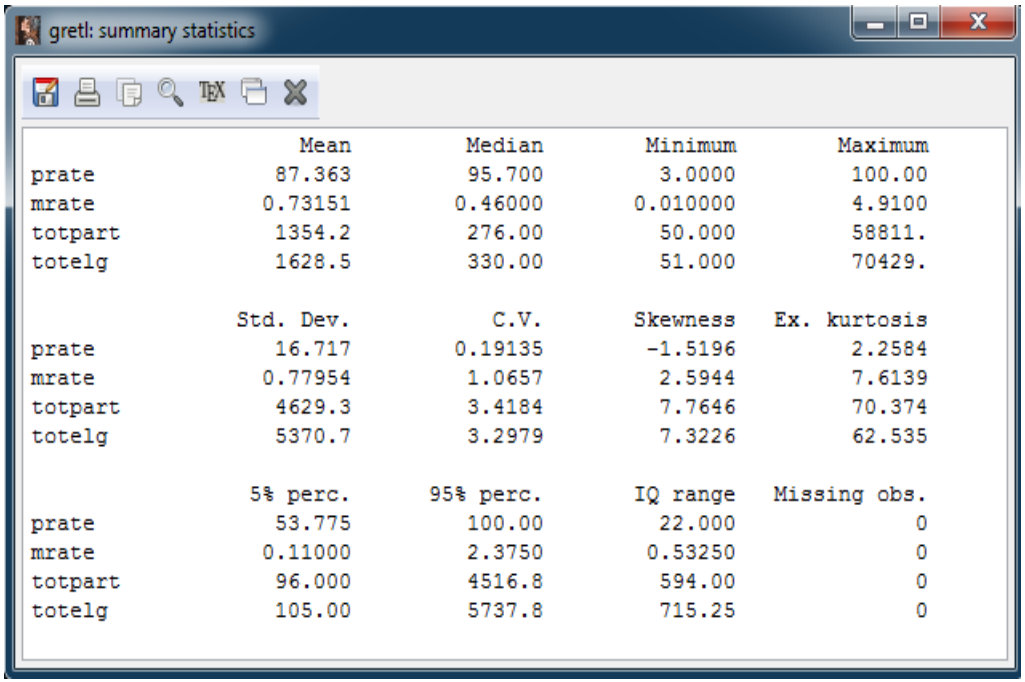
Descriptive statistics.

Summary statistics in Gretl.

Go up to the Menu Bar and click

`View -> Summary statistics`

The output appears in the figure below.



	Mean	Median	Minimum	Maximum
prate	87.363	95.700	3.0000	100.00
mrate	0.73151	0.46000	0.010000	4.9100
totpart	1354.2	276.00	50.000	58811.
totelg	1628.5	330.00	51.000	70429.

	Std. Dev.	C.V.	Skewness	Ex. kurtosis
prate	16.717	0.19135	-1.5196	2.2584
mrate	0.77954	1.0657	2.5944	7.6139
totpart	4629.3	3.4184	7.7646	70.374
totelg	5370.7	3.2979	7.3226	62.535

	5% perc.	95% perc.	IQ range	Missing obs.
prate	53.775	100.00	22.000	0
mrate	0.11000	2.3750	0.53250	0
totpart	96.000	4516.8	594.00	0
totelg	105.00	5737.8	715.25	0

Descriptive statistics.

Summary statistics in Gretl.

Consider a sample of N observations of the random variable X: $X_1, X_2, \ldots, X_N$.

Mean	$\hat{\mu}_X = \bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
Variance	$\hat{V}(X) = \hat{\sigma}_X^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}$
Standard deviation	$\hat{\sigma}_X = \sqrt{\hat{\sigma}_X^2}$
Coefficient of variation	C.V. $= \frac{\hat{\sigma}_X}{\bar{X}}$
Skewness	$\frac{1}{N-1} \frac{\sum_{i=1}^{N} (X_i - \bar{X})^3}{\hat{\sigma}_X^3}$
Excess kurtosis	$\frac{1}{N-1} \frac{\sum_{i=1}^{N} (X_i - \bar{X})^4}{\hat{\sigma}_X^4} - 3$

Descriptive statistics.

Summary statistics in Gretl.

The **quantile** or order p $(0 < p < 1)$ of a ranked set of data is the value x_p such that the probability that the random variable will be less or equal than x_p is at most p.

The **quartiles** of a ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data: the 25 % quantile, (Q1), 50 % quantile (Q2) and the 75 % quantile (Q3).

The **percentiles** of a ranked set of data values are the ninety nine points that divide the data set into 100 equal groups, each group comprising a hundreth of the data.

Median	The middle value of a ranked set of data values: half the figures will fall below the median and half above. It is the 50 % percentile and the second quartile.
IQ range	The inter quartile range is the difference between the third and the first quartiles.
5 % Percentile	Value of the variable such that 5 % of the ranked set of data are smaller.
95 % Percentile	Value of the variable such that 95 % of the ranked set of data are smaller.

See **Example 3.3** for applications.

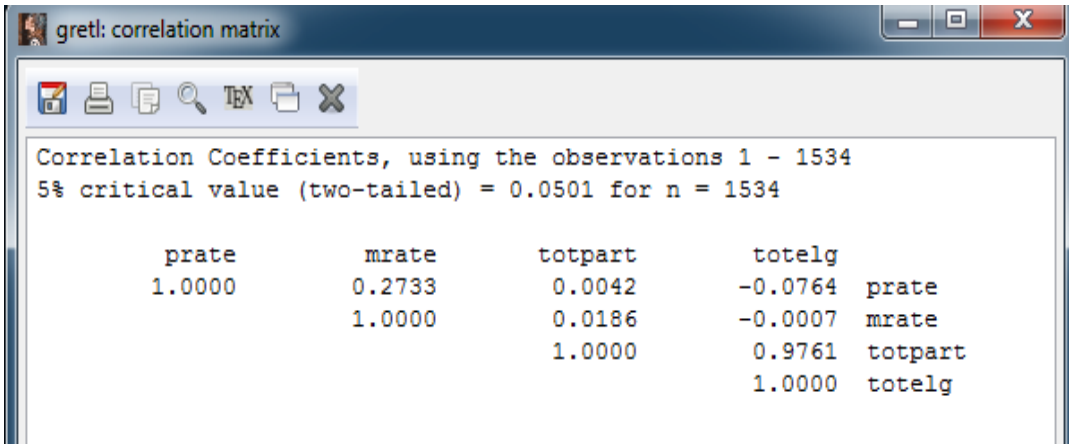
Descriptive statistics.

Summary statistics in Gretl.

Go up to the Menu Bar and click

`View -> Correlation matrix`

The output is shown in the figure below.



Each element of this matrix is the estimated correlation coefficient between two variables:

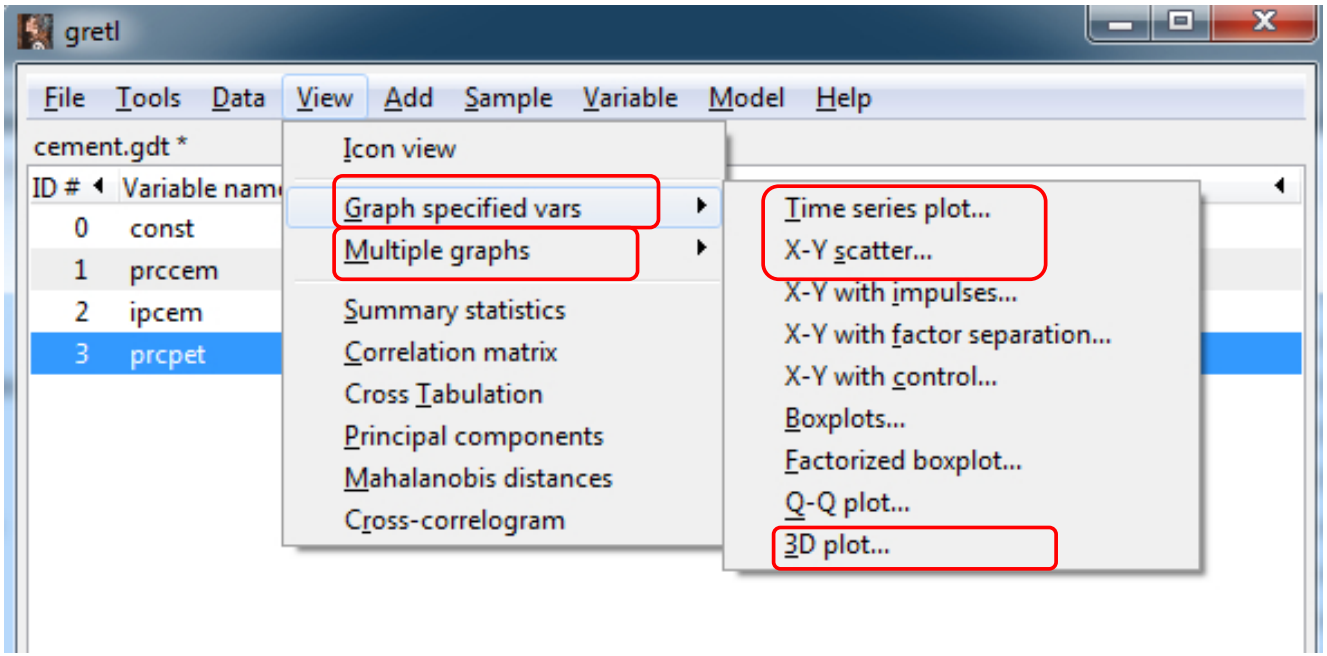
$$\hat{\rho}_{Y,X2} = 0.9472 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X2_i - \bar{X2})}{\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \ \sum_{i=1}^{N}(X2_i - \bar{X2})^2}}$$

Contents

Graphics for data analysis in Gretl.

Graphs in Gretl.

The pulldown menu **View** in the Main Bar includes the commands required to make graphs in Gretl.



Graphics for data analysis in Gretl.

Time series graphs.

A time series graph is the plot of a time series variable against time.

1. You may plot several time series on one single graph, clicking

   `View -> Graph specified vars -> Time series plot ...`

2. You may plot several time series in separate graphs, clicking

   `View -> Multiple graphs -> Time series ...`

See **Example 3.4.1** for applications.

X-Y scatter.

A scatter represents the cloud of points (X, Y), where the variable X is in the abscissas axes and the variable Y is in the ordinate axes.

1. You may plot several pairs of variables on one single graph, clicking

   `View -> Graph specified vars -> X-Y scatter ...`

2. You may plot several pairs of variables in separate graphs, clicking

   `View -> Multiple graphs -> X-Y scatters ...`

See **Example 3.4.2** for applications.

Graphics for data analysis in Gretl.

3D Graphs.

Go up to the Menu Bar and click

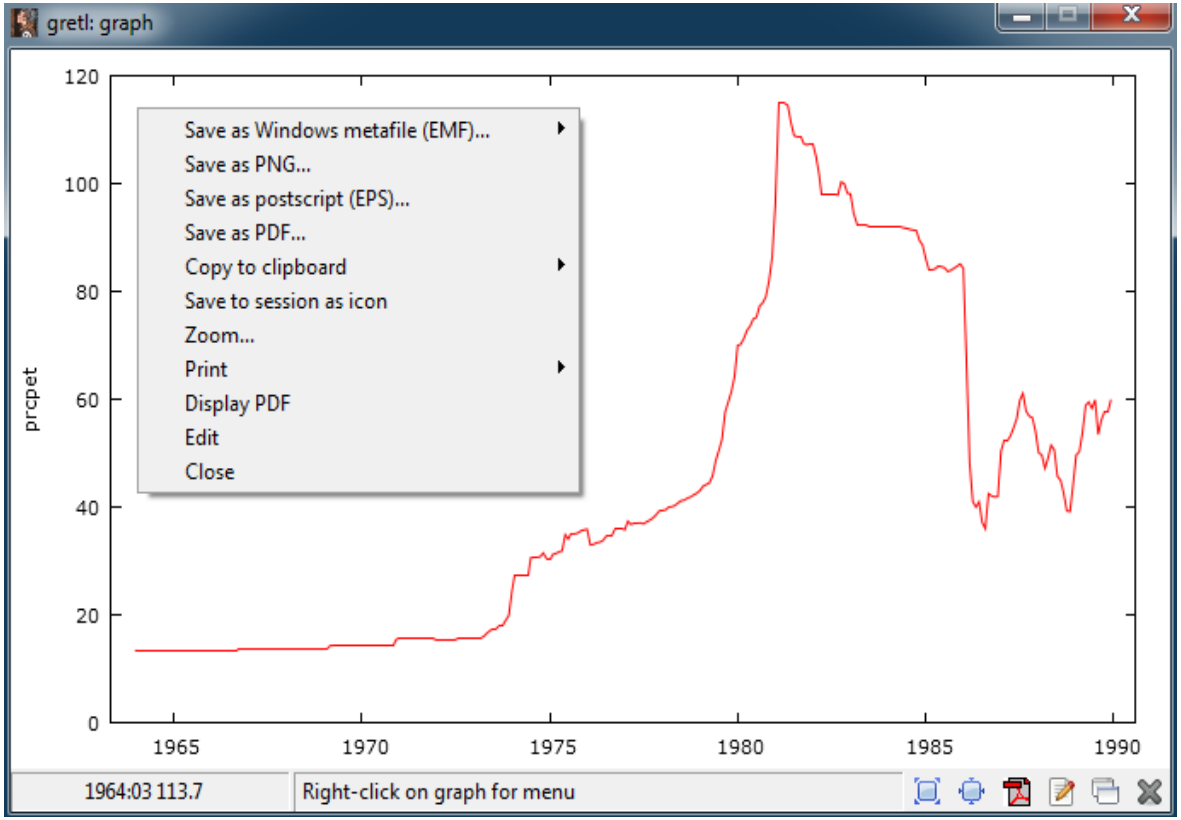
```
View -> Graphs specified vars    -> 3D plot ...
```

See **Example 3.4.3** for applications.

Graphics for data analysis in Gretl.

Edit a graph in Gretl.

Right-clicking on the graph yields a pulldown menu that offers a number of options.



Graphics for data analysis in Gretl.

Edit a graph in Gretl.

1. Save the graphs in a variety of formats:
   - Windows Metafile (EMF), color or monochrome.
   - PNG file.
   - Postscript (EPS) file.
   - PDF file.
2. Copy to clipboard to export to another file (Microsoft Word file, ...).
3. Save the graph in the session as an icon.
4. Zoom ... This option enables us to select a section of the graph to see it more clearly. Right-clicking on the graph and selecting the option *Restore full view*, the original graph is recovered.
5. Print the graph (color or monochrome).
6. Display the pdf file.
7. Edit the graph: colours, labels, dimensions, ...

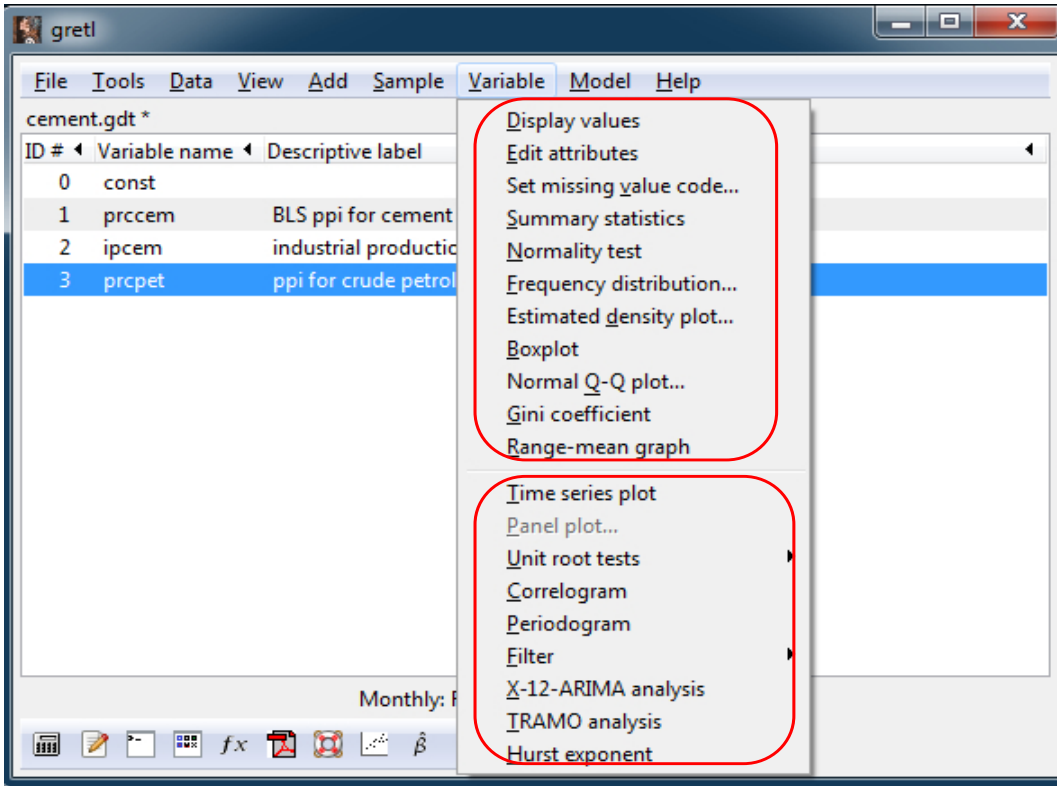
See **Example 3.4.4** for applications.

Contents

Univariate analysis.

Variable menu.

First, highlight the variable you want to analyse using the cursor. The pulldown menu **Variable** in the Menu Bar becomes active allowing us to obtain information about different aspects of this variable.



Univariate analysis.

The pulldown menu **Variable** includes a complete set of options which is split in two sections.

The first subset of options can be used with both cross-section and time series data. The second subset of options is relevant only for time series data.

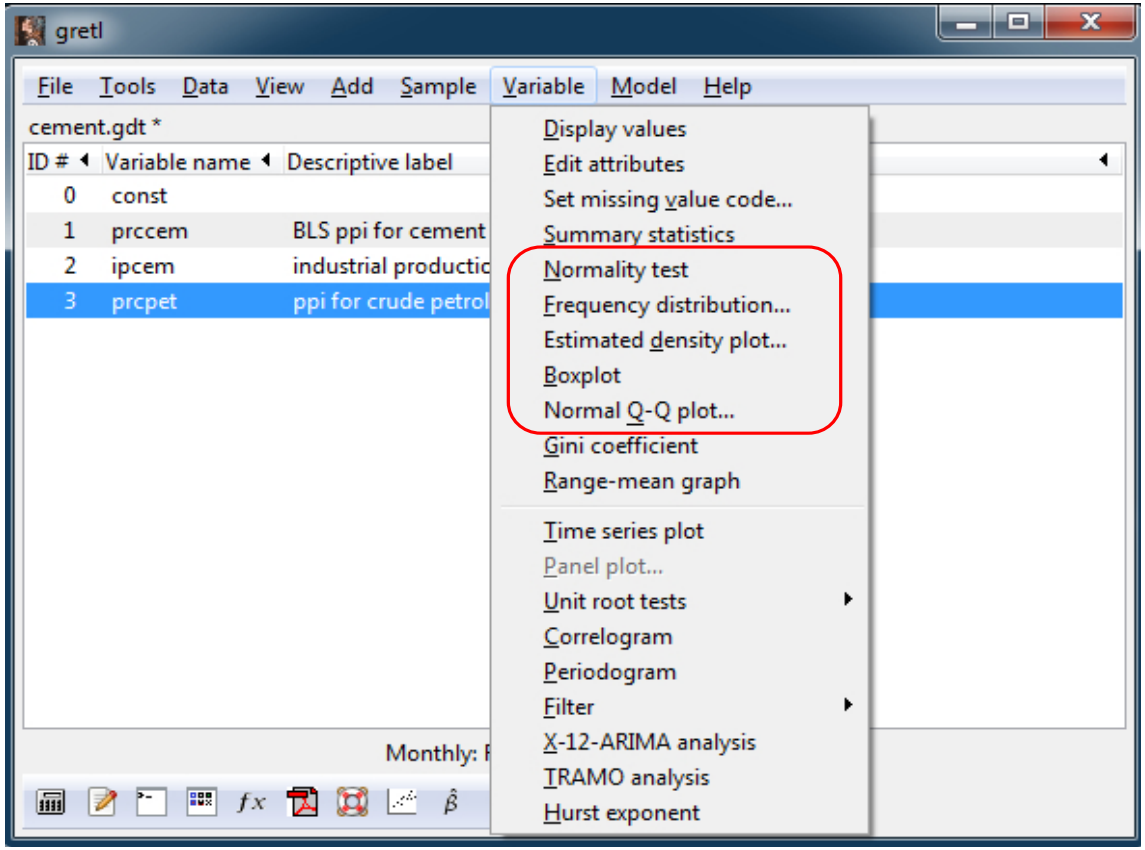
It should be noted that:

- Some options of the menu **Variable** are included in other menus (Display values, Edit attributes, Set missing value code, Summary statistics, Time series plot...).
- The only options revised in this section are the ones related to the distribution of the selected variable.
- The rest of the options of this menu correspond to and advance level of Econometrics.

Univariate analysis.

Distributions.

The pulldown menu **Variable** includes several options that help analyse the characteristics of the distribution of the selected variable.



Univariate analysis.

Normality tests.

Clicking **`Variable -> Normality test`**, the value of the statistics and the p-values are obtained for a number of normality tests:

- Doornik-Hansen test
- Lilliefors test
- Shapiro-Wilk W
- Jarque-Bera test

Null hypothesis: the selected variable follows a normal distribution

Alternative hypothesis: the selected variable does not follow a normal distribution

The p-value indicates whether there is evidence in the sample that the variable follows a normal distribution. If the p-value is smaller than α, it may be concluded that the selected variable does not follow a normal distribution at an $\alpha\,\%$ significance level.

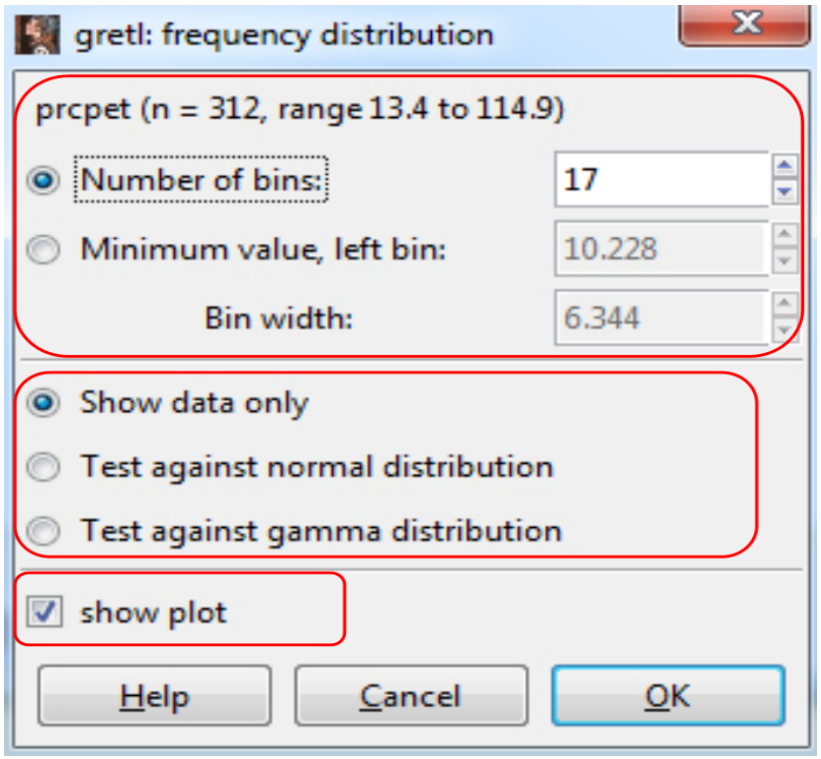
See **Example 3.5.1** for applications.

Univariate analysis.

Frequency distributions.

Go up to the Menu Bar and click

`Variable -> Frequency distribution ...`



See **Example 3.5.2** for applications.

Univariate analysis.

Distributions.

The dialog box is split into three sections:

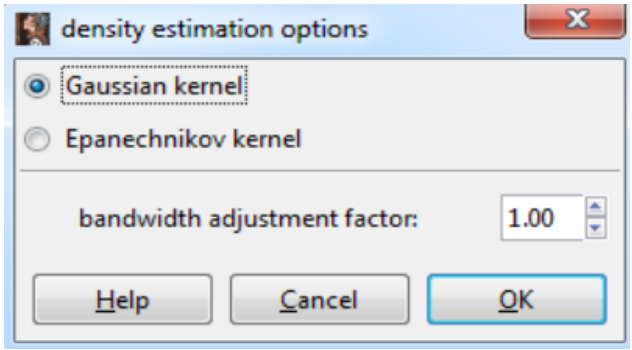
- In the first section we find:
  - The number of observations and the range of the selected variable.
  - Two options for the number of bins:
    - To choose a specific number of bins (Gretl proposes a number by default).
    - To choose the minimum value for the first (left) bin, and the width of the bins.
- In the second section, there are three options for the output:
  - Only the numerical results for the frequency distribution.
  - To run a test against the normal distribution (Doornik-Hansen test).
  - To run a test against the gamma distribution.
- In the third section, you may decide whether you want Gretl to show the frequency distribution plot. In the case of the tests, the graph of the estimated density is included by default.

Univariate analysis.

Estimated density.

Go up to the **Variable** pulldown menu and click

```
Variable -> Estimated density plot ...
```



You have to choose a weight function (kernel) in the dialog box: Gaussian kernel or Epanechnikov kernel.

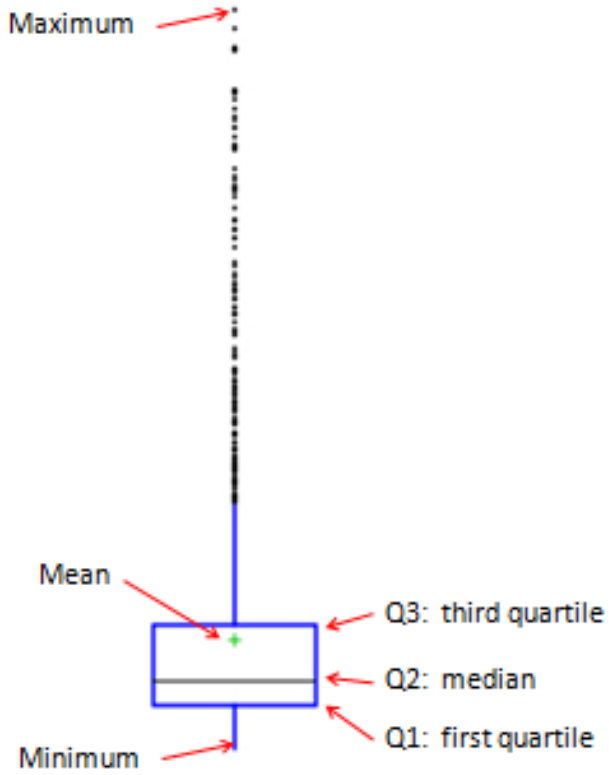
The bandwith adjustment factor is used to increase or decrease the smoothness of the weights. If you are not an expert, it is recommended not to change this parameter.

See **Example 3.5.3** for applications.

Univariate analysis.

Boxplot.

The simple boxplot shows the three quartiles, Q1, Q2 (the median) and Q3, the sample mean and the minimum and maximum values.

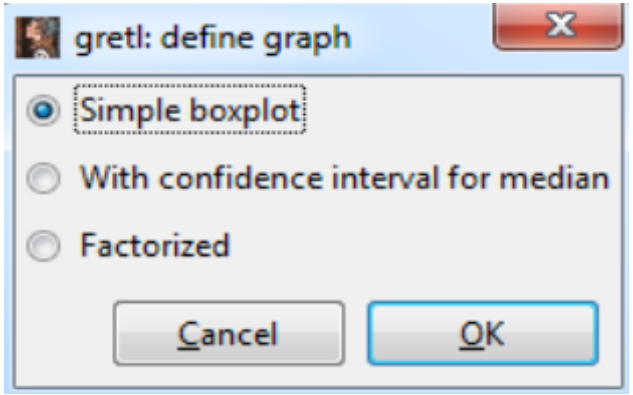


Univariate analysis.

Boxplot.

Go up to the Menu Bar and click

`Variable -> Boxplot`



You have to choose the type of boxplot in the dialog box:

- Simple boxplot.
- Boxplot including a confidence interval for the median.
- Factorized boxplot. With this option you get as many boxplots as the number of categories of the discrete variable used as factor.

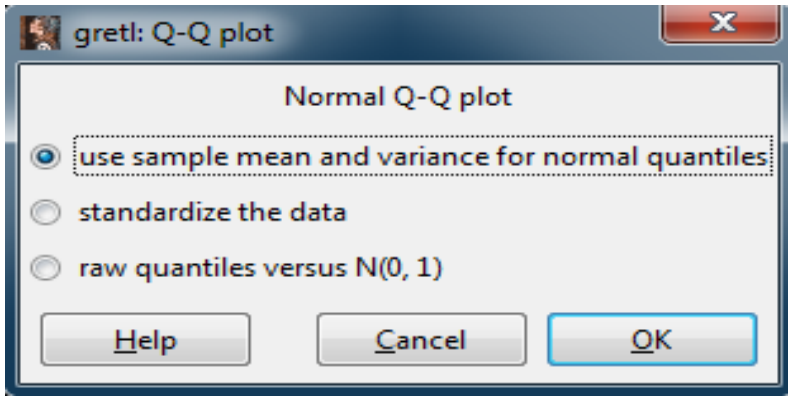
See **Example 3.5.4** for applications.

Univariate analysis.

Q-Q plot.

The Q-Q plot is a graphic tool to check whether the sample comes from a normal distribution: the closer the data are to the diagonal line, the greater the evidence in favor of the normal distribution is. Go up to the Menu Bar and click

`Variable -> Normal Q-Q plot`



Gretl offers several options:

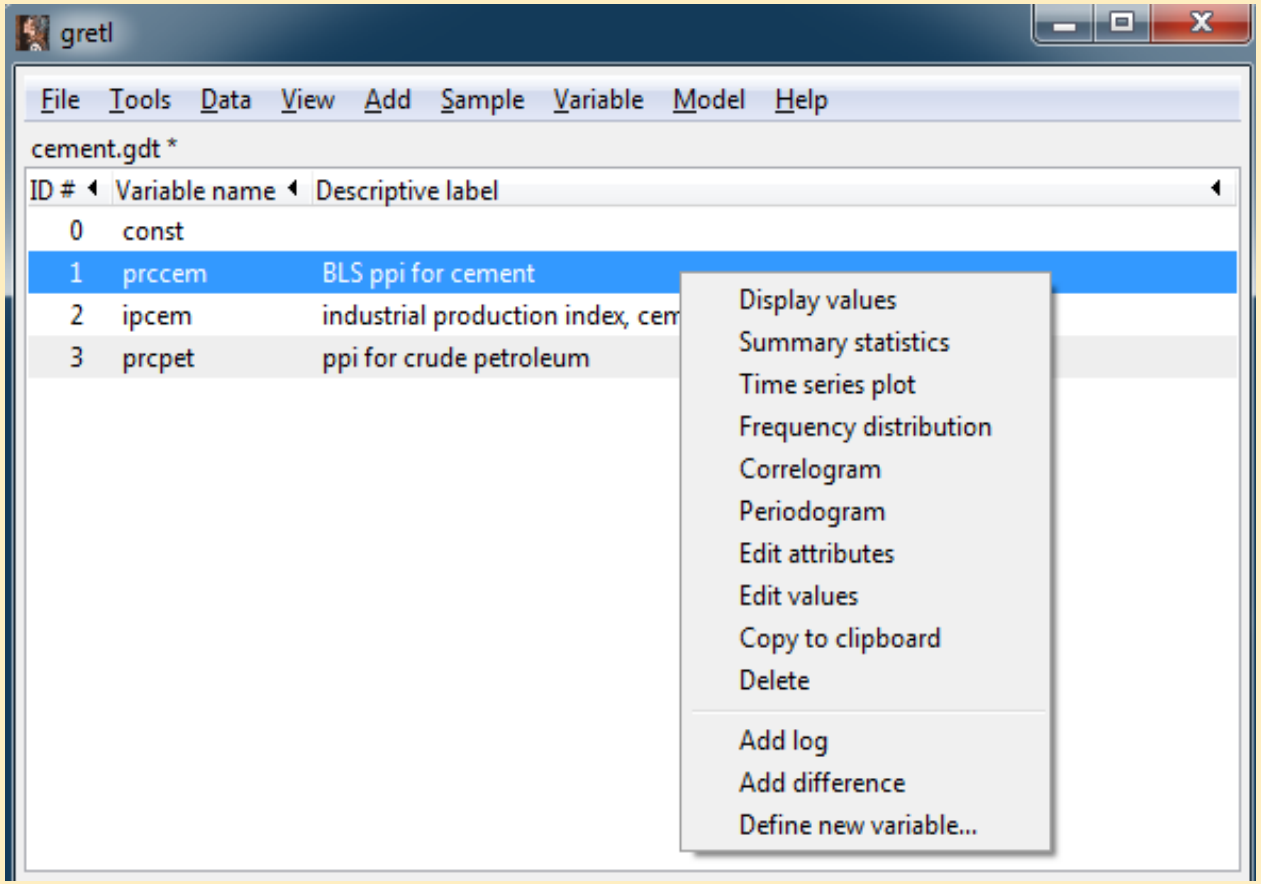
- Use sample mean and variance for normal quantiles.
- Standardize the data.
- Use the raw quantiles versus the standard normal distribution.

See **Example 3.5.5** for applications.

Univariate analysis.

A shorcut.

Right-clicking on a selected variable yields a pulldown menu showing a set of options to analyse the variable.



Contents