Forecasting Methods / Métodos de Previsão

Week 3

ISCTE - IUL, Gestão, Econ, Fin, Contab.

Diana Aldea Mendes

diana.mendes@iscte.pt

February 17, 2011

- We observe data for x_t, but since y_t also depends on u_t (ε_t), we must be specific about how the u_t (ε_t) are generated.
- We make the following set of assumptions about the u_t's for the regression methods to be valid :

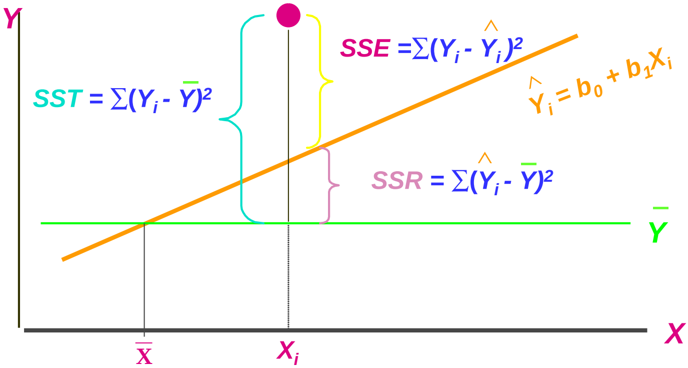
1. $E(u_t) = 0$, The errors have zero mean
2. $Var(u_t) = \sigma^2, \forall x_t$, The variance of the errors is constant and finite over all values of x_t (homoscedasticity)
3. $Cov(u_i, u_j) = 0$ ($E(u_i, u_j) = 0$, $(i \neq j)$), The errors are statistically independent of one another
4. $Cov(u_t, x_t) = 0$ ($E(u_t, x_t) = 0$), No relationship between the error and corresponding x variable (alternative assumption: the x_t's are non-stochastic or fixed in repeated samples.)
5. u_t is normally distributed ($u \sim N(0, \sigma^2)$) (this assumption is required if we want to make inferences about the population parameters from the sample parameters)

- If assumptions 1. through 4. hold, then the estimators $(\hat{\beta}_0, \hat{\beta}_1)$ determined by OLS are known as **Best Linear Unbiased Estimators** (BLUE).
  - "Estimator" - $\hat{\beta}$ is an estimator of the true value of β.
  - "Linear" - $\hat{\beta}$ is a linear estimator
  - "Unbiased" - On average, the actual value of $\hat{\beta}$ will be equal to the true values.
  - "Best" - means that the OLS estimator $\hat{\beta}$ has minimum variance among the class of linear unbiased estimators. The **Gauss-Markov theorem** proves that the OLS estimator is best.

- **Consistent :** The least squares estimators $\hat{\beta}$ are consistent. That is, the estimates will converge to their true values as the sample size increases to infinity (Need the assumptions 2 and 4 to prove this).
- **Unbiased:** The least squares estimates are unbiased. That is $E\left(\widehat{\beta}\right) - \beta = 0$, thus on average the estimated value will be equal to the true values (to prove this we need assumption 1). Unbiasedness is a stronger condition than consistency.
- **Efficiency:** An estimator of parameter β is said to be efficient if it is unbiased and no other unbiased estimator has a smaller variance, i.e. $Var\left(\widehat{\beta}\right) < Var\left(\widetilde{\beta}\right)$. If the estimator is efficient, we are minimizing the probability that it is a long way off from the true value of β.

- In addition to determining the coefficients of the least squares line, we need to assess it to see how well it "fits" the data. They're based on the sum of squares for errors (SSE).

$$
\begin{aligned}
SST &= SS_{yy} = \sum_{t=1}^{T} (y_t - \bar{y})^2 \text{ (total sum of squares, measure the deviation of the observations from their mean)} \\
SSR &= \sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2 \text{ (explained sum of squares, measure the deviation of predicted values from the mean)} \\
SSE &= \sum_{t=1}^{T} (u_t)^2 = \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \text{ (residual sum of squares)}
\end{aligned}
$$



Y
SSE =∑(Yi - Ŷi)²
SST = ∑(Yi - Ȳ)²
Ŷi = b0 + b1Xi
SSR = ∑(Ŷi - Ȳ)²
Ȳ
X
X̄
Xi

- How do we think about how well our sample regression line fits our sample data?
- Compute the fraction of the total sum of squares (SST) that is explained by the model, call this the **R-squared** of regression (Coefficient of Determination), to judge the adequacy of the regression model

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- $0 \leq R^2 \leq 1$, represents the percent of the data that is the closest to the line of best fit.
- The higher the R^2, the more useful the model.

- For example, $R^2 = 0.850$, means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.
- If x contributes lots of information about y then SSE is very small.
- Interpretation: R^2 tells us how much better we do by using the regression equation rather than just $\bar{y}$ to predict y

- Another measure of how well the model fits the data is the **Standard Error** of the $\hat{\beta}$ estimate (measures the spread of the actual points around the fitted line)
- Since σ^2 (population variance) is best estimated by s^2 (sample variance) we have that

$$
\begin{aligned}
s &= \sqrt{\frac{SSE}{df}} = \sqrt{\frac{\sum_{t=1}^{T}\left(y_t - \widehat{y}_t\right)^2}{T-2}} = \sqrt{SS_{yy} - \widehat{\beta}_1 SS_{xy}} \text{ where} \\
SS_{yy} &= \sum_{t=1}^{T}\left(y_t - \overline{y}\right)^2 = \sum_{t=1}^{T} y_t^2 - \frac{\left(\sum_{t=1}^{T} y_t\right)^2}{T}; \\
SS_{xx} &= \sum_{t=1}^{T} x_t^2 - \frac{\left(\sum_{t=1}^{T} x_t\right)^2}{T}, SS_{xy} = \sum_{t=1}^{T} x_t y_t - \frac{\left(\sum_{t=1}^{T} x_t\right)\left(\sum_{t=1}^{T} y_t\right)}{T};
\end{aligned}
$$

df degrees of freedom

- **Standard errors (**measure the precision (reliability) of the estimate**)** for the β_1 and β_2 estimates are given by

$$s_{\widehat{\beta}_1} = \sqrt{\frac{s^2}{SS_{xx}}}; \quad s_{\widehat{\beta}_0} = \sqrt{s^2\left(\frac{1}{T} + \frac{\overline{x}^2}{SS_{xx}}\right)}$$

since

$$\begin{aligned}
E\left(\widehat{\beta}_1\right) &= \beta_1,\ Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{SS_{xx}} = \frac{\sigma^2}{\sum\left(x_t - \overline{x}\right)^2} \\
\widehat{\beta}_1 &\sim N\left(\beta_1, \frac{\sigma^2}{\sum\left(x_t - \overline{x}\right)^2}\right);\ E\left(\widehat{\beta}_0\right) = \beta_0, \\
Var\left(\widehat{\beta}_0\right) &= \sigma^2\left[\frac{1}{T} + \frac{\overline{x}^2}{\sum\left(x_t - \overline{x}\right)^2}\right]; \widehat{\beta}_0 \sim N\left(\beta_0, Var\left(\widehat{\beta}_0\right)\right)
\end{aligned}$$

- $s_{\widehat{\beta}_0}$ and $s_{\widehat{\beta}_1}$ depend on s^2. The greater the variance s^2, then the more dispersed the errors are about their mean value and therefore the more dispersed y will be about its mean value.
- The larger the SS_{xx} (sum of squares of x), the smaller the coefficient variances.
- The larger the sample size, T, the smaller will be the coefficient variances.

- We can use the information in the sample to make inferences about the population.
- We will always have two hypotheses that go together, the **null hypothesis** (denoted H_0) and the **alternative hypothesis** (denoted H_1).
- The null hypothesis is the statement or the statistical hypothesis that is actually being tested. The alternative hypothesis represents the remaining outcomes of interest.

- **For example**: we are interested in the hypothesis that the true value of β_1 is in fact 0.5. We would use the notation

$$\begin{aligned} H_0 &: \quad \beta_1 = 0.5 \\ H_1 &: \quad \beta_1 \neq 0.5 \end{aligned}$$

- This is known as a two sided test.
- There are two ways to conduct a hypothesis test: via the **test of significance** approach or via the **confidence interval** approach.

The steps involved in doing a test of significance are:

- Estimate $\hat{\beta}_0, \hat{\beta}_1$, and $s_{\widehat{\beta}_0}$, $s_{\widehat{\beta}_1}$, in the usual way
- Calculate the test statistic. This is given by the formula

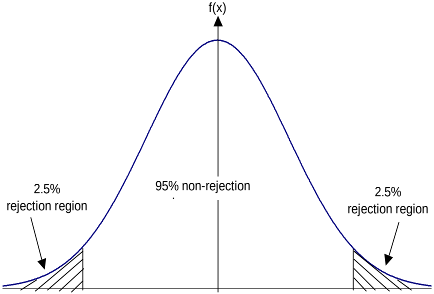
$$test\ statistic = \frac{\hat{\beta} - \beta^{'}}{s_{\widehat{\beta}}}$$

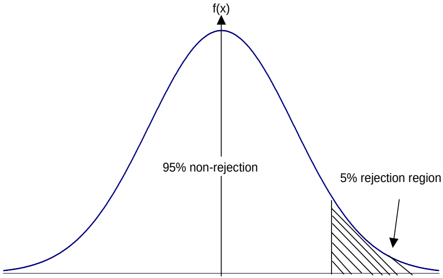
  where β^* is the value of β under the null hypothesis $H_0 : \hat{\beta} = \beta^{'}$
- Since σ^2 is unknown, we will use the $t - Student$ statistics (instead of normal statistics), that is

$$\frac{\widehat{\beta_0} - \beta_0'}{s_{\widehat{\beta_0}}} \sim t_{T-2}, \quad \frac{\widehat{\beta_1} - \beta_1'}{s_{\widehat{\beta_1}}} \sim t_{T-2}$$

- We need some tabulated distribution with which to compare the estimated test statistics ($t - Student$ distribution with $T - 2$ degrees of freedom).
- We need to choose a "significance level", often denoted α (the size of the test and it determines the region where we will reject or not reject the null hypothesis that we are testing). It is conventional to use a significance level of 5% (10% and 1% are also commonly used).
- Given a significance level, we can determine a rejection region and non-rejection region.
- Use the t-tables to obtain a critical value or values with which to compare the test statistic.
- Finally perform the test. If the test statistic lies in the rejection region then reject the null hypothesis (H_0), else do not reject H_0.

- Two-sided and one-sided (upper tail) test of significance





- For simple linear regression, we must test if y depends or not on x (x cause y)

$$
\begin{aligned}
H_0 &: \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0 \\
\text{test statistic } t &= \frac{\widehat{\beta_1}}{s_{\widehat{\beta_1}}} = \frac{\widehat{\beta_1}}{s/\sqrt{SS_{xx}}} \sim t_{n-2} \quad (t\text{-ratio test}) \\
\text{rejection region} &: t_{calc} < -t_{\alpha/2} \text{ or } t_{calc} > t_{\alpha/2} \quad (\alpha \text{ significance level} \\
p - \text{value (prob)} &: p = P\left(|t_{n-2}| > |t_{calc}|\right)
\end{aligned}
$$

- Variables that are not significant (not reject H_0) are usually removed from regression model

- For simple linear regression, we also have to test if the intercept is zero:

$$
\begin{aligned}
H_0 \quad &: \quad \beta_0 = 0 \ \text{ versus } \ H_1 : \beta_0 \neq 0 \\
\text{test statistic } t \quad &= \quad \frac{\widehat{\beta}_0}{s_{\widehat{\beta}_0}} \sim t_{n-2} \ \text{ (sob } H_0\text{)}
\end{aligned}
$$

- Estimate $\hat{\beta}_0, \hat{\beta}_1$, and $s_{\widehat{\beta}_0}$, $s_{\widehat{\beta}_1}$, in the usual way
- Choose a significance level, α, (again the convention is 5%). This is equivalent to choosing a $(1-\alpha)\times 100\%$ confidence interval, i.e.

$$\boxed{5\% \textbf{ significance level}} = \boxed{95\% \textbf{ confidence interval}}$$

- Use the t-tables to find the appropriate critical value, which will again have $T-2$ degrees of freedom.
- The confidence interval is given by

$$CI_{100(1-\alpha)\%} = \left]\widehat{\beta} - t_{crit}\ s_{\widehat{\beta}}, \widehat{\beta} + t_{crit}\ s_{\widehat{\beta}}\right[$$

- Perform the test: If the hypothesized value of β (β') lies outside the confidence interval, then reject the null hypothesis that $\beta = \beta'$, otherwise do not reject the null.
- Note that the Test of Significance and Confidence Interval approaches always give the same answer.
- If we reject the null hypothesis at the 5% level, we say that the result of the test is **statistically significant**

Example

Consider the following regression results

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t = \underset{(14.38)}{20.3} + \underset{(0.2561)}{0.5091}\, x_t, \quad T = 22$$

Using both the test of significance and confidence interval approaches, test the hypothesis that $\beta_1 = 1$ against a two-sided alternative, that is

$$\begin{aligned} H_0 &: \beta_1 = 1 \\ H_1 &: \beta_1 \neq 1 \end{aligned}$$

- The first step is to obtain the critical value. We want $t_{crit} = t_{20;5\%}$
- Test of significance approach

$$\text{test stat} \quad = \quad \frac{\widehat{\beta}_1 - \beta_1'}{s_{\widehat{\beta}_1}} = \frac{0.5091 - 1}{0.2561} = -1.917$$

Do not reject H_0 since test stat lies within **non-rejection** region (t_{crit} are ± 2.086)

- Confidence interval approach

$$\left]\widehat{\beta}_1 - t_{crit}\ s_{\widehat{\beta}_1}, \widehat{\beta}_1 + t_{crit}\ s_{\widehat{\beta}_1}\right[\quad = \quad]0.509 - 2.086 \times 0.256, 0.509 + 2.086$$

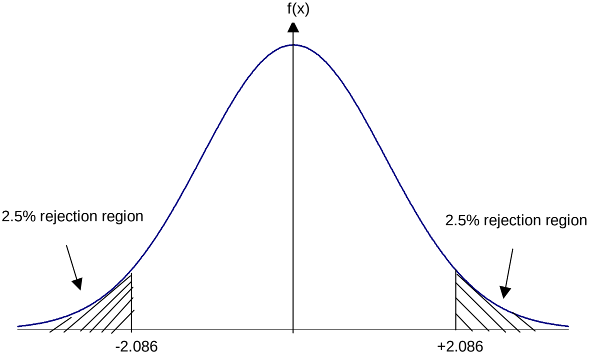
$$= \quad]-0.0251,\ 1.0433[$$

since 1 lies within the confidence inter

do not reject H_0

Linear regression

Hypothesis Testing - Example



- We usually reject H_0 if the test statistic is statistically significant at a chosen significance level.
- There are two possible errors we could make:
  - Rejecting H_0 when it was really true. This is called a **type I error**.
  - Not rejecting H_0 when it was in fact false. This is called a **type II error**

		Reality	
		H_0 is true	H_0 is false
Result	Significant (reject H_0)	Type I error (α)	OK
of test	Insignificant (not reject H_0)	OK	Type II error (β)

- The probability of a type I error is just α, the significance level or size of test we chose.
- What happens if we reduce the size of the test (e.g. from a 5% test to a 1% test)?

$$\begin{array}{lll} \text{reduce size of test} & \longrightarrow & \text{more strict criterion for rejection} \longrightarrow \\ & \longrightarrow & \text{reject null hypothesis less often} \longrightarrow \\ & \longrightarrow & \begin{array}{l}\text{less likely to falsely reject} \\ \text{more likely to incorrectly not reject}\end{array} \end{array}$$

- So there is always a trade off between type I and type II errors when choosing a significance level. The only way we can reduce the chances of both is to **increase the sample size**.

- The **p-value** is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. The lower the p-value, the less likely the result is if the null hypothesis is true, and consequently the more "significant" the result is, in the sense of statistical significance.
- This is equivalent to choosing an infinite number of critical t-values from tables. It gives us the marginal significance level where we would be indifferent between rejecting and not rejecting the null hypothesis.
- We reject the null hypothesis when the **p-value** is less than 0.05 or 0.01(significance level)

$$\boxed{\text{reject} \quad H_0 \quad \text{if} \quad p\text{-value} \leq \alpha}$$

- Once a model has been estimated (and carefully validated using economic and statistical tests) it can be used for prediction or forecasting.

Linear regression

Main steps in regression:

- Plot the data (scatter plot)
- Basic descriptive statistics (data)
- Model selection (parameter estimation)
- Evaluation of Assumptions (residuals)
- Model validation (assessment of the goodness of fit)
- Forecasting

Linear regression

Example

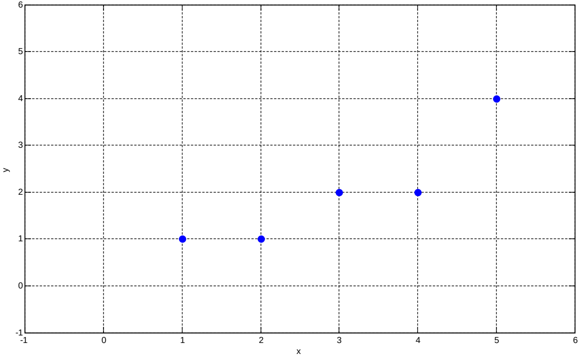
- A shop is doing an experience over 5 months to determine the effects of advertising on sales. Data are presented in the next table

Month	advertising costs×100€	sale revenue×1000€
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4
	x	y

Linear regression

Example

- scatter plot



Linear regression

Example

- We assume that the (linear) relationship between the revenue y and the costs of advertising x is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Parameters estimation by the method of least squares (OLS)

$$\begin{aligned} SS_{xy} &= \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 7 \text{ and } SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 10 \\ \widehat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}} = \frac{7}{10} = 0.7 \quad \text{and} \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} = \frac{\sum y_i}{5} - \widehat{\beta}_1 \frac{\sum x_i}{5} = -0.1 \end{aligned}$$

then, the regression line is given by

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_0 x = -0.1 + 0.7x$$

Linear regression

Example

- Estimation of SSE

$$\begin{aligned} SSE &= \sum_{i=1}^{5}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i\right)^2=(1+0.1-0.7)^2 \\ ...+(4+0.1-0.7\times 5)^2 &= 1.10 \end{aligned}$$

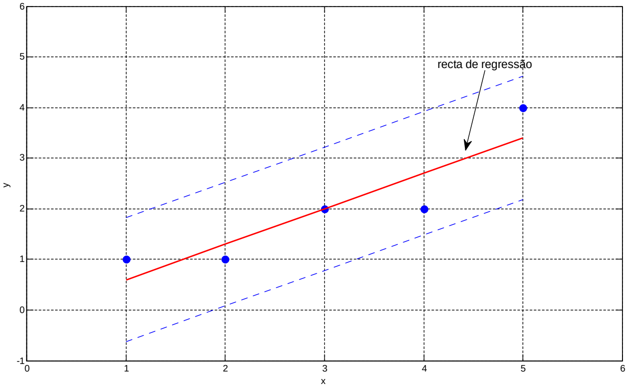
- Estimation of the population variance σ^2

$$s^2=\frac{SSE}{n-2}=\frac{1.10}{3}=0.367 \quad \Rightarrow s=0.61$$

therefore the majority of observations (data) belong to the interval of amplitude $2s=1.22$ (around the regression line, no outliers)

Linear regression

Example



- Hypothesis test: we consider the following null hypothesis

$$H_0 \quad : \quad \beta_1 = 0$$
$$H_1 \quad : \quad \beta_1 \neq 0$$

and we choose $\alpha = 0.05$ and since $T = 5$ we have $5 - 2 = 3$ degrees of freedom. Then the rejection region (for the two-sided test) is $t < t_{0.025} = -3.182$ or $t > t_{0.025} = 3.182$

- Since

$$t = \frac{\widehat{\beta}_1}{s/\sqrt{SS_{xx}}} = 3.7 > t_{0.025}$$

we reject H_0, and we have that $\beta_1 \neq 0$, which means that x contributes with information for the prediction of y in the regression model.

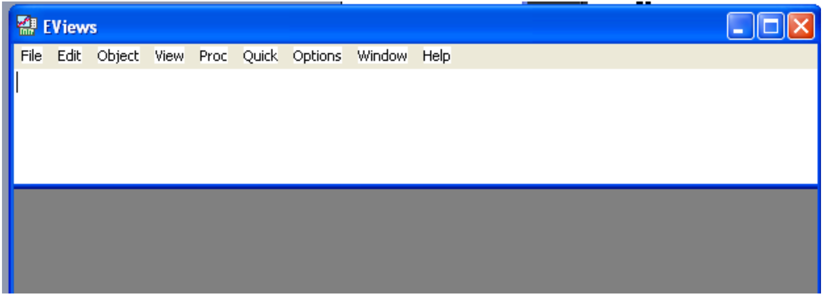
- Confidence interval: 95%

$$\widehat{\beta}_1 \pm t_{0.025} s_{\widehat{\beta}_1} = 0.7 \pm 3.182 \left(\frac{s}{\sqrt{SS_{xx}}} \right) = 0.7 \pm 0.61$$

so, with 95% confidence we can say that the parameter β_1 belongs to the interval $[0.09, 1.31]$ (The average revenue for each 100 euro spent on advertising is between 90 and 1310 euro)

Eviews

- Open the program by doubleclicking on the EViews icon
- You will be confronted by the following view (Command Window)



- The main menu options are shown at the top (File Edit Objects View . . .). If you click on any of these words a drop-down menu will appear with further options.

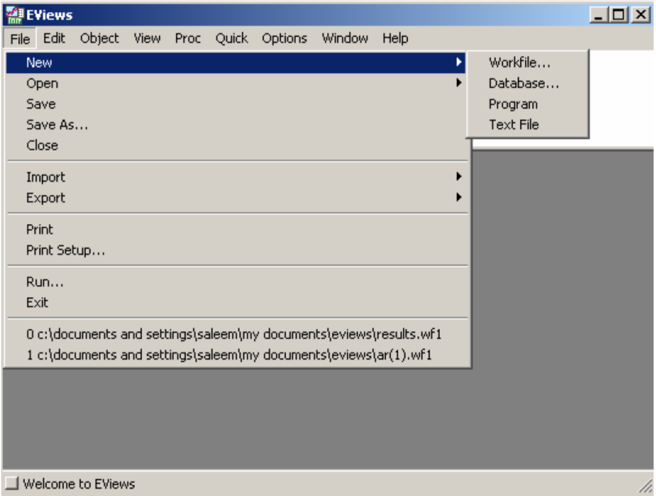
Introduction to Eviews

- The first step in a project is to read the data into an EViews workfile. EViews can import data from an Excel spreadsheet
  - **File** -> **New** -> **Workfile**
  - **Workfile Range** -> **Workfile frequency** -> **Start date** - **End date**
- save the Workfile: **File** -> **Save As**
- EViews describes data the following way: year (e.g. 1981); year:quarter (e.g. 1992:1); year:month. (e.g. 1990:11); month:day:year (e.g. 8:10:97). Cross-sectional data is stored as undated or irregular.

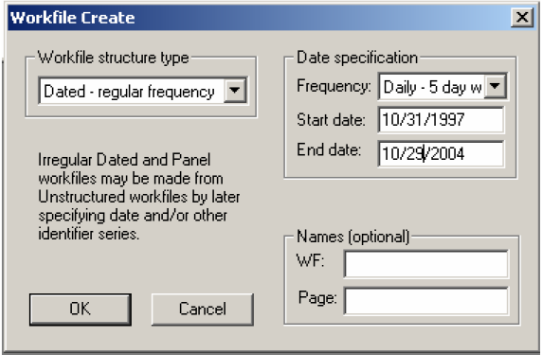
Introduction to Eviews

Creating a work file:

Open Eviews 5
Click File ----New----Work File



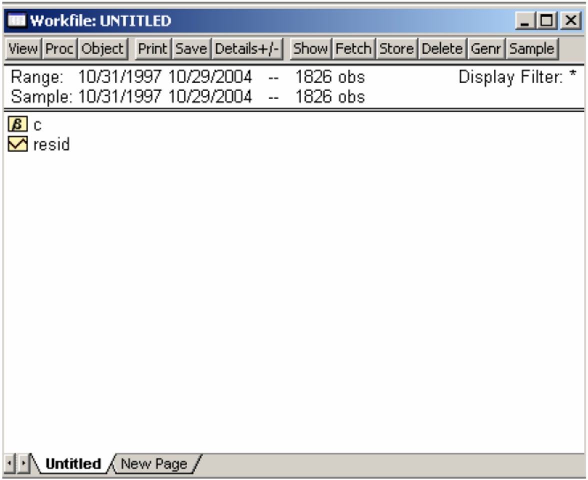
Introduction to Eviews



For example you have daily prices of KSE-100 from 31-10-1997 to 29-10-2004 with 5 days week. Enter the start and ending date and choose the frequency according to the data.

Note: Date should be mm/dd/yyyy format

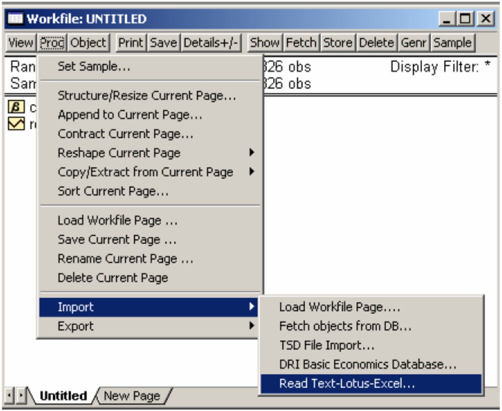
Click OK
You have created your work file with constant C and Residual



Introduction to Eviews

Importing Data:

The next step is to import data in Eviews.

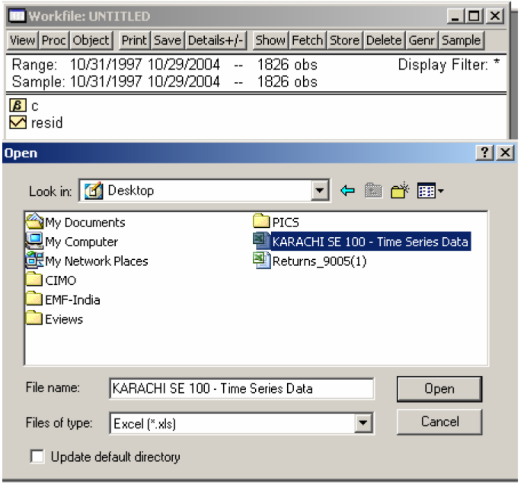


Click PROC....IMPORT....Read Text-Lotus-Excel
Go to the data file

Introduction to Eviews

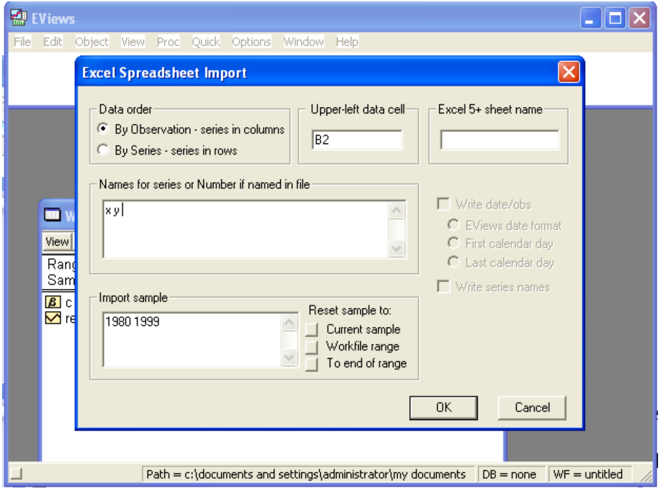
- Importing data: **Proc** -> **Import** -> **Read Text-Lotus-Excel** (from the workfile window)
- Importing data: **File** -> **Import** -> **Read Text-Lotus-Excel** (from the command window)
- Enter the names of the series that you wish to read into the edit box (alternatively, if the names that you wish to use for your series are contained in the file, you can simply provide the number of series to be read)
- If the data are organized by row and the starting cell is B2, then the names must be in column A, beginning at cell A2.
- If the data are organized by column beginning in B2, then the names must be in row 1, starting in cell B1.
- The name of the uploaded series will appear in the Command Window (joining **c** and **resid**)
- Save the workfile

Introduction to Eviews

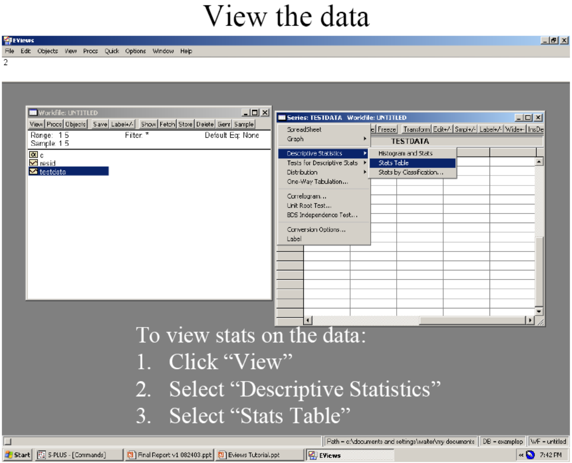


Remember to choose Excel(*.xls) as file type
Double click on the Data File
Note: The excel sheet must be closed while you are importing data to Eviews, other wise it will give you error.

Introduction to Eviews

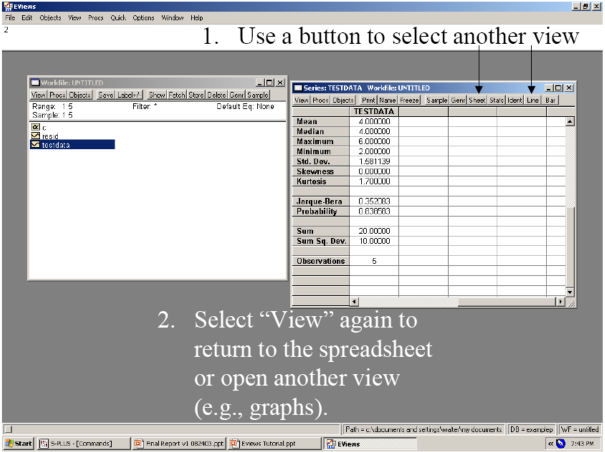


- Double-click on the variable and **View** -> **Graph** (variable window)



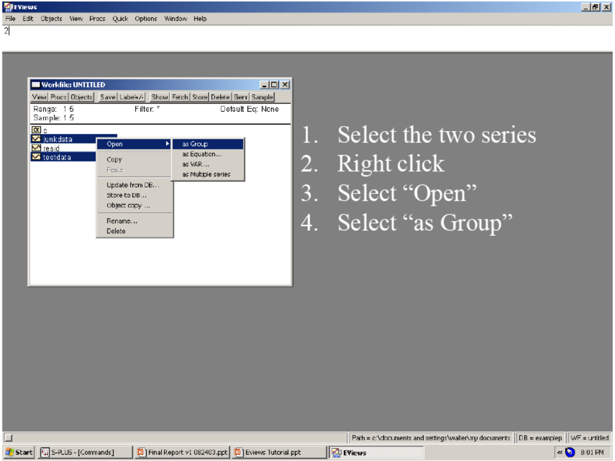
- Descriptive statistics: **View** -> **Descriptive Statistics** -> **Stats Table** (variable window))

A Statistics View



- Open a group of variables:

Viewing two or more series



- Important operators

+	add	x+y adds the contents of X and Y
–	subtract	x–y subtracts the contents of Y from X
*	multiply	x*y multiplies the contents of X by Y
/	divide	x/y divides the contents of X by Y
^	raise to the power	x^y raises X to the power of Y
>	greater than	x>y takes the value 1 if X exceeds Y, and 0 otherwise
<	less than	x<y takes the value 1 if Y exceeds X, and 0 otherwise
=	equal to	x=y takes the value 1 if X and Y are equal, and 0 otherwise
<>	not equal to	x<>y takes the value 1 if X and Y are not equal, and 0 if they are equal
<=	less than or equal to	x<=y takes the value 1 if X does not exceed Y, and 0 otherwise
>=	greater than or equal to	x>=y takes the value 1 if Y does not exceed X, and 0 otherwise
and	logical and	x and y takes the value 1 if both X and Y are nonzero, and 0 otherwise
or	logical or	x or y takes the value 1 if either X or Y is nonzero, and 0 otherwise
log(x)	natural logarithm	ln(X)

- Transform, define a new variable: **Quick** -> **Generate Series**.

  Example: Genr -> y=dlog(x)

- Exporting table of results:
  - Save table using **Freeze/Name**
  - Open table, select all, then copy and choose "formatted", thus paste in word file.

Regression with Eviews

- Graphical representation of data
  - **View** -> **Graphs** -> **Line** (one or several variables)
  - **View** -> **Graph** -> **Scatter** (a group of variables)
- Estimate the regression model
  - **Quick** -> **Estimate Equation** (main Eviews Window).
    - equation specification: dependent variable, c, independent variable(s);
    - estimation setting: LS - Least squares
  - Saving equation: using **Name** Option
- Interpret the output
- Test: Click **View->Coefficient Tests->Wald-Coefficient Restrictions**
- View results in the equation window: Regression output: click **View**, then **Estimation output**
- Residuals: click **View**, then **Actual, Fitted residual**