QF5210A Financial Time Series

Chao ZHOU

Department of Mathematics, NUS
matzc@nus.edu.sg

3rd Lecture, 12 May 2019

Textbook

This lecture follows closely Chirs Brooks's book Chapter 2 and Chapter 3.

Introductory Econometrics for Finance, 2nd Edition, Chris Brooks, Cambridge University Press, 2008. Chapters 2-8, 11. Software : EViews
(E-book available on the website of NUS Libraries)

Outline

Linear regression model

Regression analysis is almost certainly the most important tool in econometrics.

Univariate regression is an attempt to explain movements in a variable (dependent variable or response variable y) by reference to movements in one or more other variables (independent variable(s), predictor variable(s) or explanatory variable(s) $x_1, \cdots, x_p$).

One important goal of the regression is the prediction of future y values when the corresponding values of $x_1, \cdots, x_p$ are already available.

Data : y_t and $x_{t,1}, \cdots, x_{t,p}$, the values of these variables for the tth observation.

Simple linear regression model

Suppose that it is believed that y depends on only one x variable.

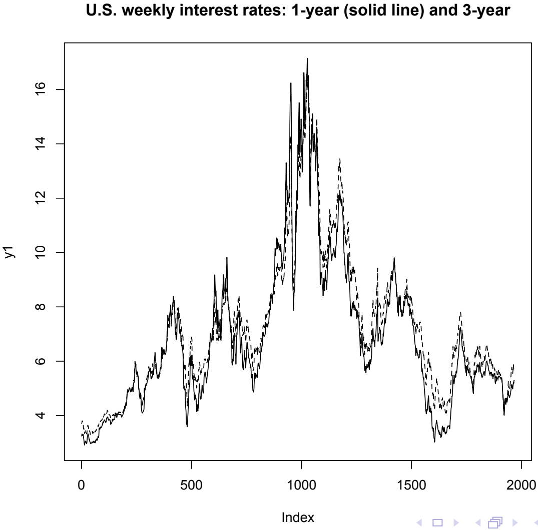
The model is

$$y_t = \alpha + \beta x_t + \epsilon_t \tag{1}$$

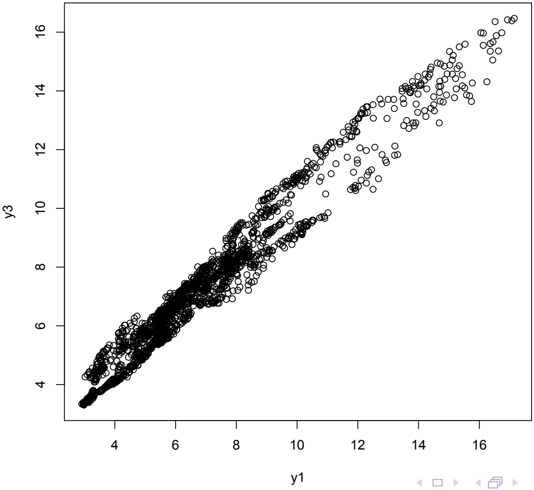
where the subscript t $(= 1, 2, 3, ..., T)$ denotes the observation number and ϵ a random disturbance (or noise, error) term.

α and β are also called intercept and slope of the line.

U.S. weekly interest rates



U.S. weekly interest rates : y1 y3



Notations

A hat (ˆ) over a variable or parameter is used to denote a value estimated by a model.

$\hat{\alpha}$ and $\hat{\beta}$ are the estimated values for α and β.

Let y_t denote the actual data point for observation t and let $\hat{y}_t$ denote the fitted value from the regression line.

Let $\hat{\epsilon}_t$ denote the residual, which is the difference between the actual value of y and the value fitted by the model for this data point, i.e. $(y_t - \hat{y}_t)$. $\hat{\epsilon}$ can be understood as an estimation for ϵ.

Least squares estimation

The residual sum of squares (RSS) (or the sum of squared residuals) is defined as

$$\sum_t \hat{\epsilon}_t^2$$

So minimising RSS is equivalent to minimising $\sum_t (y_t - \hat{y}_t)^2$.

The equation for the fitted line is given by $\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$.

Let L denote the RSS, which is also know as a loss function.

$$L = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{T} (y_t - \hat{\alpha} - \hat{\beta} x_t)^2 \tag{2}$$

where T is the number of observations.

L is minimised with respect to $\hat{\alpha}$ and $\hat{\beta}$ to find the values of α and β that give the line that is closest to the data.

The first order condition says that

$$\frac{\partial L}{\partial \hat{\alpha}} = -2\sum_{t=1}^{T}(y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \tag{3}$$

$$\frac{\partial L}{\partial \hat{\beta}} = -2\sum_{t=1}^{T}x_t(y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \tag{4}$$

(3) is equivalent to

$$\sum_{t=1}^{T} y_t - T\hat{\alpha} - \hat{\beta}\sum_{t=1}^{T} x_t = 0$$

$$T\bar{y} - T\hat{\alpha} - \hat{\beta}T\bar{x} = 0$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Substituting $\hat{\alpha}$ into (4)

$$\sum_{t=1}^{T} x_t(y_t - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_t) = 0$$

$$\sum_{t=1}^{T} x_t y_t - \bar{y}\sum_{t=1}^{T} x_t + \hat{\beta}\bar{x}\sum_{t=1}^{T} x_t - \hat{\beta}\sum_{t=1}^{T} x_t^2 = 0$$

$$\sum_{t=1}^{T} x_t y_t - T\bar{y}\bar{x} + \hat{\beta}T\bar{x}^2 - \hat{\beta}\sum_{t=1}^{T} x_t^2 = 0$$

Rearranging for $\hat{\beta}$,

$$\hat{\beta}\left(T\bar{x}^2 - \sum_{t=1}^{T} x_t^2\right) = T\bar{y}\bar{x} - \sum_{t=1}^{T} x_t y_t$$

Therefore the coefficient estimators for the slope β and the intercept α are the minimisers of L, given by

$$\hat{\beta} = \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2} = \frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sum (x_t - \bar{x})^2} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

We can rewrite $\hat{\beta} = \frac{\frac{1}{(T-1)} \sum (x_t - \bar{x})(y_t - \bar{y})}{\frac{1}{(T-1)} \sum (x_t - \bar{x})^2}$, which is equivalent to the sample covariance between x and y divided by the sample variance of x.

This method of finding the optimum is known as ordinary least squares (OLS) estimation.

Least squares line

The least squares line is given by

$$\begin{aligned}
\hat{y}_t &= \hat{\alpha} + \hat{\beta} x_t = \bar{y} + \hat{\beta}(x_t - \bar{x}) \\
&= \bar{y} + \frac{\frac{1}{(T-1)} \sum (x_t - \bar{x})(y_t - \bar{y})}{\frac{1}{(T-1)} \sum (x_t - \bar{x})^2}(x_t - \bar{x}) \\
&= \bar{y} + \frac{s_{xy}}{s_x^2}(x_t - \bar{x})
\end{aligned}$$

where s_{xy} is the sample covariance between x and y and s_x^2 is the sample variance of x.

Interpretation

The coefficient estimate for β, $\hat{\beta}$, is interpreted as saying that, if x increases by 1 unit, y will be expected, everything else being equal, to increase by $\hat{\beta}$ units.

$\hat{\alpha}$, the intercept coefficient estimate, is interpreted as the value that would be taken by the dependent variable y if the independent variable x took a value of zero.

Non-linear models with OLS

In order to use OLS, a model that is linear is required. More specifically, the model must be linear in the parameters (α and β).

Models that are not linear in the variables can often be made to take a linear form by applying a suitable transformation or manipulation.

(*i*) Exponential regression model : $Y_t = AX_t^{\beta} e^{\epsilon_t}$
Taking logarithms, we have $\ln Y_t = \ln(A) + \beta \ln X_t + \epsilon_t$. Let $\alpha = \ln(A)$, $y_t = \ln Y_t$ and $x_t = \ln X_t$

$$y_t = \alpha + \beta x_t + \epsilon_t$$

(*ii*)

$$y_t = \alpha + \frac{\beta}{z_t} + \epsilon_t$$

The regression can be estimated using OLS by setting $x_t = \frac{1}{z_t}$.

Non-linear models

On the other hand, some models are intrinsically non-linear, e.g.

$$y_t = \alpha + \beta x_t^{\gamma} + \epsilon_t$$

Such models might be estimated using a non-linear method.

Classical linear regression model

The model $y_t = \alpha + \beta x_t + \epsilon_t$ with the assumptions

(1) The errors have zero mean, i.e. $E(\epsilon_t) = 0$.

(2) The variance of the errors is constant and finite, i.e. $\mathrm{var}(\epsilon_t) = \sigma^2 < \infty$.

(3) The errors are linearly independent of one another, i.e. $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$.

(4) There is no relationship between the error and corresponding x variate, i.e. $\mathrm{cov}(\epsilon_t, x_t) = 0$.

(5) ϵ_t is normally distributed, i.e. $\epsilon_t \sim N(0, \sigma^2)$.

Assumption (5) is required to make valid inferences about the population parameters (the actual α and β) from the sample parameters ($\hat{\alpha}$ and $\hat{\beta}$) estimated using a finite amount of data.

Estimators

With these assumptions, another way to find a estimator for β is to take the covariance with x_t on both sides of the model.

$$\begin{aligned} y_t &= \alpha + \beta x_t + \epsilon_t \\ \mathrm{cov}(x_t, y_t) &= \mathrm{cov}(x_t, \alpha + \beta x_t + \epsilon_t) \\ \mathrm{cov}(x_t, y_t) &= \mathrm{cov}(x_t, \beta x_t) \\ \beta &= \frac{\mathrm{cov}(x_t, y_t)}{\mathrm{cov}(x_t, x_t)} \end{aligned}$$

Then, again, we have $\hat{\beta} = \frac{\frac{1}{(T-1)}\sum(x_t - \bar{x})(y_t - \bar{y})}{\frac{1}{(T-1)}\sum(x_t - \bar{x})^2}$.

Properties of the OLS estimator

Under assumptions (1)-(4) listed above, the OLS estimator can be shown to have the desirable properties that it is consistent, unbiased and efficient.

- Consistency
- Unbiasedness
- Efficiency

Estimating the variance of the error term (σ^2)

The variance of a random variable ϵ_t is given by

$$\mathrm{var}(\epsilon_t) = \sigma^2 = E[(\epsilon_t - E(\epsilon_t))^2]$$

Under Assumption (1) above, it reduces to $\mathrm{var}(\epsilon_\mathrm{t}) = \sigma^2 = \mathrm{E}[\epsilon_\mathrm{t}^2]$.
To estimate $E[\epsilon_t^2]$, the sample counterpart to ϵ_t, $\hat{\epsilon}_t$ is used

$$\hat{\sigma}^2 = \frac{1}{T} \sum \hat{\epsilon}_t^2$$

An unbiased estimator is $\hat{\sigma}^2 = \frac{1}{T-2} \sum \hat{\epsilon}_t^2$, where T is the number of observations and 2 is the number of model parameters (i.e. α and β).

Standard error of the regression

$\hat{\sigma} = \sqrt{\frac{1}{T-2}\sum \hat{\epsilon}_t^2}$ is known as the standard error of the regression.

It is sometimes used as a broad measure of the fit of the regression equation.

Everything else being equal, the smaller this quantity is, the closer is the fit of the line to the actual data.

Standard errors of the estimators

Standard errors are used as measure of the reliability or precision of the estimators ($\hat{\alpha}$ and $\hat{\beta}$). It's to have an idea of how 'good' these estimates of α and β.
Given Assumptions (1)-(4) above, valid estimators of the standard errors can be shown to be given by

$$\mathrm{SE}(\hat{\alpha}) = \hat{\sigma}\sqrt{\frac{\sum x_t^2}{T\sum(x_t-\bar{x})^2}} = \hat{\sigma}\sqrt{\frac{\sum x_t^2}{T\left(\sum x_t^2 - T\bar{x}^2\right)}} \quad (5)$$

$$\mathrm{SE}(\hat{\beta}) = \hat{\sigma}\sqrt{\frac{1}{\sum(x_t-\bar{x})^2}} = \hat{\sigma}\sqrt{\frac{1}{(T-1)s_x^2}} \quad (6)$$

Testing single hypothesis : t-test

Hypothesis testings : What can we say about the population (true) values based on the estimated regression parameters from the sample data ?

There are always two hypotheses that go together, known as the null hypothesis (denoted by H_0) and the alternative hypothesis (denoted by H_a).

Example

Given the regression results, it is of interest to test the hypothesis that the true value of β is in fact 0.5 :

$$H_0 : \beta = 0.5, \quad H_a : \beta \neq 0.5$$

This would be known as a two-sided test, since the outcomes are both $\beta < 0.5$ and $\beta > 0.5$ under H_a.

Sometimes, some prior information may suggest that $\beta > 0.5$ would be expected rather that $\beta < 0.5$, then a one-sided test would be :

$$H_0 : \beta = 0.5, \quad H_a : \beta > 0.5$$

When $\beta < 0.5$ is expected, another one-sided test would be :

$$H_0 : \beta = 0.5, \quad H_a : \beta < 0.5$$

Test statistics

In very general terms, if the estimated value is a long way away from the hypothesised value, the null hypothesis is likely to be rejected ; if the value under the null hypothesis and the estimated value are close to one another, the null hypothesis is less likely to be rejected.

Under Assumption (5), it can be shown that the coefficient estimates will also be normally distributed,

$$\hat{\alpha} \sim N(\alpha, \mathrm{var}(\hat{\alpha})) \quad \text{and} \quad \hat{\beta} \sim N(\beta, \mathrm{var}(\hat{\beta}))$$

When the assumption (5) doesn't hold, the coefficient estimates still follow a normal distribution if all the other assumptions hold and the sample size is sufficiently large.

Thus we have

$$\frac{\hat{\alpha}-\alpha}{\sqrt{\mathrm{var}(\hat{\alpha})}} \sim N(0,1) \quad \text{and} \quad \frac{\hat{\beta}-\beta}{\sqrt{\mathrm{var}(\hat{\beta})}} \sim N(0,1)$$

Since the standard errors $\mathrm{var}(\hat{\alpha})$ and $\mathrm{var}(\hat{\beta})$ are unknown, one should replace them by the sample estimated version $SE(\hat{\alpha})$ and $SE(\hat{\beta})$.

Then the statistics follow a t-distribution with $T-2$ degrees of freedom rather that a normal distribution

$$\frac{\hat{\alpha}-\alpha}{SE(\hat{\alpha})} \sim t_{T-2} \quad \text{and} \quad \frac{\hat{\beta}-\beta}{SE(\hat{\beta})} \sim t_{T-2}$$

These are called the t-test statistics.

As in Lecture 2, we can use the confidence interval to conduct the tests.

For instance, the null hypothesis that $\beta = \beta^*$ will not be rejected if the test statistic lies within the non-rejection region

$$-t_{1-\alpha/2} \leq \frac{\hat{\beta} - \beta^*}{SE(\hat{\beta})} \leq t_{1-\alpha/2}$$

which is equivalent to

$$-t_{1-\alpha/2} \cdot SE(\hat{\beta}) \leq \hat{\beta} - \beta^* \leq t_{1-\alpha/2} \cdot SE(\hat{\beta})$$

i.e. one would not reject if

$$-t_{1-\alpha/2} \cdot SE(\hat{\beta}) + \beta^* \leq \hat{\beta} \leq t_{1-\alpha/2} \cdot SE(\hat{\beta}) + \beta^*$$

Significance

If the null hypothesis is rejected at the 5% level, it would be said that the result of the test is 'statistically significant'.

If the null hypothesis is not rejected, it would be said that the result of the test is 'not significant', or that it is 'insignificant'.

Finally, if the null hypothesis is rejected at the 1% level, the result is termed 'highly statistically significant'.

The exact significance level p-value

The exact significance level is also commonly known as the p-value. If the test statistic is 'large' in absolute value, the p-value will be small, and vice versa.

In fact, the null hypothesis is rejected if the p-value is smaller than the significance level (α).

p-values are almost always provided automatically by software packages.

Review : p-value

There are two possible errors that could be made :

1. Rejecting H_0 when it was really true ; this is called a type I error.
2. Not rejecting H_0 when it was false ; this is called a type II error.

Informally, the p-value is also often referred to as the probability of making type I error.

Thus, for example, if a p-value of 0.05 or less leads the researcher to reject the null (equivalent to a 5% significance level), this is equivalent to saying that if the probability of incorrectly rejecting the null hypothesis is more than 5%, do not reject it.

t-ratio test

If the test is

$$H_0 : \beta = 0, \quad H_a : \beta \neq 0$$

this is known as t-ratio test. Here $\beta^* = 0$, then

$$\text{test statistic} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

This ratio is known as the *t*-ratio.

Multiple linear regression

It is very easy to generalise the simple linear regression model to the multiple case :

$$y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t = \mathbf{x}_t^{'} \boldsymbol{\beta} + \varepsilon_t, \; t = 1, ..., T,$$

where $\boldsymbol{\beta}$ is $(k+1) \times 1$ vector. So the variables $\{x_1, \cdots x_k\}$ are a set of k explanatory variables which influence y.

In the simple linear regression model, there are 2 regressors (1 and x), while in the above setting, there are $k+1$ regressors (1, and $x_1, \cdots, x_k$).

Centering the explanatory variables

β_0 is the expected value of y when all of the explanatory variables are equal to 0.

However, frequently, 0 is outside the range of some explanatory variables, making the interpretation of β_0 of little real interest

In practice, we can center the explanatory variables.

If $x_{i,1}, \cdots, x_{i,T}$ are the values of the ith explanatory variable and $\bar{x}_i$ is their mean, then $(x_{i,1} - \bar{x}_i), \cdots, (x_{i,T} - \bar{x}_i)$ are values of the centered explanatory variable.

If all explanatory variables are centered, then β_0 is the expected value of y when all of the explanatory variables are equal to their mean. This gives β_0 an interpretable meaning.

Matrix notation

In matrix notation we have $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,where $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are $T \times 1$ vectors, and $\mathbf{x}$ is a $(T \times (k+1))$ matrix.

For example, a simple linear regression can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{1T} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix}$$

OLS estimations

Minimise the RSS $L = \sum_{t=1}^{T} \hat{\varepsilon}_t^2 = \sum_{t=1}^{T} (y_t - \mathbf{x}_t \hat{\beta})^2$.
Write L in matrix notation

$$L = \hat{\epsilon}' \hat{\epsilon} = (\mathbf{y} - \mathbf{x}\hat{\beta})'(\mathbf{y} - \mathbf{x}\hat{\beta}) = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{x}'\mathbf{y} - \mathbf{y}'\mathbf{x}\hat{\beta} + \hat{\beta}'\mathbf{x}'\mathbf{x}\hat{\beta} \quad (7)$$

We can check easily that $\beta' x' y = y' x \beta$, thus (7) can be written

$$L = \mathbf{y}'\mathbf{y} - 2\hat{\beta}'\mathbf{x}'\mathbf{y} + \hat{\beta}'\mathbf{x}'\mathbf{x}\hat{\beta}$$

The first order condition for the minimisation of L is

$$\frac{\partial L}{\partial \hat{\beta}} = -2\mathbf{x}'\mathbf{y} + 2\mathbf{x}'\mathbf{x}\hat{\beta} = 0 \text{ which is } \mathbf{x}'\mathbf{y} = \mathbf{x}'\mathbf{x}\hat{\beta}.$$

Pre-multiplying both sides of the above equation by the inverse of $\mathbf{x}'\mathbf{x}$, the vector of OLS coefficient estimates is given by

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

Testing multiple hypotheses : the F-test

The t-test was used to test single hypotheses, i.e. hypotheses involving only one coefficient.

To test more than one coefficient simultaneously, F-test is employed. Under the F-test framework, two regressions are required, known as the unrestricted and the restricted regressions.

The unrestricted regression is the one in which the coefficients are freely determined by the data, as has been constructed previously. The restricted regression is the one in which the coefficients are restricted, i.e. the restrictions are imposed on some β's.

F-test statistic

The F-test statistic for testing multiple hypotheses is given by

$$F-\text{test statistic} = \frac{\mathrm{RRSS} - \mathrm{URSS}}{\mathrm{URSS}} \times \frac{\mathrm{T} - \mathrm{k} - 1}{\mathrm{m}} \tag{8}$$

where

$$\begin{aligned} URSS &= \text{residual sum of squares from unrestricted regression} \\ RRSS &= \text{residual sum of squares from restricted regression} \\ m &= \text{number of restrictions} \\ T &= \text{number of observations} \\ k+1 &= \text{number of regressors in unrestricted regression} \end{aligned}$$

It can be shown that $URSS \leq RRSS$.

Examples

Informally, the number of restrictions can be seen as 'the number of equality signs under the null hypothesis'. For example,

H_0 :	Number of restriction m
$\beta_1 + \beta_2 = 1$	1
$\beta_1 = 1$ and $\beta_2 = 0$	2
$\beta_1 = 0, \beta_2 = 0$ and $\beta_3 = 0$	3

"The" regression F-statistic

With softwares (e.g. R), in general, the F-statistic given in the output of the regression function tests the null hypothesis that all of the coefficients except the intercept coefficient are zero.
For example, if the model is,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

then the null hypothesis tested by F-statistic is

$$\beta_1 = 0 \text{ and } \beta_2 = 0 \text{ and } \beta_3 = 0,$$

and the alternative hypothesis is

$$\beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \beta_3 \neq 0,$$

If this null hypothesis cannot be rejected, it would imply that none of the explanatory variables in the model was able to explain variations in y.

Ideas of F-test

(*i*) If, after imposing constraints on the model, a residual sum of squares results that is not much higher than the unconstrained model's residual sum of squares, it would be concluded that the restrictions were supported by the data.

(*ii*) If the residual sum of squares increased considerably after the restrictions were imposed, it would be concluded that the restrictions were not supported by the data and therefore that the hypothesis should be rejected.

The test statistic follows the F-distribution under the null hypothesis, more precisely, $F_{m,T-k-1}$.

Single hypotheses involving one coefficient can be tested using a t- or an F-test (they always give the same conclusion, see definitions in Lecture 1), but multiple hypotheses can be tested only using an F-test.

It is not possible to test hypotheses that are not linear or that are multiplicative using this framework - for example, $H_0 : \beta_1\beta_2 = 3$, or $H_0 : \beta_1^2 = 4$ cannot be tested.

Goodness of fit statistics

Goodness of fit statistics are used to test how well the estimated regression model fits the data.

OLS selected the coefficient estimates that minimised the RSS, so the lower was the minimised value of the RSS, the better the model fitted the data.

Goodness of fit statistics : R^2

A scaled version of the RSS is usually employed. The most common goodness of fit statistic is known as R^2.

$$Total\ SS(TSS) = Explained\ SS(ESS) + RSS$$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t (y_t - \hat{y}_t)^2$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2 can be understood as the portion of the total fluctuation of the dependent variable, y, explained by the regression relation.

R^2 must always lie between zero and one. A higher R^2 implies, everything else being equal, that the model fits the data better.

Problems with R^2

However, there are a number of problems with R^2 as a goodness of fit measure :

(i) R^2 never falls if more regressors are added to the regression. It's impossible to use R^2 as a determinant of whether a given variable should be present in the model or not.

(ii) R^2 can take values of 0.9 or higher for time series regressions, and hence it is not good at discriminating between models.

In order to get around the problem (i), one can use the adjusted R^2, denoted by $\bar{R}^2$

$$\bar{R}^2 = 1 - \left[\frac{T-1}{T-k-1}(1-R^2) \right] \tag{9}$$

If an extra regressor (explanatory variable) is added to the model, k increases and unless R^2 increases by a more than off-setting amount, $\bar{R}^2$ will fall.

The rule is : include the variable if $\bar{R}^2$ rises and do not include it if $\bar{R}^2$ falls.

$\bar{R}^2$ is always smaller than R^2. Also, $\bar{R}^2$ may take negative values, even with an intercept in the regression, if the model fits the data very poorly.

Model selection criteria

When there are many potential explanatory variables, often we wish to find a subset of them that provides a good regression model.

- For linear regression, AIC (Akaike's information criterion) is

$$AIC = T\log(\hat{\sigma}^2) + 2(1+k)$$

  where $1+k$ is the number of regressors.
- BIC (Bayesian IC) replaces $2(1+k)$ in AIC by $\log(T)(1+k)$.
- Suppose there are M explanatory variables. Let $\hat{\sigma}^2_M$ be the estimate of σ^2 using all of them, and let $RSS(p)$ be the residual sum of squares for a model with some subset of only $p \leq M$ of the predictors.

$$C_p = \frac{RSS(p)}{\hat{\sigma}^2_M} - T + 2(p+1)$$

- Ajusted R^2 ($\bar{R}^2$)

Model selection criteria

With C_p, AIC, and BIC, smaller values are better, but for adjusted R^2 ($\bar{R}^2$), larger values are better.

Linear Market Model

Linear market model :

$$r_t = \alpha + \beta r_{m,t} + \epsilon_t,$$

where $\{r_t\}$ are the log returns of a single stock (e.g. GM) and $\{r_{m,t}\}$ the log returns of an index (e.g. S&P composite index).

```
> gm=log(da[,2]+1) ; sp=log(da[,3]+1) #compute log returns
> m1=lm(gm ~ sp) ; summary(m1) #fit the market model
Call : lm(formula = gm ~ sp)
Coefficients :
```

	Estimate	*Std.Error*	*t value*	$Pr(>\|t\|)$
(*Intercept*)	−0.004861	0.003434	−1.415	0.158
sp	1.072508	0.077177	13.897	$< 2e-16 ***$

Residual standard error : 0.07652 on 500 degrees of freedom
Multiple R-squared : 0.2786, Adjusted R-squared : 0.2772

Nonlinear Market Model

Simple nonlinear market model :

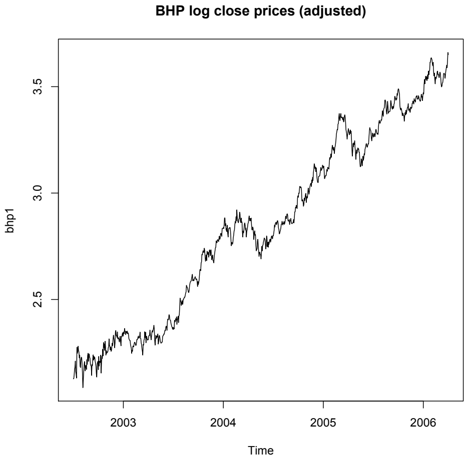
$$\begin{aligned} r_t &= \alpha_1 + \beta_1 r_{m,t} + \epsilon_t, \text{ if } r_{m,t} \leq 0, \\ &= \alpha_2 + \beta_2 r_{m,t} + \epsilon_t, \text{ if } r_{m,t} > 0. \end{aligned}$$

```
> idx=c(1 :502)[sp <= 0] # locate the non-positive market returns.
> nsp=rep(0,502) ; nsp[idx]=sp[idx] ; c1=rep(0,502) ; c1[idx]=1
> m2=lm(gm~ c1+sp) ; m3=lm(gm~nsp+sp)
> m4=lm(gm~ sp+c1+nsp) ; summary(m4)
```

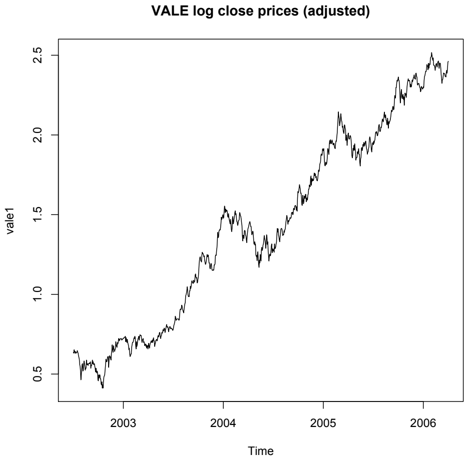
Coefficients :

	Estimate	*Std.Error*	*t value*	*Pr(> |t|)*
(Intercept)	−0.007778	0.007369	−1.055	0.2917
sp	1.041129	0.176838	5.887	7.21e − 09 * **
c1	0.020713	0.010550	1.963	0.0502.
nsp	0.387630	0.236399	1.640	0.1017

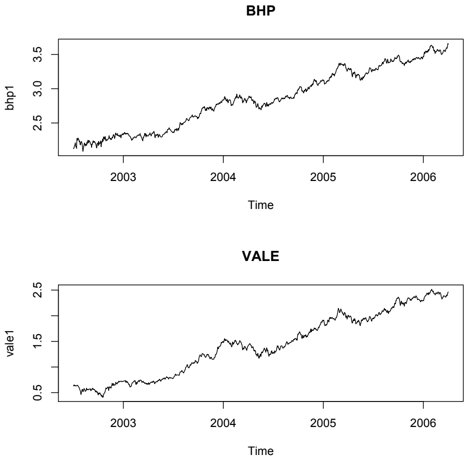
Log prices (d) of BHP (01/07/2002 - 31/03/2006)



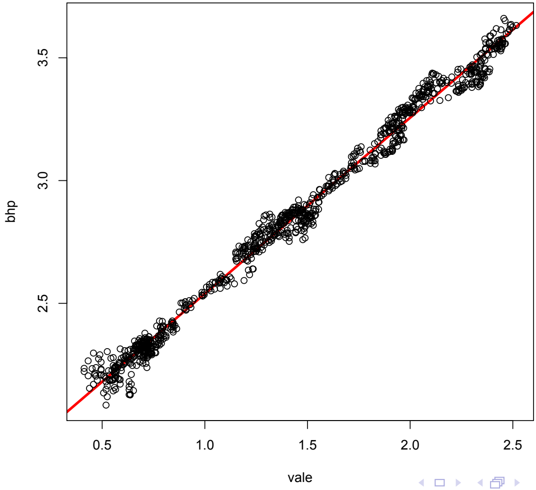
Log prices of VALE (01/07/2002 - 31/03/2006)



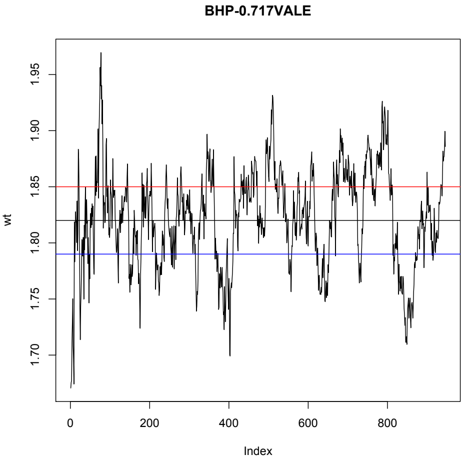
Log prices (daily) (01/07/2002 - 31/03/2006)



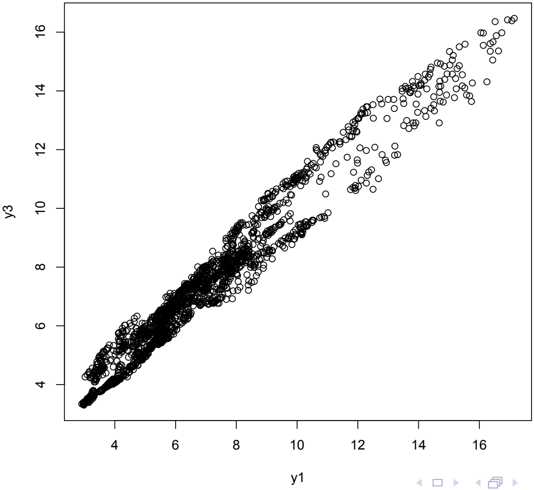
Log prices (daily) (01/07/2002 - 31/03/2006)



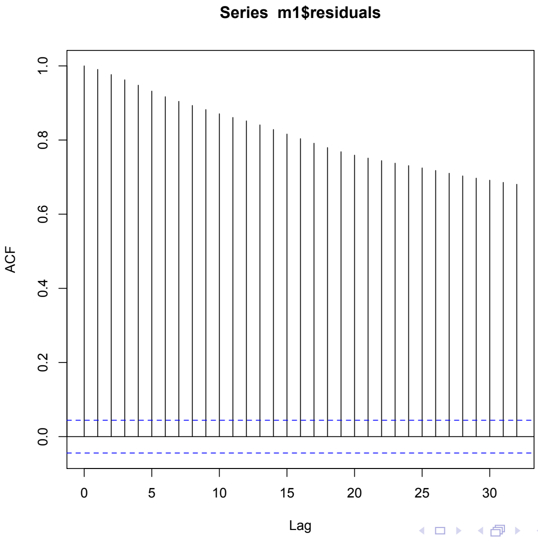
Pairs trading of BHP and VALE : Linear regression



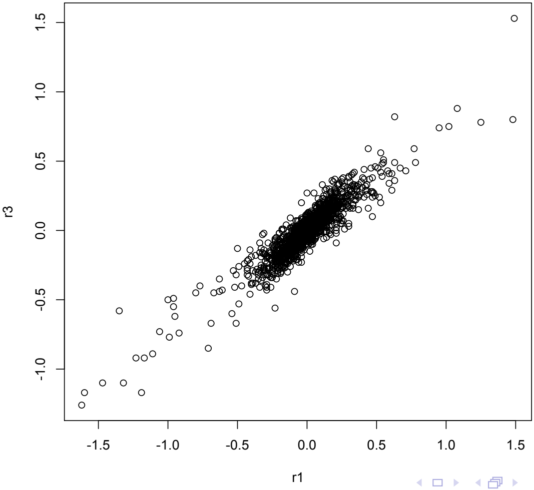
U.S. weekly interest rates : y1 y3



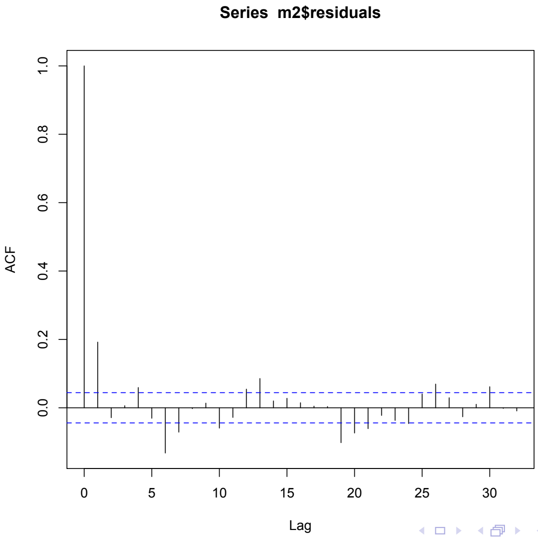
U.S. weekly interest rates : Regression y1 y3



U.S. weekly interest rates : r1 r3



U.S. weekly interest rates : Regression r1 r3



Thank you for your attention.