Introduction to Econometrics

Hannu Kahra <hannu.kahra@oulu.fi>

February 2013

Preamble

Two of the cornerstones of econometrics are the so-called **linear regression model** and the **ordinary least squares** (**OLS**) estimation method.

1 An Introduction to Linear Regression

1.1 Ordinary Least Squares as an Algebraic Tool

1.1.1 Ordinary least squares

Suppose we have a sample with N observations on individual wages and some background characteristics. Our main interest lies in the question as to how in this sample wages are related to the other observations. In this example, wages are *functions* of the underlying characteristics. Similarly, equity returns are functions of companany characteristics, e.g. the size of the company, the book to market ratio, dividend yield, price-earnings ratio, etc.

Let's denote wages by y and the $K-1$ characteristics by $x_2, \ldots, x_K$. Now, we may ask the question: which linear combination of $x_2, \ldots, x_K$ and a constant gives a good approximation (*fit*) of y? To answer the question, first consider an arbitrary linear combination, including a constant, which can be written as

$$\tilde{\beta}_1 + \tilde{\beta}_2 x_2 + \ldots + \tilde{\beta}_K x_K,$$

where $\tilde{\beta}_1, \ldots, \tilde{\beta}_K$ are constants (*parameters*) to be choosen. Let's index the observations by i such that $i = 1, \ldots, N$. Now, the difference between an observed value y_i and its linear approximation (fit) is

$$y_i - \left[\tilde{\beta}_1 + \tilde{\beta}_2 x_{i2} + \ldots + \tilde{\beta}_K x_{iK}\right]. \tag{1}$$

We simplify using vector notation. First, we collect the x-values for individual i in a vector x_i, which includes the constant. That is

$$x_i = (1\ x_{i2}\ x_{i3} \ldots x_{iK})'.$$

Collecting the $\tilde{\beta}$coefficients in a K-dimensional vector $\tilde{\beta} = \left(\tilde{\beta}_1, \ldots, \tilde{\beta}_K\right)'$ we can briefly write (1) as

$$y_i - x_i'\tilde{\beta}.$$

Clearly, we would like to choose values for $\tilde{\beta}_1, \ldots, \tilde{\beta}_K$ such that these differences are small. Although different measures can be used to define what we mean by 'small', the most common approach is to choose $\tilde{\beta}$ such that the sum of squared differences is as small as possible. In this case we determine $\tilde{\beta}$ to minimize the objective function $S\left(\tilde{\beta}\right)$:

$$\min_{\tilde{\beta}} S\left(\tilde{\beta}\right) = \sum_{i=1}^{N} \left(y_i - x_i'\tilde{\beta}\right)^2. \tag{2}$$

That is, we minimize the sum of squared approximation errors. This approach is referred to as the **ordinary least squares** or **OLS** approach. Taking squares makes sure that positive and negative deviations do not cancel out when taking the summation.

To solve the minimization problem, we consider the first-order conditions, obtained by differentiating $S\left(\tilde{\beta}\right)$ with respect to the vector $\tilde{\beta}$. This gives the following system of K conditions:

$$-2\sum_{i=1}^{N} x_i \left(y_i - x_i^{'}\tilde{\beta}\right) = 0$$

or

$$\left(\sum_{i=1}^{N} x_i x_i^{'}\right)\tilde{\beta} = \sum_{i=1}^{N} x_i y_i.$$

These equations are sometimes referred to as normal equations. As this system has K unknows, one can obtain a unique solution to $\tilde{\beta}$ provided that the symmetrix matrix $\sum_{i=1}^{N} x_i x_i^{'}$ which contains the sum of squares and cross products of the regressors x_i, can be inverted. For the moment, we shall assume that this is the case. The solution to the minimazation problem, which we shall denote by b (or usually by $\hat{\beta}$), is then given by

$$b = \left(\sum_{i=1}^{N} x_i x_i^{'}\right)^{-1} \sum_{i=1}^{N} x_i y_i. \tag{3}$$

By checking the second-order conditions, it is easily verified that b indeed corresponds to a minimum of (2).

The resulting linear combination of x_i is thus given by

$$\hat{y}_i = x_i^{'} b,$$

which is the **best linear approximation** of y from $x_2, \ldots, x_K$ and a constant. The phrase 'best' refers to the fact that the sum of squared differences between the observed values y_i and fitted values $\hat{y}_i$ is minimal for the least squares solution b.

In deriving the linear approximation, we have not used any economic or statistical theory. It is simply an algebraic tool and it holds irrespective of the way the data are generated. That is, given a set of variables we can always determine the best linear approximation of one variable using the other variables.

Defining a **residual** e_i as the difference between the observed and the approximated value, $e_i = y_i - \hat{y}_i = y_i - x_i^{'} b$, we can decompose the observed y_i as

$$y_i = \hat{y}_i + e_i = x_i^{'} b + e_i.$$

This allows us to write the minimum value for the objective function as

$$S(b) = \sum_{i}^{N} e_i^2,$$

which is referred to as the **residual sum of squares**.

1.1.2 Simple linear regression

In the case where $K = 2$ we only have one regressor and a constant. In this case, the observations (y_i, x_i) can be drawn in a two-dimensional graph with x-values on the horizontal axis and y-values on the vertical one. This is done for the US National Longitudial Survey (NLS) that relates to 1987, and we have a sample of 3294 young working individuals, of which 1569 are female.[1]

```
> my.data <- read.table("H:/721364P/Rdata/wages1.dat", header=T)
> head(my.data)
```

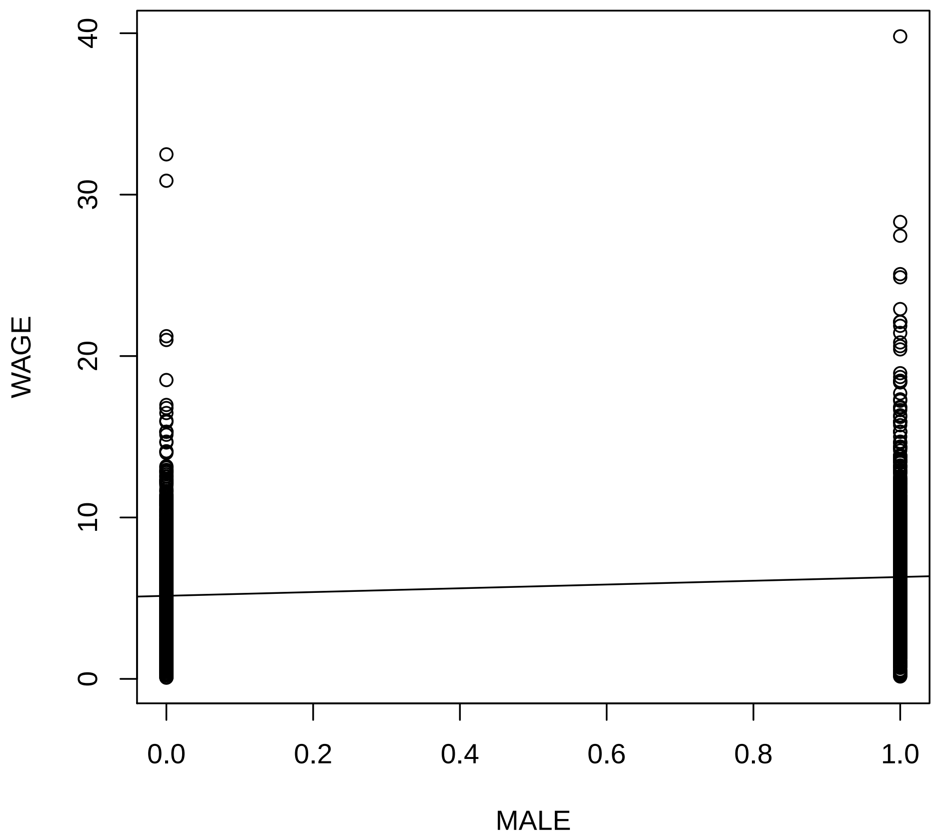
[1] The data for this example are available as WAGES1.DAT, and it is taken from Marno Verbeek (2012), A Guide to Modern Econometrics, 4th edition, John Wiley & Sons.

```
  EXPER MALE SCHOOL     WAGE
1     9    0     13 6.315296
2    12    0     12 5.479770
3    11    0     11 3.642170
4     9    0     14 4.593337
5     8    0     14 2.418157
6     9    0     14 2.094058

> tail(my.data)

     EXPER MALE SCHOOL     WAGE
3289     5    1      8 5.512004
3290     6    1      9 4.287114
3291     5    1      9 7.145190
3292     6    1      9 4.538784
3293    10    1      8 2.909113
3294     7    1      7 4.153974

> attach(my.data)
> summary(my.data)

     EXPER             MALE             SCHOOL            WAGE
 Min.   : 1.000   Min.   :0.0000   Min.   : 3.00   Min.   : 0.07656
 1st Qu.: 7.000   1st Qu.:0.0000   1st Qu.:11.00   1st Qu.: 3.62157
 Median : 8.000   Median :1.0000   Median :12.00   Median : 5.20578
 Mean   : 8.043   Mean   :0.5237   Mean   :11.63   Mean   : 5.75759
 3rd Qu.: 9.000   3rd Qu.:1.0000   3rd Qu.:12.00   3rd Qu.: 7.30451
 Max.   :18.000   Max.   :1.0000   Max.   :16.00   Max.   :39.80892

> m1 <- lm(WAGE~MALE)
> summary(m1)

Call:
lm(formula = WAGE ~ MALE)

Residuals:
   Min     1Q Median     3Q    Max
-6.160 -2.102 -0.554  1.487 33.496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.14692    0.08122   63.37   <2e-16 ***
MALE         1.16610    0.11224   10.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.217 on 3292 degrees of freedom
Multiple R-squared: 0.03175,        Adjusted R-squared: 0.03145
F-statistic: 107.9 on 1 and 3292 DF,  p-value: < 2.2e-16
```

The best linear approximation of y (salary) from x (gender) and a constant is obtained by minimizing the sum of squared residuals, which – in the two-dimensional case – equal the vertical distances between an observation and the fitted value. All fitted values are on a straight line, the **regression line**.

Because a 2×2 matrix can be inverted analytically, we can derive solutions for b_1 and b_2 in this special case from the general expression for b above. Equivalently, we can minimize the residual sum of squares

Figure 1: Simple linear regression: fitted line and observation points



with respect to the unkowns directly. Thus we have

$$S\left(\tilde{\beta}_1, \tilde{\beta}_2\right) = \sum_{i=1}^{N} \left(y_i - \tilde{\beta}_i - \tilde{\beta}_2 x_i\right)^2.$$

The basic elements in the derivation of the OLS solutions are the first-order conditions

$$\frac{\partial S\left(\tilde{\beta}_1, \tilde{\beta}_2\right)}{\partial \tilde{\beta}_1} = -2\sum_{i=1}^{N} \left(y_i - \tilde{\beta}_i - \tilde{\beta}_2 x_i\right) = 0 \tag{4}$$

$$\frac{\partial S\left(\tilde{\beta}_1, \tilde{\beta}_2\right)}{\partial \tilde{\beta}_2} = -2\sum_{i=1}^{N} x_i \left(y_i - \tilde{\beta}_i - \tilde{\beta}_2 x_i\right) = 0 \tag{5}$$

From (4) we can write

$$b_1 = \frac{1}{N}\sum_{i=1}^{N} y_i - b_2 \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{y} - b_2 \bar{x}, \tag{6}$$

where $b_2$is solved from (5) and (6). First, from (5) we write

$$\sum_{i=1}^{N} x_i y_i - b_1 \sum_{i=1}^{N} x_i - \left(\sum_{i=1}^{N} x_i^2\right) b_2 = 0$$

and then substitute (6) to obtain

$$\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y} - \left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right) b_2 = 0$$

such that we can solve for the slope coefficient b_2 as

$$b_2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}.$$

By dividing both numerator and denominator by $N-1$ it appears that the OLS solution b_2 is the ratio of the sample covariance between x and y and the sample variance of x. From (6), the intercept is determined so as to make the average approximation error (residual) equal to zero.

1.1.3 Example: Individual wages

The following examples are based on a sample of individual wages with background characteristics, like gender, race and years of schooling. The average hourly wage rate in this sample equals \$6.31 for males and \$5.15 for females. Now suppose we try to approximate wages by a linear combination of a constant and a (binary) $0-1$ variable denoting whether the individual is male or not. That is, $x_i = 1$ if individual i is male and zero otherwise. Such a variable, which can only take on the values of zero and one, is called a **dummy variable**. Using the OLS approach the result is

$$\hat{y}_i = 5.15 + 1.17x_i.$$

This means that for females our best approximation is \$5.15 and for males it is \$5.15+\$1.17 = \$6.31. It is not a coincidence that these numbers are exactly equal to the sample means in the two subsamples. It is easily verified from the results above that

$$\begin{aligned} b_1 &= \bar{y}_i \\ b_2 &= \bar{y}_m - \bar{y}_f, \end{aligned}$$

where $\bar{y}_m$ is the sample average of the wage for males, and $\bar{y}_f$ is the average for females.

1.1.4 Matrix notation

Using matrices, deriving the least squares solution is fasterm but it requires some knowledge of matrix differential calculus. We introduce the following notation:

$$X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N2} & \cdots & x_{NK} \end{pmatrix} = \begin{pmatrix} x_1' \\ \vdots \\ x_N' \end{pmatrix},\ y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}.$$

So, in the $N \times K$ matrix X the ith row refers to observation i, and the kth column refers to the kth explanatory variable (regressor).

The criterion to be minimized can be rewritten in matrix notation using the fact that the inner product of a given vector a with itself $(a'a)$ is the sum of its squared elements. That is,

$$S\left(\tilde{\beta}\right) = \left(y - X\tilde{\beta}\right)'\left(y - X\tilde{\beta}\right) = y'y - 2y'X\tilde{\beta} + \tilde{\beta}'X'X\tilde{\beta},$$

from which the least squares solution follows from differentiating with respect to $\tilde{\beta}$ and setting the result to zero:

$$\frac{\partial S\left(\tilde{\beta}\right)}{\partial \tilde{\beta}} = -2\left(X'y - X'X\tilde{\beta}\right) = 0. \tag{7}$$

Solving (7) gives the OLS solution

$$b = \left(X'X\right)^{-1} X'y$$

which is exactly the same as the one derived in (3). Here we assume that $X'X$ is invertible, i.e. that there is no exact (or perfect) **multicollinearity**.

As before, we can decompose y as

$$y = Xb + e,$$

where e is an N-dimensional vector of residuals.

1.2 The Linear Regression Model

Usually, economists want more than just finding the best linear approximation of one variable given a set of others. They want economic relationships that are more generally valid than the sample they happen to have. They want to draw conclusions about what happens if one of the variables actually changes. That is: they want to say something about things that are not observed (yet). In this case, we want the relationship that is found to be more than just a historical coincidence; it should reflect a fundamental relationship. To do this it is assumed that there is a general relationship that is valid for all possible observations from a well-defined population. Restricting attention to linear relationships, we specify a **statistical model** as

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots \beta_K x_{iK} + \varepsilon_i$$

or

$$y_i = x_i'\beta + \varepsilon_i, \tag{8}$$

where y_i and x_i are observable variables and ε_i is unobserved and referred to as an **error term** or disturbance term. The elements of β are unknown population parameters. The equality in (8) is supposed to hold for any possible observation, while we only observe a **sample** on N observations.

We shall consider this sample as one realization of all possible samples of size N that could have been drawn from the same population. In this way we can view y_i and ε_i (and often x_i) as **random variables**. Each observation corresponds to a realization of these random variables. Again we can use matrix notation and stack all observations to write

$$y = X\beta + \varepsilon, \tag{9}$$

where y and ε are N-dimensional vectors and X, as before, is of dimension $N \times K$. Equations (8) and (9) are population relationships, where β is a vector of unknown parameters characterizing the population.

We need to impose some assumptions to give the model a meaning. A common assumption is the expected value of ε_i given all the explanatory variables in x_i is zero, that is $E\{\varepsilon_i|x_i\} = 0$. Usually, people refer to this assumption by saying that the x variable is **exogenous**. Under this assumption it holds that

$$E\{y_i|x_i\} = x_i^{'}\beta,$$

so that the regression line $x_i^{'}\beta$ describes the conditional expectation of y_i given the values for x_i. The coefficients β_k measure how the expected value of y_i is affected if the value of x_k is changed, keeping the other elements in x_i constant. This is referred to as the **ceteris paribus** condition.

Now that our β coefficients have a meaning, we can try to use the sample (y_i, x_i) $i = 1, \ldots N$, to say something about them. The rule that says how a given sample is translated into an approximate value for β is referred to as an **estimator**. The result for a given sample is called an **estimate** (usually denoted as $\hat{\beta}$ or b). The *estimator* is a vector of random variables, because the sample changes, while the *estimate* is a vector of numbers. The most widely used estimator in econometrics is the **ordinary least squares** (**OLS**) estimator. The OLS estimator is given by

$$b = \left(\sum_{i=1}^{N} x_i x_i^{'}\right)^{-1} \sum_{i=1}^{N} x_i y_i.$$

1.3 Small Sample Properties of the OLS Estimator

1.3.1 The Gauss-Markov assumptions

Whether or not the OLS estimator b provides a good approximation to the unknown parameter vector β depends crucially upon the assumptions that are made about the distribution of ε_i and its relation to x_i. A standard case in which the OLS estimator has good properties is characterized by the Gauss-Markov conditions. They constitute a simple case in which the small sample properties of b are easily derived.

For the linear regression model, given by

$$y_i = x_i^{'}\beta + \varepsilon_i$$

the **Gauss-Markov conditions** are

$$E\{\varepsilon_i\} = 0, \text{ for } i = 1, \ldots, N \tag{10}$$

$$\{\varepsilon_1, \ldots, \varepsilon_N\} \text{ and } \{x_1, \ldots, x_N\} \text{ are independent} \tag{11}$$

$$V\{\varepsilon_i\} = \sigma^2, \text{ for } i = 1, \ldots, N \tag{12}$$

$$\text{cov}\{\varepsilon_i, \varepsilon_j\} = 0, \text{ for } i, j = 1, \ldots, N, \ i \neq j. \tag{13}$$

Assumption (10) says that the expected value of the error term is zero, which means that, *on average*, the regression line should be correct. Assumption (12) states that all error terms have the same variance, which is referred to as **homoskedasticity**, while assumption (13) imposes zero correlation between different error terms. This excludes any form of **autocorrelation**. Taken together, (10), (12) and (13) imply that the error terms are uncorrelated drawings from the distribution with expectation zero and constant variance σ^2.

1.3.2 Properties of the OLS estimator

Under assumptions (10)–(13), the OLS estimator b for β has several desirable properties. First of all, it is **unbiased**. This means that, in repeated sampling, we can expect that the OLS estimator is on average equal to the true (and unknown) value β. We formulate this as $E\{b\} = \beta$. It is instructive to see the proof:

$$\begin{aligned} E\{b\} &= E\left\{(X'X)^{-1} X'y\right\} = E\left\{\beta + (X'X)^{-1} X'\varepsilon\right\} \\ &= \beta + E\left\{(X'X)^{-1} X'\varepsilon\right\} = \beta. \end{aligned}$$

The latter step here is essential and it follows from

$$E\left\{\left(X'X\right)^{-1}X'\varepsilon\right\} = E\left\{\left(X'X\right)^{-1}X'\right\}E\{\varepsilon\} = 0,$$

because, from assumption (11), X and ε are independent and, from (10), $E\{\varepsilon\} = 0$.

In addition to knowing that we are, on average, correct, we would also like to make statements about how (un)likely it is to be far off in a given sample. This means we would like to know the distribution of b. First of all, the variance of b (conditional upon X) is given by

$$V\{b|X\} = \sigma^2\left(X'X\right)^{-1} = \sigma^2\left(\sum_{i=1}^{N} x_i x_i^{'}\right)^{-1}, \tag{14}$$

which, for simplicity, we shall denote by $V\{b\}$. The $K \times K$ matrix $V\{b\}$ is a variance-covariance matrix, containing the variances of $b_1, b_2, \ldots, b_K$ on the diagonal and their covariances as off-diagonal elements. The poof is fairly easy and goes as follows:

$$\begin{aligned} V\{b\} &= E\left\{(b-\beta)(b-\beta)'\right\} = E\left\{\left(X'X\right)^{-1}X'\varepsilon\varepsilon'X\left(X'X\right)^{-1}\right\} \\ &= \left(X'X\right)^{-1}X'\left(\sigma^2 I_N\right)X\left(X'X\right)^{-1} = \sigma^2\left(X'X\right)^{-1}, \end{aligned}$$

where I_N is the $N \times N$ identity matrix. To estimate the variance of b, $V\{b\}$, we have to replace the unknown error variance σ^2 with an estimate. An obvious candidate is the sample variance of the residuals $e_i = y_i - x_i^{'}b$, that is

$$\tilde{s}^2 = \frac{1}{N-1}\sum_{i=1}^{N} e_i^2$$

(recalling the average residual is zero). However, because e_i different from ε_i, this estimator is biased for σ^2.An unbiased estimator is given by

$$s^2 = \frac{1}{N-K}\sum_{i=1}^{N} e_i^2. \tag{15}$$

The estimator has a degrees of freedom correction as it divides by the number of observations minus the number of regressors (including the intercept). The variance of b can thus be estimated by

$$\hat{V}\{b\} = s^2\left(X'X\right)^{-1} = s^2\left(\sum_{i=1}^{N} x_i x_i^{'}\right)^{-1}.$$

The estimated variance of an element b_k is given by $s^2 c_{kk}$, where c_{kk} is the (k,k) element in $\left(\sum_i x_i x_i^{'}\right)^{-1}$. The square root of this estimated variance is usually referred to as the **standard error** of b_k. We denote it by $se(b_k)$. It is the *estimated* standard deviation of b_k and is a measure for the accuracy of the estimator. When the error terms are not homoskedastic and/or exhibit autocorrelation, the standard error of the OLS estimator b_k will have to be computed in a different way (to be discussed later).

Assumptions (10)–(13) state that the error term ε_i are mutually uncorrelated, are independent of X, have zero mean and have constant variance, but do not specify the shape of the distribution. For exact statistical inference from a given sample of N observations, explicit distributional assumptions have to be made. The most common assumption is that the errors are jointly normally distributed. In this case the uncorrelatednes of (13) is equivalent to independence of all error terms. The precise assumption is as follows:

$$\varepsilon \sim N\left(0, \sigma^2 I_N\right), \tag{16}$$

saying that the vector of error terms ε has a N-variate normal disribution with mean vector 0 and covariance matrix $\sigma^2 I_N$. An alternative way of formulating (16) is

$$\varepsilon \sim NID\left(0, \sigma^2\right), \tag{17}$$

which is a shorthand way of saying that the error terms ε_i are independent drawings from a normal distribution (n.i.d.) with mean zero and variance σ^2.

To make things simpler, let's consider the X matrix as fixed and deterministic or, alternatively, let's work conditionally upon the outcomes of X. Then the following result holds. Under assumptions (11) and (17) the OLS estimator b is normally distributed with mean vector β and covariance matrix $\sigma^2 \left(X'X\right)^{-1}$, i.e.

$$b \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right). \tag{18}$$

The result in (18) implies that each element in b is normally distributed, for example

$$b_k \sim N\left(\beta_k, \sigma^2 c_{kk}\right), \tag{19}$$

where, as before, c_{kk} is the (k, k) element in $\left(X'X\right)^{-1}$.These results provide the basis for statistical tests based upon the OLS estimator b.

1.3.3 Example: Individual wages (continued)

Let's now turn back to our wage example. We can formulate a (fairly trivial) statistical model as

$$wage_i = \beta_1 + \beta_2 male_i + \varepsilon_i,$$

where $wage_i$ denotes hourly wage rate for individual i and $male_i = 1$ if i is male and 0 otherwise. Imposing that $E\{\varepsilon_i\} = 0$ and $E\{\varepsilon_i|male\} = 0$ gives β_1 the interpretation of the expected wage rate for females, while $E\{wage_i|male = 1\} = \beta_1 + \beta_2$ is the expected wage rate for males. Thus β_2 is the expected wage differential between an arbitrary male and female. We can now say that our estimate of the expected wage differential β_2 between males and females is \$1.17 with a standard error of \$0.11. Combined with the normal distribution, this allows us to make statements about β_2. For example, we can test the hypothesis that $\beta_2 = 0$. If this hypothesis is true, the wage differential between males and females in our sample is nonzero only by chance. We'll discuss hypotheses testing shortly.

1.4 Goodness-of-fit

Having estimated a particular linear model, a natural question that comes up is: how well does the estimated regression line fit the observations? A popular measure for the goodness-of-fit is the proportion of the (sample) variance of y that is explained by the model. This variable is called the R^2 (R squared) and is defined as

$$R^2 = \frac{\hat{V}\left(\hat{y}_i\right)}{\hat{V}\left(y_i\right)} = \frac{1\left(N-1\right)\sum_{i=1}^{N}\left(\hat{y}_i - \bar{y}\right)^2}{1\left(N-1\right)\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}, \tag{20}$$

where $\hat{y}_i = x_i^{'} b$ and $\bar{y}_i = \frac{1}{N}\sum_i y_i$ denotes the sample mean of y_i. Note that $\bar{y}$ also corresponds to the sample mean of $\hat{y}_i$, because for the average observation $\bar{y} = \bar{x}'b$.

R^2 can be rewritten as

$$R^2 = 1 - \frac{\hat{V}\{e_i\}}{\hat{V}\{y_i\}} = 1 - \frac{\frac{1}{N-1}\sum_{i=1}^{N} e_i^2}{\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}. \tag{21}$$

In the exceptional cases where the model does not contain an intercapt term, the two expressions for R^2 are not equivalent. If there is no intercept, we apply the uncentered R^2 which is defines as

$$\text{uncentered } R^2 = \frac{\sum_i^N \hat{y}_i^2}{\sum_{i=1}^{N} y_i^2} = 1 - \frac{\sum_i^N e_i^2}{\sum_{i=1}^{N} y_i^2}. \tag{22}$$

Generally, the uncentered R^2 is higher than the standard R^2.

1.4.1 R2

Sometimes R^2 is interpreted as a measure of quality of the *statistical* model, while in fact it measures nothing more than the quality of the linear approximation. For later use, we'll present an alternative definition for R^2, which for OLS is equivalent to (20) and (21), and for any other estimator is guaranteed to be between zero and and one. It is given by

$$R^2 = \text{corr}^2\{y_i, \hat{y}_i\} = \frac{\left(\sum_{i=1}^{N}(y_i-\bar{y})(\hat{y}_i-\bar{y})\right)^2}{\left(\sum_{i=1}^{N}(y_i-\bar{y})^2\right)\left(\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2\right)}, \tag{23}$$

which denotes the squared (sample) correlation coefficient between the actual and fitted values. Written this way, the R^2 can be interpreted to measure how well the variation in $\hat{y}_i$ reflects the quality of the linear approximation and not necessarily that of the statistical model in which we are interested. As a result, the R^2 is typically not the most important aspect of our estimation result.

Another drawback of the R^2 is that it will never decrease if the number of regressors is increased, even if the additional variables have no real explanatory power. A common way to solve this is to correct the variance estimates in (21) . This gives the so-called **adjusted** R^2, or $\bar{R}^2$,defined as

$$\bar{R}^2 = 1 - \frac{1/(N-K)\sum_{i=1}^{N} e_i^2}{1/(N-1)\sum_{i=1}^{N}(y_i-\bar{y})^2}. \tag{24}$$

The goodness-of-fit measure has some punishment for the inclusion of additional explanatory variables in the model and therefore does not automatically increase when regressors are added to the model. In fact, it may decline when a variable is added to the set of regressors. Note that, in extreme cases, the $\bar{R}^2$ is strictly smaller than R^2 unless $K = 1$ and the model only includes an intercept.

1.5 Hypothesis Testing

Under the Gauss-Markov assumptions (10)-(13) and normality of the error term (17), we saw that the OLS estimator b has a normal distribution with mean β and covariance matrix $\sigma^2 (X'X)^{-1}$. We can use this result to develop tests for hypotheses regarding the unknown population parameter β. Starting from (19), it follows that the variable

$$z = \frac{b_k - \beta_k}{\sigma\sqrt{c_{kk}}} = \frac{b_k - \beta_k}{s.e.(b_k)}$$

has a standard normal distribution (i.e., a normal distribution with mean 0 and variance 1, $z \sim N(0,1)$). If we replace the unknown σ by its estimate s, this is no longer exactly true. It can be shown that the unbiased estimator s^2 defined in (15) is independent of b and has a Chi-squared distribution with $N-K$ degrees of freedom. In particular,

$$(N-K)s^2/\sigma^2 \sim \chi^2_{N-K}.$$

Consequently, the random variable

$$t_k = \frac{b_k - \beta_k}{s\sqrt{c_{kk}}} = \frac{b_k - \beta_k}{\widehat{s.e.(b_k)}}$$

is the ratio of a standard normal variable and the square root of an independent Chi-squared variable and therefore follows Student's t-distribution with $N-K$ degrees of freedom. The t-distribution is close to the standard normal distribution except that it has fatter tails, particularly when the number of degrees of freedom $N-K$ is 'small'. The larger the $N-K$, the more closely the t-distribution resembles the standard normal, and for sufficiently large $N-K$ the two distributions are identical.

1.5.1 A simple t-test

The result above can be used to construct test statistics and confidence intervals. The general idea of hypothesis testing is as follows. Starting from a given hypothesis, the **null hypothesis**, a test statistic is computed that has a known distribution *under the assumption that the null hypothesis is valid.* Next, it is decided whether the computed value of the test statistic is unlikely to come from this distribution, which indicates that the null hypothesis is unlikely to hold. Let's illustrate this with an example.

Suppose we have a null hypothesis that specifies the value of β_k, say $H_0 : \beta_k = \beta_k^0$, where β_k^0 is a specific value chosen by the researcher. If this hypothesis is true, we know that the statistic

$$t_k = \frac{b_k - \beta_k^0}{s.e.(b_k)} \tag{25}$$

has a t-distribution with $N - K$ degrees of freedom. If the null hypothesis is not true, the alternative hypothesis $H_1 : \beta_k \neq \beta_k^0$ holds. The quantity in (25) is a **test statistic** and it is computed from the estimate b_k, its standard error $s.e.(b_k)$, and the hypothesized value β_k^0 under the null hypothesis. If the test statistic realizes a value that is very unlikely under the null distribution, we reject the null hypothesis. In this case this means very large absolute values for t_k. To be precise, one rejects the null hypothesis if the probability of observing a value of $|t_k|$ or larger is smaller than a given significance level α, often 5%. From this, one can define the **critical values** $t_{N-K:\alpha/2}$ using

$$P\left\{|t_k| > t_{N-K:\alpha/2}\right\} = \alpha.$$

For $N - K$ not too small, these critical values are only slightly larger than those of the standard normal distribution, for which the two-tailed critical value for $\alpha = 0.05$ is 1.96. Consequently, at the 5% level the null hypothesis will be rejected if

$$|t_k| > 1.96.$$

The above test is referred to as a two-sided test since the alternative hypothesis allows for values of β_k on both sides of β_k^0. Occasionally, the alternative hypothesis is one-sided, for example: the expected wage for a man is larger than for a woman. Formally, we define the null hypothesis as $H_0 : \beta_k \leq \beta_k^0$ with alternative $H_1 : \beta_k > \beta$. Next, we consider the distribution of the test statistic t_k at the boundary of the null hypothesis (i.e., under $\beta_k = \beta_k^0$, as before) and we reject the null hypothesis if t_k is too large (note that large values for b_k lead to large values of t_k). Large negative values for t_k are compatible with the null hypothesis and do not lead to its rejection. Thus for this **one-sided test** the critical value is determined by

$$P\left\{t_k > t_{N-K:\alpha}\right\} = \alpha.$$

Using the standard normal approximation again, we reject the null hypothesis at the 5% level if

$$t_k > 1.64.$$

Regression packages (R, too) typically report the following t-value:

$$t_k = \frac{b_k}{s.e.(b_k)},$$

sometimes referred to as the t-ratio, which is the point estimate divided by its standard error. The t-ratio is the t-statistics one would compute to test the null hypothesis that $\beta_k = 0$, which may be a hypothesis that is of economic interest as well. If it is rejected, it is said that 'b_k differs significantly from zero', or the corresponding variable 'x_{ik} has statistically significant impact on y_i'. Often we simply say that (the effect of) 'x_{ik} is statistically significant'. Note that, if an economic variable is statistically significant, this does not necessarily imply that its impact is economically meaningful. Therefore, it is good practice to pay attention to the magnitude of the coefficients as well as to their statistical significance.

A confidence interval can be defined as the interval of all values for β_k^0 for which the null hypothesis $\beta_k = \beta_k^0$ is not rejected by the t-test. Loosely speaking, given the estimate b_k and its associated standard error, a confidence interval gives a range of values which are likely to contain the true value β_k. It is derived from the fact that the following inequalities hold with probability $1 - \alpha$:

$$-t_{K-N:\alpha/2} < \frac{b_k - \beta_k}{s.e.(b_k)} < t_{N-K:\alpha/2},$$

or

$$b_k - t_{N-K:\alpha/2} s.e.(b_k) < \beta_k < b_k + t_{N-K:\alpha/2} s.e.(b_k).$$

Consequently, using the standard normal approximation, a 95% condidence interval (setting $\alpha = 0.05$) for β_k is given by the interval

$$b_k - 1.96 \times s.e.(b_k), b_k + 1.96 \times s.e.(b_k).$$

In repeated sampling, 95% of those intervals will contain the true value β_k which is a fixed but unknown number.

1.5.2 Example: Individual wages (continued)

From the results of the previous example we can compute t-ratios and perform simple tests. For example, if we want to test whether $\beta_2 = 0$, we construct the t-statistics as the estimate divided by its standard error to get $t_2 = 1.16610/0.11224 = 10.39$. Given the large number of observations, the appropriate t-distribution is virtually identical to the standard normal one, so that the 5% two-tailed critical value is 1.96. This means that we clearly reject the null hypothesis that $\beta_2 = 0$. That is, we reject that in the US population the expected wage differential between males and females is zero. We can also compute a confidence interval, which has bounds $1.17 \pm 1.96 \times 0.11$. This means that with 95% confidence we can say that over the entire US population the expected wage differential between males and females is between \$0.95 and \$1.39 per hour.

1.5.3 A joint test of significance of regression coefficients

A standard test that is typically automatically suplied by a regression package (also suplied by R) is a test for the joint hypothesis that all coefficients, except the intercept β_1, are equal to zero. Without loss of generality, assume that these are the last J coefficients in the model

$$H_0 : \beta_{K-J+1} = \cdots = \beta_K = 0.$$

The alternative hypothesis in this case is that H_0 is not true, i.e., at least one of these J coefficients is not equal to zero.

We can define the following test statistic:

$$F = \frac{(S_0 - S_1)/J}{S_1/(N-K)}, \tag{26}$$

where S_1 is the residual sum of squares of the full model and S_0 is the residual sum of squares of the restricted model. Under the null hypothesis, F has and F-distribution with J and $N-K$ degrees of freedom. denoted F^J_{N-K}. If we use the definition of the R^2 from (2.42), we can also write this F-statistic as

$$F = \frac{\left(R_1^2 - R_0^2\right)/J}{\left(1 - R_1^2\right)/(N-K)}, \tag{27}$$

where R_1^2 and R_0^2 are the usual goodness-of-fit measures for the unrestricted and the restricted model, respectively. This shows that the test can be interpreted as testing whether the increase in R^2 moving from the restricted model to the more general model is significant.

It is clear that in this case only very large values for the test statistic imply rejection of the null hypothesis. Despite the two-sided alternative hypothesis, the critical values $F^J_{N-K:\alpha}$ for this test are one-sided and defined by the following equality:

$$P\left\{F > F^J_{N-K:\alpha}\right\} = \alpha,$$

where α is the significance level of the test. For example, if $N-K = 60$ and $J = 3$ the critical value at the 5% level is 2.76. The resulting test is referred to a the ***F*-test**.

The F-statistic is routinely provided by the majority of all regression packages. Note that it is a simple function of the R^2 of the model (see, (27)).

1.5.4 Example: Individual wages (continued)

The fact that we concluded above that there was a significant difference between expected wage rates for males and females does not necessarily point to discrimination. It is possible that working males and females differ in terms of their characteristics, for example their years of schooling. To analyze this, we can extend the regression model with additional explanatory variables, for example $school_i$, which denotes the years of schooling, and $expr_i$, which denotes experience in years. The model is now interpreted to describe the conditional expected wage of an individual given his or her gender, years of schooling and experience and can be written a

$$wage_i = \beta_1 + \beta_2 male_i + \beta_3 school_i + \beta_4 exper_i + \varepsilon_i.$$

The coefficient β_2 for $male_i$ now measures the difference in expected wage between a male and a female *with the same schooling and experience*. Similarly, the coefficient β_3 for $school_i$ gives the expected wage difference between two individual with the same experience and gender where one has one additional year of schooling. In general, the coefficients in a multiple regression model can only be interpreted under a **ceteris paribus condition**, which says that that the other variables that are included in the model are constant.

```
> #my.data <- read.table("H:/721364P/Rdata/wages1.dat", header=T)
> m2 <- lm(WAGE~MALE+SCHOOL+EXPER)
> summary(m2)

Call:
lm(formula = WAGE ~ MALE + SCHOOL + EXPER)

Residuals:
   Min     1Q Median     3Q    Max
-7.654 -1.967 -0.457  1.444 34.194

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.38002    0.46498  -7.269 4.50e-13 ***
MALE         1.34437    0.10768  12.485  < 2e-16 ***
SCHOOL       0.63880    0.03280  19.478  < 2e-16 ***
EXPER        0.12483    0.02376   5.253 1.59e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 3290 degrees of freedom
Multiple R-squared: 0.1326,        Adjusted R-squared: 0.1318
F-statistic: 167.6 on 3 and 3290 DF,  p-value: < 2.2e-16

> # males
> males <- subset(my.data,MALE==1)
> m.males <- lm(WAGE~SCHOOL+EXPER,data=males)
> summary(m.males)

Call:
lm(formula = WAGE ~ SCHOOL + EXPER, data = males)

Residuals:
   Min     1Q Median     3Q    Max
-7.850 -2.120 -0.539  1.553 34.203

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -2.62390    0.68924  -3.807 0.000146 ***
SCHOOL       0.69349    0.04760  14.569  < 2e-16 ***
EXPER        0.12032    0.03505   3.433 0.000612 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.303 on 1722 degrees of freedom
Multiple R-squared: 0.1099,        Adjusted R-squared: 0.1089
F-statistic: 106.3 on 2 and 1722 DF,  p-value: < 2.2e-16

> # females
> females <- subset(my.data,MALE==0)
> m.females <- lm(WAGE~SCHOOL+EXPER,data=females)
> summary(m.females)

Call:
lm(formula = WAGE ~ SCHOOL + EXPER, data = females)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9093 -1.7883 -0.4244  1.3091 27.0794

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.59337    0.60235  -4.305 1.77e-05 ***
SCHOOL       0.56123    0.04491  12.496  < 2e-16 ***
EXPER        0.14184    0.03184   4.455 8.98e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.733 on 1566 degrees of freedom
Multiple R-squared: 0.09805,        Adjusted R-squared: 0.0969
F-statistic: 85.12 on 2 and 1566 DF,  p-value: < 2.2e-16

> # test the schooling parameters are equal:
> t.school <- (0.69349-0.56123)/0.04760 # test statistic
> t.school

[1] 2.778571

> pval <- 2*(1-pnorm(t.school)) # p-value
> pval # reject the null

[1] 0.005459851

> # test the experience parameters are equal:
> t.exper <-  (0.12032-0.14184)/0.03505 # test statistic
> t.exper

[1] -0.61398

> pval <- 2*pnorm(t.exper) # p-value
> pval # no not reject the null

[1] 0.5392285
```

The coefficient for $male_i$ now suggests that, if we compare an arbitrary male and female with the same years of schooling and experience, the expected wage differential is \$1.34 compared with \$1.17 before. With a standard error of \$0.11, this difference is still statistically highly significant. The null hypothesis that schooling has no effect on a person's wage, given gender and experience, can be tested using the t-test described above, with a test statistic of 19.48. Clearly the null hypothesis has to be rejected. The estimated wage increase from one additional year of schooling, keeping years of experience fixed, is \$0.64.

It should not be surprising, given these results, that the joint hypothesis that all three partial slope coefficients ($\beta_2, \beta_3, \beta_4$) are zero, that is, wages are not affected by gender, schooling or experience, has to be rejected as well. The F-statistic takes the value of 167.6, while the appropriate 5% critical value is 2.60.

Finally, we can use the above results to compare this model with the simpler one, the first model. The R^2 has increased from 0.0317 to 0.1326, which means that the current model is able to explain 13.3% of the within-sample variation in wages. We can perform a joint test on the hypothesis that the two additional variables, schooling and experience, *both* have zero coefficients, by performing the F-test described above. The test statistic in (27) can be computed from the R^2s reported in the OLS outputs as

$$F = \frac{(0.1326 - 0.0317/2)}{(1-0.1326)\,/\,(3294-4)} = 191.35.$$

With 5% critical value of 3.00, the null hypothesis is obviously rejected. We can thus conclude that the model that included gender, schooling and experience performs significantly better than the model that only includes gender.

1.5.5 More testing examples:

```
> # F-test
> # Estimating the restricted (restricting some (or all) of slope coefficients to be zero)
> # and the unrestricted model (allowing non-zero as well as zero coefficients). You can use
> # anova() (analysis of variance) to test the joint hypotheses defined as in the restricted model.
> mod.restricted <- lm(WAGE~MALE)  # restricted model
> # summary(mod.restricted)   # output suppressed
> mod.unrestricted <- lm(WAGE~MALE+SCHOOL+EXPER)  # unrestricted model
> # summary(mod.unrestricted) # output suppressed
> anova(mod.restricted,mod.unrestricted)

Analysis of Variance Table

Model 1: WAGE ~ MALE
Model 2: WAGE ~ MALE + SCHOOL + EXPER
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   3292 34077
2   3290 30528  2      3549 191.24 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> library(lmtest)
> # t-test for coefficients
> mod <- lm(WAGE~MALE+SCHOOL+EXPER)
> # summary(mod) # output suppressed coeftest(mod)
> # Wald test
> mod1 <- lm(WAGE~MALE+SCHOOL+EXPER,data=my.data)
> #summary(mod1)
> mod2 <- lm(WAGE~SCHOOL+EXPER,data=my.data)
> #summary(mf)
> waldtest(mod1,mod2)
```

```
Wald test

Model 1: WAGE ~ MALE + SCHOOL + EXPER
Model 2: WAGE ~ SCHOOL + EXPER
  Res.Df Df      F    Pr(>F)
1   3290
2   3291 -1 155.88 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.6 Illustration: The Capital Asset Pricing Model (Verbeek: pp. 38–42)

```
> capm.data <- read.table("H:/721364P/Rdata/capm.dat", header=T)
> head(capm.data)
```

```
   month constrrf durblrf foodrf   hml jan   rf  rmrf   smb
1 196001    -6.82    0.96  -4.61  2.69   1 0.33 -6.97  2.04
2 196002     2.63    3.64   2.79 -1.98   0 0.29  1.13  0.56
3 196003    -0.33   -2.25  -1.80 -2.91   0 0.35 -1.63 -0.43
4 196004    -1.99    2.56   0.84 -2.39   0 0.19 -1.72  0.43
5 196005     3.75    6.79   7.38 -3.79   0 0.27  3.13  1.34
6 196006     2.06   -1.29   4.96 -0.33   0 0.24  2.06 -0.16
```

```
> tail(capm.data)
```

```
     month constrrf durblrf foodrf   hml jan   rf  rmrf   smb
605 201005    -8.27   -5.42  -4.86 -2.36   0 0.01 -8.00 -0.03
606 201006   -14.28   -8.82  -1.97 -4.28   0 0.01 -5.21 -2.05
607 201007     5.40    4.07   6.67  0.13   0 0.01  7.24 -0.08
608 201008    -4.59   -3.81  -0.65 -1.71   0 0.01 -4.40 -2.92
609 201009    10.47   11.55   3.17 -3.14   0 0.01  9.24  3.97
610 201010    -1.03    1.58   3.98 -2.14   0 0.01  3.89  0.91
```

```
> attach(capm.data)
> summary(capm.data)
```

```
     month           constrrf            durblrf              foodrf
 Min.   :196001   Min.   :-29.8100   Min.   :-25.9000   Min.   :-18.800
 1st Qu.:197209   1st Qu.: -3.1800   1st Qu.: -2.8500   1st Qu.: -1.603
 Median :198506   Median :  0.4200   Median :  0.4250   Median :  0.715
 Mean   :198498   Mean   :  0.4379   Mean   :  0.3275   Mean   :  0.653
 3rd Qu.:199802   3rd Qu.:  3.7475   3rd Qu.:  4.0000   3rd Qu.:  3.223
 Max.   :201010   Max.   : 25.5200   Max.   : 29.4500   Max.   : 19.520
      hml                jan                 rf              rmrf
 Min.   :-12.7800   Min.   :0.00000   Min.   :0.000   Min.   :-23.140
 1st Qu.: -1.1500   1st Qu.:0.00000   1st Qu.:0.280   1st Qu.: -2.205
 Median :  0.4350   Median :0.00000   Median :0.410   Median :  0.840
 Mean   :  0.4022   Mean   :0.08361   Mean   :0.427   Mean   :  0.437
 3rd Qu.:  1.7975   3rd Qu.:0.00000   3rd Qu.:0.530   3rd Qu.:  3.462
 Max.   : 13.8400   Max.   :1.00000   Max.   :1.350   Max.   : 16.050
      smb
 Min.   :-16.670
 1st Qu.: -1.460
 Median :  0.070
 Mean   :  0.230
 3rd Qu.:  2.038
 Max.   : 22.190
```

```
> # Table 2.3 in Verbeek
> # CAPM regressions without intercept
> m1 <- lm(foodrf~rmrf-1)   # Food
> summary(m1)

Call:
lm(formula = foodrf ~ rmrf - 1)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.539  -1.026   0.141   1.745  15.924 

Coefficients:
     Estimate Std. Error t value Pr(>|t|)    
rmrf  0.75774    0.02579   29.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 2.884 on 609 degrees of freedom
Multiple R-squared: 0.5864,        Adjusted R-squared: 0.5857 
F-statistic: 863.5 on 1 and 609 DF,  p-value: < 2.2e-16 

> m2 <- lm(durblrf~rmrf-1)  # Durables
> summary(m2)

Call:
lm(formula = durblrf ~ rmrf - 1)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.6504 -1.9420 -0.3069  1.7332 17.8871 

Coefficients:
     Estimate Std. Error t value Pr(>|t|)    
rmrf  1.04736    0.02775   37.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 3.105 on 609 degrees of freedom
Multiple R-squared: 0.7005,        Adjusted R-squared:   0.7 
F-statistic:  1424 on 1 and 609 DF,  p-value: < 2.2e-16 

> m3 <- lm(constrrf~rmrf-1) # Construction
> summary(m3)

Call:
lm(formula = constrrf ~ rmrf - 1)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.9414  -1.7193  -0.1866   1.4458  11.6551 

Coefficients:
     Estimate Std. Error t value Pr(>|t|)    
rmrf  1.16662    0.02535   46.01   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.836 on 609 degrees of freedom
Multiple R-squared: 0.7766,	Adjusted R-squared: 0.7763
F-statistic:  2117 on 1 and 609 DF,  p-value: < 2.2e-16

> # Table 2.4 in Verbeek
> # CAPM regressions with intercept
> m4 <- lm(foodrf~rmrf)     # Food
> summary(m4)

Call:
lm(formula = foodrf ~ rmrf)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.8088  -1.3498  -0.1708   1.4423  15.5687 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.32486    0.11669   2.784  0.00554 ** 
rmrf         0.75082    0.02576  29.142  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.869 on 608 degrees of freedom
Multiple R-squared: 0.5828,	Adjusted R-squared: 0.5821
F-statistic: 849.2 on 1 and 608 DF,  p-value: < 2.2e-16

> m5 <- lm(durblrf~rmrf)    # Durables
> summary(m5)

Call:
lm(formula = durblrf ~ rmrf)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5355 -1.8116 -0.1857  1.8485 17.9876 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.13141    0.12628  -1.041    0.298    
rmrf         1.05016    0.02788  37.664   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.104 on 608 degrees of freedom
Multiple R-squared:   0.7,	Adjusted R-squared: 0.6995
F-statistic:  1419 on 1 and 608 DF,  p-value: < 2.2e-16

> m6 <- lm(constrrf~rmrf)   # Construction
> summary(m6)

Call:
lm(formula = constrrf ~ rmrf)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-12.879  -1.641  -0.115   1.520  11.725

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07259    0.11542  -0.629     0.53
rmrf         1.16817    0.02549  45.837   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.837 on 608 degrees of freedom
Multiple R-squared: 0.7756,        Adjusted R-squared: 0.7752
F-statistic:  2101 on 1 and 608 DF,  p-value: < 2.2e-16

> # Table 2.5 in Verbeek
> # CAPM regressions with intercept and January dummy
> m7 <- lm(foodrf~jan+rmrf)     # Food
> summary(m7)

Call:
lm(formula = foodrf ~ jan + rmrf)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8969  -1.3599  -0.1552   1.4408  15.5047

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.39745    0.12140   3.274  0.00112 **
jan         -0.87849    0.41870  -2.098  0.03631 *
rmrf         0.75277    0.02571  29.280  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.861 on 607 degrees of freedom
Multiple R-squared: 0.5858,        Adjusted R-squared: 0.5844
F-statistic: 429.2 on 2 and 607 DF,  p-value: < 2.2e-16

> m8 <- lm(durblrf~jan+rmrf)    # Durables
> summary(m8)

Call:
lm(formula = durblrf ~ jan + rmrf)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5223 -1.8001 -0.1801  1.8415 18.0025

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.14287    0.13184  -1.084    0.279
jan          0.13872    0.45473   0.305    0.760
rmrf         1.04985    0.02792  37.600   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.107 on 607 degrees of freedom
Multiple R-squared:   0.7,        Adjusted R-squared: 0.699
F-statistic: 708.3 on 2 and 607 DF,  p-value: < 2.2e-16
```

```
> m9 <- lm(constrrf~jan+rmrf)   # Construction
> summary(m9)
```

```
Call:
lm(formula = constrrf ~ jan + rmrf)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8203  -1.6622  -0.0673   1.5535  11.7772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.12246    0.12031  -1.018    0.309
jan          0.60354    0.41494   1.455    0.146
rmrf         1.16683    0.02548  45.797   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.835 on 607 degrees of freedom
Multiple R-squared: 0.7763,       Adjusted R-squared: 0.7756
F-statistic:  1054 on 2 and 607 DF,  p-value: < 2.2e-16
```

1.6.1 The World's largest hedge fund (Verbeek, pp. 42–43)

```
> madoff.data <- read.table("H:/721364P/Rdata/madoff.dat", header=T)
> head(madoff.data)
```

```
      month  fsl fslrf   hml   rf  rmrf   smb
1 01dec1990 2.77  2.17 -1.50 0.60  2.35  0.77
2 01jan1991 3.01  2.49 -1.73 0.52  4.39  3.85
3 01feb1991 1.40  0.92 -0.59 0.48  7.10  3.89
4 01mar1991 0.52  0.08 -1.19 0.44  2.45  3.92
5 01apr1991 1.32  0.79  1.43 0.53 -0.20  0.52
6 01may1991 1.82  1.35 -0.56 0.47  3.60 -0.33
```

```
> tail(madoff.data)
```

```
        month   fsl fslrf   hml   rf   rmrf   smb
210 01may2008  0.81  0.64 -0.31 0.17   2.22  2.87
211 01jun2008 -0.06 -0.23 -1.05 0.17  -8.03  1.08
212 01jul2008  0.72  0.57  3.61 0.15  -1.47  3.71
213 01aug2008  0.71  0.59  1.46 0.12   0.99  3.76
214 01sep2008  0.50  0.35  4.48 0.15  -9.96 -0.24
215 01oct2008 -0.06 -0.14 -3.13 0.08 -18.54 -2.12
```

```
> attach(madoff.data)
```

```
The following object(s) are masked from 'capm.data':

    hml, month, rf, rmrf, smb
```

```
> summary(madoff.data)

       month          fsl              fslrf              hml
 01apr1991:  1   Min.   :-0.6400   Min.   :-1.0100   Min.   :-12.7800
 01apr1992:  1   1st Qu.: 0.2950   1st Qu.:-0.0400   1st Qu.: -1.3450
 01apr1993:  1   Median : 0.7300   Median : 0.3900   Median :  0.3100
 01apr1994:  1   Mean   : 0.8422   Mean   : 0.5246   Mean   :  0.4164
 01apr1995:  1   3rd Qu.: 1.2700   3rd Qu.: 0.9400   3rd Qu.:  1.9550
 01apr1996:  1   Max.   : 3.2900   Max.   : 3.1400   Max.   : 13.8400
 (Other)  :209
       rf              rmrf              smb
 Min.   :0.0600   Min.   :-18.5400   Min.   :-16.670
 1st Qu.:0.2200   1st Qu.: -2.0800   1st Qu.: -1.635
 Median :0.3500   Median :  1.0300   Median :  0.050
 Mean   :0.3177   Mean   :  0.4795   Mean   :  0.254
 3rd Qu.:0.4200   3rd Qu.:  3.3600   3rd Qu.:  2.185
 Max.   :0.6000   Max.   : 10.3000   Max.   : 22.190

> # Table 2.6 CAPM regression (with intercept) Madoff's returns
> madoff.capm <-lm(fslrf~rmrf)
> summary(madoff.capm)

Call:
lm(formula = fslrf ~ rmrf)

Residuals:
     Min       1Q   Median       3Q      Max
-1.34773 -0.48005 -0.08337  0.38865  2.97276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.50495    0.04570  11.049  < 2e-16 ***
rmrf         0.04089    0.01072   3.813  0.00018 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6658 on 213 degrees of freedom
Multiple R-squared: 0.06388,        Adjusted R-squared: 0.05949
F-statistic: 14.54 on 1 and 213 DF,  p-value: 0.0001801
```

1.7 Asymptotic Properties of the OLS Estimator

This section lists briefly the asymptotic properties of the OLS estimator

1.7.1 Consistency

Let us start with the linear model under the Gauss-Markov assumptions. In this case we know that the OLS estimator b has the following first two moment:

$$E\{b\} = \beta$$

$$V\{b\} = \sigma^2 \left(\sum_{i=1}^{N} x_i x_i^{'} \right)^{-1} = \sigma^2 \left(X'X \right)^{-1}.$$

What happens when the sample size N grows to infinity? It is clear that $\sum_{i=1}^{N} x_i x_i^{'}$ increases as the number of terms grows, so that the variance of b decreases as the sample size increases. If we assume that

$$\frac{1}{N}\sum_{i=1}^{N} x_i x_i^{'} \text{ converges to a finite nonsingular matrix } \Sigma_{xx} \tag{28}$$

if the sample size N becomes infinitely large. It follows directly that 'the probability limit of b is β', or 'b converges in probability to β', or just

$$\text{plim } b = \beta.$$

When an estimator for β converges to the true value, we say that it is a **consistent estimator**.

1.7.2 Asymptotic normality

If the small sample distribution of an estimator is unknown, the best we can do is try to find some approximation. In most cases, one uses an asymptotic approximation (for N growing to infinity) based on the **asymptotic distribution**. Most estimators in econometrics can be shown to be asymptotically normally distributed.

For the OLS estimator it can be shown that under the Gauss-Markov conditions (10)–(13) combined with (28) we have

$$\sqrt{N}\,(b-\beta) \to N\left(0, \sigma^2 \Sigma_{xx}^{-1}\right),$$

where $\to$ means 'is asymptotically distributed as' and $\sqrt{N}$ is referred to as the **rate of convergence**. Thus, the OLS estimator b is asymptotically normally distributed with variance-covariance matrix $\sigma^2 \Sigma_{xx}^{-1}$.

2 Interpreting and Comparing Regression Models

2.1 Interpreting the Linear Model

As already stressed the linear model

$$y_i = x_i^{'}\beta + \varepsilon_i \tag{29}$$

has little meaning unless we complement it with additional assumption on ε_i. It is common to state that ε_i has expectation zero and that the x_is are taken as given. A formal way of stating this is that it is assumed that the expected value of ε_i given X, or expected value of ε_i given x_i is zero; that is

$$E\{\varepsilon_i|X\} = 0 \text{ or } E\{\varepsilon_i|x_i\} = 0 \tag{30}$$

respectively, where the latter condition is implied by the first. Under $E\{\varepsilon_i|x_i\} = 0$, we can interpret the regression model as describing the conditional expected value of y_i given values for the explanatory variables x_i.

For example, what is the expected wage for an *arbitrary* woman of age 40, with a university education and 14 years of experience? Or, what is the expected unemployment rate given wage rates, inflation and total output in the economy? Or, what is the expected return on a stock, if the expected return on the market is 12%, the risk-free rate is 5% and the asset's beta is 0.8?

The first consequence of (30) is the interpretation of the individual β coefficient. For example, β_k measures the expected change in y_i if x_{ik} changes with one unit but all the other variables in x_i do not change. This is,

$$\frac{\partial E\{y_i|x_i\}}{\partial x_{ik}} = \beta_k.$$

It is important to realize that we had to state explicitly that the other variables in x_i did not change. This is the so-called **ceteris paribus condition**. An important consequence of this condition is that *it is impossible to interpret a single coefficient in a regression model without knowing what the other variables in the equation are.* If interest is focused on the relationship between y_i and x_{ik}, the other variables inx_i act as **control variables**.

2.2 Selecting the Set of Regressors

2.2.1 Misspecifying the set of regressors

If one is (implicitely) assuming that the conditioning set of the model contains more variables than the ones that are included, it is possible that the set of explanatory variables is 'misspecified'. This means that one or more of the omitted variables are relevant, i.e. have nonzero coefficients.

2.2.2 Selecting regressors

Again, it should be stressed that, if we interpret the regression model as describing the conditional expectation of y_i given the *included* variables x_i, there is no issue of misspecified set of regressors, although there might be a problem of functional form. This implies that statistically there is nothing to test here. The set of x_i variables will be chosen on the basis of what we find interesting, and often economic theory or common sense guides us in our choice. Interpreting the model in a broader sense implies that there may be relevant regressors that are excluded or irrelevant ones that are included. To find potentially relevant variables, we can use economic thery again.

It is a good practice to select the set of *potentially* relevant variables on the basis of economic arguments rather that statistical ones. Although it is sometimes suggested otherwise, statistical arguments are never certainty arguments. That is, there is always a small (but not ignorable non-zero) probability of drawing the wrong conclusion. For example, there is always a probability of rejecting the null hypothesis that a coefficient is zero, while the null is is actually true. Such **type I errors**[2] are rather likely to happen if we use a sequence of many tests to select the regressors to include in the model. This process is referred to as **data snooping** or **data mining** and in economics it is not a compliment if someone accuses you of doing it. In general, data snooping refers to the fact that a given set of data is used more than once to choose a model specification and to test hypotheses. You can imagine, for example, that, if you have a set of 20 potential regressors and you try each one of them, it is quite likely to conclude that one of them is significant, even though there is no true relationship between any of these regressors and the variable you are explaining.[3] The probability of making incorrect choises is high, and it is not unlikely that your 'model' captures some peculiarities (e.g. 'calendar anomalies') in the data that have no real meaning outside the sample. In practice, however, it is hard to prevent some amount of data snooping from entering your work.[4]

Besides formal statistical tests there are other criteria that are sometimes used to select a set of regressors. First of all, theR^2, discussed earlier , measures the proportion of the sample variation in y_i that is explained by variation in x_i. However, using R^2 as the criterion would not be optimal, since with too many variables we will not be able to say very much about the model's coefficients , as they may be estimated rather inaccurately. Because the R^2 does not 'punish' the inclusion of many variables, it would be better to use a measure that incorporates a trade-off between goodness-of-fit and the number of regressors. The adjusted $\bar{R}^2$ (24) is such a measure.

There exist a number of alternative criteria that provide such trade-off, the most common ones being **Akaike's Information Criterion** (AIC)

$$AIC = \ln \frac{1}{N} \sum_{i=1}^{N} e_i^2 + \frac{2K}{N} \tag{31}$$

and the **Rissanen-Schwartz Bayesian Information Criterion** (BIC)

$$BIC = \ln \frac{1}{N} \sum_{i=1}^{N} e_i^2 + \frac{K}{N} \ln N. \tag{32}$$

[2] A **type II error** is such that that the null hypothesis is not rejected while the alternative is true.

[3] Although statistical software packages sometimes provide mechanical routines, e.g. stepwise regression, to select regressors, these are typically *not recommended* in economic work.

[4] In recent years, the possibility of data snooping biases has played an important role in empirical studies modelling financial asset pricing models. Lo and MacKinlay (Lo and MacKinlay (1990), Data-Snooping Biases in Tests of Financial Asset Pricing Models, *Review of Financial Studies*, 3, 431–469), for example, analyse such biases in tests of financial asset pricing models, while Sullivan, Timmerman and White (Sullivan, Timmerman and White (2001), Dangers of Data-Driven Inference: The Case of Calendar Effects in Stock Returns, *Journal of Econometrics*, 105, 249–286) analyse the extent to which the presence of calendar effects in stock returns, like the January effect can be attributed to data snooping.

Models with lower AIC or BIC are typically preferred. Note that both criteria add a penaty that increases with the number of regressors. Because the penaly is larger for BIC, the latter criterion tend to favour more parsimomious ('less parameters') than AIC.

Alternatively, it is possible to test whether the increase in R^2 is statistically significant using the F-test (27).

2.3 Illustration: Explaining House Prices (Verbeek, pp. 72–76)

```
> house <- read.table("H:/721364P/Rdata/HOUSING.dat", header=T)
> attach(house)
> names(house)
```

```
 [1] "price"    "lotsize"  "bedrooms" "bathrms"  "stories"  "driveway"
 [7] "recroom"  "fullbase" "gashw"    "airco"    "garagepl" "prefarea"
```

```
> head(house)
```

```
  price lotsize bedrooms bathrms stories driveway recroom fullbase gashw airco
1 42000    5850        3       1       2        1       0        1     0     0
2 38500    4000        2       1       1        1       0        0     0     0
3 49500    3060        3       1       1        1       0        0     0     0
4 60500    6650        3       1       2        1       1        0     0     0
5 61000    6360        2       1       1        1       0        0     0     0
6 66000    4160        3       1       1        1       1        1     0     1
  garagepl prefarea
1        1        0
2        0        0
3        0        0
4        0        0
5        0        0
6        0        0
```

```
> # Table 3.1 page 73
> house1 <- lm(log(price)~log(lotsize)+bedrooms+bathrms+airco)
> summary(house1)
```

```
Call:
lm(formula = log(price) ~ log(lotsize) + bedrooms + bathrms +
    airco)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81782 -0.15562  0.00778  0.16468  0.84143

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.09378    0.23155  30.636  < 2e-16 ***
log(lotsize)  0.40042    0.02781  14.397  < 2e-16 ***
bedrooms      0.07770    0.01549   5.017 7.11e-07 ***
bathrms       0.21583    0.02300   9.386  < 2e-16 ***
airco         0.21167    0.02372   8.923  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2456 on 541 degrees of freedom
Multiple R-squared: 0.5674,	Adjusted R-squared: 0.5642
F-statistic: 177.4 on 4 and 541 DF,  p-value: < 2.2e-16
```

```
> AIC(house1)

[1] 23.05703

> BIC(house1)

[1] 48.87274

> # Table 3.2 page 74
> house2 <- lm(log(price)~log(lotsize)+bedrooms+bathrms+airco+driveway+recroom+
+ fullbase+gashw+garagepl+prefarea+stories)
> summary(house2)

Call:
lm(formula = log(price) ~ log(lotsize) + bedrooms + bathrms +
    airco + driveway + recroom + fullbase + gashw + garagepl +
    prefarea + stories)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68355 -0.12247  0.00802  0.12780  0.67564

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.74509    0.21634  35.801  < 2e-16 ***
log(lotsize)  0.30313    0.02669  11.356  < 2e-16 ***
bedrooms      0.03440    0.01427   2.410 0.016294 *
bathrms       0.16576    0.02033   8.154 2.52e-15 ***
airco         0.16642    0.02134   7.799 3.29e-14 ***
driveway      0.11020    0.02823   3.904 0.000107 ***
recroom       0.05797    0.02605   2.225 0.026482 *
fullbase      0.10449    0.02169   4.817 1.90e-06 ***
gashw         0.17902    0.04389   4.079 5.22e-05 ***
garagepl      0.04795    0.01148   4.178 3.43e-05 ***
prefarea      0.13185    0.02267   5.816 1.04e-08 ***
stories       0.09169    0.01261   7.268 1.30e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2104 on 534 degrees of freedom
Multiple R-squared: 0.6865,     Adjusted R-squared: 0.6801
F-statistic: 106.3 on 11 and 534 DF,  p-value: < 2.2e-16

> AIC(house2)

[1] -138.8234

> BIC(house2)

[1] -82.88931

> #  Table 3.3 page 75
> house3 <- lm(price~lotsize+bedrooms+bathrms+airco+driveway+recroom+
+ fullbase+gashw+garagepl+prefarea+stories)
> summary(house3)
```

```
Call:
lm(formula = price ~ lotsize + bedrooms + bathrms + airco + driveway +
    recroom + fullbase + gashw + garagepl + prefarea + stories)

Residuals:
   Min     1Q Median     3Q    Max
-41389  -9307   -591   7353  74875

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4038.3504  3409.4713  -1.184 0.236762
lotsize         3.5463     0.3503  10.124  < 2e-16 ***
bedrooms     1832.0035  1047.0002   1.750 0.080733 .
bathrms     14335.5585  1489.9209   9.622  < 2e-16 ***
airco       12632.8904  1555.0211   8.124 3.15e-15 ***
driveway     6687.7789  2045.2458   3.270 0.001145 **
recroom      4511.2838  1899.9577   2.374 0.017929 *
fullbase     5452.3855  1588.0239   3.433 0.000642 ***
gashw       12831.4063  3217.5971   3.988 7.60e-05 ***
garagepl     4244.8290   840.5442   5.050 6.07e-07 ***
prefarea     9369.5132  1669.0907   5.614 3.19e-08 ***
stories      6556.9457   925.2899   7.086 4.37e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15420 on 534 degrees of freedom
Multiple R-squared: 0.6731,     Adjusted R-squared: 0.6664
F-statistic: 99.97 on 11 and 534 DF,  p-value: < 2.2e-16
```

> AIC(house3)

```
[1] 12094.19
```

> BIC(house3)

```
[1] 12150.12
```

2.4 Illustration: Predicting Stock Index Returns (Verbeek, pp. 76–78)

```
> stocks <- read.table("H:/721364P/Rdata/PREDICTSP.dat", header=T)
> attach(stocks)
> names(stocks)
```

```
 [1] "OBS"    "CS_1"   "DY_1"   "EXRET"  "I12_1"  "I12_2"  "I3_1"   "I3_2"
 [9] "INF_2"  "IP_2"   "MB_2"   "PE_1"   "TS_1"   "WINTER"
```

> head(stocks)

```
      OBS     CS_1     DY_1      EXRET     I12_1     I12_2   I3_1   I3_2
1 1966M01 0.027130 0.246132  0.3556919 0.3978448 0.3850723 0.3747 0.3579
2 1966M02 0.025523 0.246734 -1.8896210 0.4026298 0.3978448 0.3795 0.3747
3 1966M03 0.027106 0.253051 -2.3094014 0.4050214 0.4026298 0.3747 0.3795
4 1966M04 0.031842 0.260563  1.9798476 0.3994401 0.4050214 0.3771 0.3747
5 1966M05 0.035802 0.257156 -5.5604503 0.4018325 0.3994401 0.3787 0.3771
6 1966M06 0.039764 0.273811 -1.7151910 0.4050214 0.4018325 0.3675 0.3787
     INF_2     IP_2     MB_2   PE_1      TS_1 WINTER
```

```
1 0.026786 0.100527 0.046695 0.1785  0.002380      1
2 0.032738 0.096325 0.051422 0.1743 -0.003219      1
3 0.032641 0.093860 0.051778 0.1712  0.019155      1
4 0.038576 0.095886 0.055034 0.1674  0.019947      1
5 0.038576 0.094733 0.060271 0.1671  0.008768      0
6 0.032448 0.093014 0.058601 0.1581  0.022364      0

> tail(stocks)

        OBS     CS_1     DY_1      EXRET     I12_1     I12_2     I3_1     I3_2
475 2005M07 0.071464 0.146370  3.4772740 0.2983874 0.2757784 0.264500 0.244200
476 2005M08 0.070612 0.141754 -1.2119415 0.3169172 0.2983874 0.282200 0.264500
477 2005M09 0.069013 0.148075  0.5195351 0.3418328 0.3169172 0.280628 0.282200
478 2005M10 0.071359 0.145608 -1.9379552 0.3538643 0.3418328 0.304031 0.280628
479 2005M11 0.075163 0.147520  3.4684277 0.3554674 0.3538643 0.317722 0.304031
480 2005M12 0.076693 0.149449 -0.2840651 0.3634782 0.3554674 0.318527 0.317722
       INF_2     IP_2      MB_2   PE_1     TS_1 WINTER
475 0.035594 0.023892  0.021507 0.1988 0.062874      0
476 0.045730 0.037026  0.013819 0.2050 0.059633      0
477 0.052525 0.031695  0.012422 0.1923 0.067624      0
478 0.070034 0.033067  0.004816 0.1921 0.039407      0
479 0.082044 0.021309  0.005773 0.1880 0.046557      1
480 0.041447 0.024396 -0.001824 0.1872 0.052155      1

> stock.model <- lm(EXRET/100~PE_1+DY_1+INF_2+IP_2+I3_1+I3_2+I12_1+I12_2+MB_2+CS_1+WINTER)
> summary(stock.model)

Call:
lm(formula = EXRET/100 ~ PE_1 + DY_1 + INF_2 + IP_2 + I3_1 +
    I3_2 + I12_1 + I12_2 + MB_2 + CS_1 + WINTER)

Residuals:
      Min        1Q    Median        3Q       Max
-0.204406 -0.024293  0.001638  0.027697  0.146903

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.033010   0.023259   1.419 0.156500
PE_1        -0.113984   0.064504  -1.767 0.077869 .
DY_1         0.059894   0.060119   0.996 0.319642
INF_2       -0.139880   0.068664  -2.037 0.042194 *
IP_2        -0.021111   0.056942  -0.371 0.710994
I3_1         0.191875   0.121159   1.584 0.113945
I3_2        -0.194518   0.119687  -1.625 0.104788
I12_1       -0.413432   0.120863  -3.421 0.000679 ***
I12_2        0.358917   0.125504   2.860 0.004428 **
MB_2        -0.128530   0.061594  -2.087 0.037453 *
CS_1         0.183006   0.099212   1.845 0.065727 .
WINTER       0.008057   0.003915   2.058 0.040149 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04164 on 468 degrees of freedom
Multiple R-squared: 0.1179,       Adjusted R-squared: 0.09712
F-statistic: 5.684 on 11 and 468 DF,  p-value: 1.281e-08
```

```
> AIC(stock.model)

[1] -1675.488

> BIC(stock.model)

[1] -1621.229

> # data snooping:
> # stepwise selection using the AIC
> step(stock.model, direction = "backward")

Start:  AIC=-3039.67
EXRET/100 ~ PE_1 + DY_1 + INF_2 + IP_2 + I3_1 + I3_2 + I12_1 +
    I12_2 + MB_2 + CS_1 + WINTER

         Df Sum of Sq     RSS     AIC
- IP_2    1 0.0002383 0.81175 -3041.5
- DY_1    1 0.0017210 0.81323 -3040.7
<none>                0.81151 -3039.7
- I3_1    1 0.0043489 0.81586 -3039.1
- I3_2    1 0.0045801 0.81609 -3039.0
- PE_1    1 0.0054144 0.81692 -3038.5
- CS_1    1 0.0059000 0.81741 -3038.2
- INF_2   1 0.0071962 0.81871 -3037.4
- WINTER  1 0.0073437 0.81885 -3037.3
- MB_2    1 0.0075506 0.81906 -3037.2
- I12_2   1 0.0141815 0.82569 -3033.3
- I12_1   1 0.0202895 0.83180 -3029.8

Step:  AIC=-3041.53
EXRET/100 ~ PE_1 + DY_1 + INF_2 + I3_1 + I3_2 + I12_1 + I12_2 +
    MB_2 + CS_1 + WINTER

         Df Sum of Sq     RSS     AIC
- DY_1    1 0.0018950 0.81364 -3042.4
<none>                0.81175 -3041.5
- I3_2    1 0.0044567 0.81621 -3040.9
- I3_1    1 0.0044589 0.81621 -3040.9
- PE_1    1 0.0052129 0.81696 -3040.5
- INF_2   1 0.0069583 0.81871 -3039.4
- WINTER  1 0.0071826 0.81893 -3039.3
- MB_2    1 0.0080102 0.81976 -3038.8
- CS_1    1 0.0109916 0.82274 -3037.1
- I12_2   1 0.0139694 0.82572 -3035.3
- I12_1   1 0.0206039 0.83235 -3031.5

Step:  AIC=-3042.41
EXRET/100 ~ PE_1 + INF_2 + I3_1 + I3_2 + I12_1 + I12_2 + MB_2 +
    CS_1 + WINTER

         Df Sum of Sq     RSS     AIC
<none>                0.81364 -3042.4
- I3_1    1 0.0038002 0.81744 -3042.2
- I3_2    1 0.0047271 0.81837 -3041.6
```

```
- INF_2   1 0.0056675 0.81931 -3041.1
- MB_2    1 0.0061733 0.81982 -3040.8
- WINTER  1 0.0075870 0.82123 -3039.9
- CS_1    1 0.0146547 0.82830 -3035.8
- I12_2   1 0.0159771 0.82962 -3035.1
- I12_1   1 0.0199986 0.83364 -3032.8
- PE_1    1 0.0228703 0.83651 -3031.1

Call:
lm(formula = EXRET/100 ~ PE_1 + INF_2 + I3_1 + I3_2 + I12_1 +
    I12_2 + MB_2 + CS_1 + WINTER)

Coefficients:
(Intercept)         PE_1        INF_2         I3_1         I3_2        I12_1
   0.047479    -0.159540    -0.118789     0.177635    -0.196934    -0.409218
      I12_2         MB_2         CS_1       WINTER
   0.375002    -0.095765     0.227046     0.008154

> step(stock.model, direction = "forward")
Start:  AIC=-3039.67
EXRET/100 ~ PE_1 + DY_1 + INF_2 + IP_2 + I3_1 + I3_2 + I12_1 +
    I12_2 + MB_2 + CS_1 + WINTER


Call:
lm(formula = EXRET/100 ~ PE_1 + DY_1 + INF_2 + IP_2 + I3_1 +
    I3_2 + I12_1 + I12_2 + MB_2 + CS_1 + WINTER)

Coefficients:
(Intercept)         PE_1         DY_1        INF_2         IP_2         I3_1
   0.033010    -0.113984     0.059894    -0.139880    -0.021111     0.191875
       I3_2        I12_1        I12_2         MB_2         CS_1       WINTER
  -0.194518    -0.413432     0.358917    -0.128530     0.183006     0.008057

> # or using the stepAIC function in the MASS package
> #library(MASS)
> #stepAIC(stock.model, direction = "backward") # output suppressed
> #stepAIC(stock.model, direction = "forward")  # output suppressed
```

3 OLS Diagnostics

In many cases, the Gauss-Markov conditions (10)–(13) will not all be satisfied. This is not necessarily fatal for the OLS estimator in the sense that it is consistent under fairly weak conditions. In this section we apply three approaches to validating linear regression (OLS) models:

1. A popular approach compares various statistics computed for the full data set with those obtained from deleting single observations. This is known as regression diagnostics.

2. In econometrics, diagnostic tests have played a prominent role since about 1980. The most important alternative hypotheses are **heteroskedasticity**, **autocorrelation**, and misspecification of the functional form.

3. Also, the impenetarble disturbance structures typically present in observational data have been led to the development of "robust" covariance matrix estimators (for the parameter estimates), a number of which have been available during the last 20 years.

In this section we pay attention to heteroskedasticity and autocorrelation, typical to financial data, which imply that the error terms in the model are no longer independently and identically distributed. In such cases, the OLS estimator may still be unbiased or consistent, but its covariance matrix is different from the one given by (14). Moreover, the OLS estimator may be relatively inefficient and no longer have the BLUE property. Let's go through the basic diagnostics using applications.

3.1 Regression Diagnostics

- See, Lecture-3.R and Chapter 4: Diagnostics andAlternative Methods of Regression

```
> library(sandwich)
> data("PublicSchools")
> summary(PublicSchools)

  Expenditure        Income
 Min.   :259.0   Min.   : 5736
 1st Qu.:315.2   1st Qu.: 6670
 Median :354.0   Median : 7597
 Mean   :373.3   Mean   : 7608
 3rd Qu.:426.2   3rd Qu.: 8286
 Max.   :821.0   Max.   :10851
 NA's   :1

> ps <- na.omit(PublicSchools) # remove missing values
> ps$Income <- ps$Income/10000
> #plot(Expenditure ~ Income, data = ps, ylim = c(230,830)) # figure 4.1
> ps_lm <- lm(Expenditure ~ Income, data = ps)
> #abline(ps_lm) #id <- c(2, 24, 48)
> #text(ps[id, 2:1], rownames(ps)[id], pos = 1, xpd = TRUE)
> #plot(ps_lm, which = 1:6) # figure 4.2
> ps_hat <- hatvalues(ps_lm)
> #plot(ps_hat)                # figure 4.3
> #abline(h = c(1, 3) * mean(ps_hat), col = 2)
> #id <- which(ps_hat > 3 * mean(ps_hat))
> #text(id, ps_hat[id], rownames(ps)[id], pos = 1, xpd = TRUE)
> influence.measures(ps_lm)

Influence measures of
         lm(formula = Expenditure ~ Income, data = ps) :

                 dfb.1_  dfb.Incm     dffit cov.r   cook.d    hat inf
Alabama        -1.52e-02  1.39e-02 -1.74e-02 1.103 1.55e-04 0.0543
Alaska         -2.39e+00  2.52e+00  2.65e+00 0.555 2.31e+00 0.2144   *
Arizona        -1.51e-02  9.50e-03 -4.32e-02 1.061 9.51e-04 0.0210
Arkansas        5.93e-05 -5.44e-05  6.73e-05 1.107 2.32e-09 0.0576
California      1.83e-01 -2.09e-01 -2.72e-01 1.031 3.67e-02 0.0485
Colorado       -2.90e-02  4.58e-02  1.30e-01 1.035 8.48e-03 0.0228
Connecticut    -1.83e-01  2.07e-01  2.65e-01 1.042 3.48e-02 0.0515
Delaware        3.45e-02 -4.06e-02 -5.87e-02 1.081 1.76e-03 0.0383
Florida        -2.75e-02  1.17e-02 -1.18e-01 1.035 7.01e-03 0.0202
Georgia        -1.09e-01  9.47e-02 -1.44e-01 1.056 1.05e-02 0.0353
Hawaii          3.31e-02 -4.10e-02 -6.87e-02 1.070 2.41e-03 0.0310
Idaho          -3.03e-02  2.60e-02 -4.27e-02 1.075 9.32e-04 0.0317
Illinois        3.30e-02 -3.81e-02 -5.16e-02 1.088 1.36e-03 0.0439
Indiana        -4.17e-03 -6.73e-03 -8.03e-02 1.050 3.27e-03 0.0201
```

```
Iowa           -1.08e-02  2.35e-02  9.53e-02 1.047 4.60e-03 0.0213
Kansas          2.53e-02 -4.01e-02 -1.14e-01 1.043 6.51e-03 0.0228
Kentucky       -1.15e-01  1.02e-01 -1.48e-01 1.060 1.10e-02 0.0383
Louisiana       2.38e-02 -2.10e-02  3.08e-02 1.082 4.83e-04 0.0373
Maine           1.36e-01 -1.23e-01  1.59e-01 1.076 1.28e-02 0.0501
Maryland       -7.03e-03  8.91e-03  1.60e-02 1.074 1.31e-04 0.0290
Massachusetts  -1.56e-02  2.29e-02  5.72e-02 1.062 1.66e-03 0.0238
Michigan       -5.49e-02  6.69e-02  1.07e-01 1.063 5.80e-03 0.0328
Minnesota      -1.90e-02  4.76e-02  2.13e-01 0.976 2.22e-02 0.0211
Mississippi     7.09e-02 -6.65e-02  7.61e-02 1.137 2.95e-03 0.0848   *
Missouri       -7.47e-02  4.92e-02 -1.98e-01 0.988 1.93e-02 0.0213
Montana         1.57e-01 -1.27e-01  2.68e-01 0.957 3.47e-02 0.0257
Nebraska       -5.19e-02  3.17e-02 -1.55e-01 1.016 1.19e-02 0.0209
Nevada          3.45e-01 -3.86e-01 -4.79e-01 0.949 1.08e-01 0.0575
New Hampshire  -7.62e-02  5.38e-02 -1.77e-01 1.006 1.56e-02 0.0220
New Jersey      8.20e-02 -9.38e-02 -1.24e-01 1.080 7.77e-03 0.0470
New Mexico      2.63e-01 -2.35e-01  3.23e-01 0.988 5.07e-02 0.0425
New York       -3.28e-02  4.22e-02  7.88e-02 1.063 3.16e-03 0.0280
North Carolina  7.97e-02 -7.05e-02  1.02e-01 1.073 5.24e-03 0.0385
North Dakota   -3.24e-02  1.57e-02 -1.26e-01 1.031 7.96e-03 0.0203
Ohio            8.97e-03 -3.01e-02 -1.57e-01 1.015 1.22e-02 0.0208
Oklahoma       -1.42e-02  1.18e-02 -2.20e-02 1.072 2.47e-04 0.0280
Oregon         -1.54e-03  4.05e-03  1.87e-02 1.065 1.79e-04 0.0210
Pennsylvania    1.19e-03  8.42e-03  7.08e-02 1.054 2.55e-03 0.0203
Rhode Island   -1.28e-02  4.73e-03 -5.98e-02 1.057 1.82e-03 0.0201
South Carolina  1.25e-01 -1.14e-01  1.44e-01 1.087 1.05e-02 0.0545
South Dakota    1.33e-03 -1.14e-03  1.91e-03 1.076 1.87e-06 0.0309
Tennessee      -8.06e-02  7.21e-02 -9.85e-02 1.080 4.93e-03 0.0432
Texas          -7.75e-03 -1.29e-02 -1.52e-01 1.015 1.15e-02 0.0201
Utah            2.95e-01 -2.61e-01  3.79e-01 0.934 6.80e-02 0.0380
Vermont         1.47e-01 -1.31e-01  1.83e-01 1.052 1.68e-02 0.0411
Virginia       -5.15e-03 -6.33e-04 -4.27e-02 1.060 9.27e-04 0.0200
Washington      2.56e-02 -3.10e-02 -4.94e-02 1.075 1.24e-03 0.0331
Washington DC   6.57e-01 -7.05e-01 -7.68e-01 1.014 2.77e-01 0.1277   *
West Virginia   7.72e-02 -6.94e-02  9.34e-02 1.083 4.44e-03 0.0446
Wyoming        -7.54e-02  8.41e-02  1.03e-01 1.103 5.36e-03 0.0609

> which(ps_hat > 3 * mean(ps_hat))

       Alaska Washington DC
            2            48

> summary(influence.measures(ps_lm))

Potentially influential observations of
         lm(formula = Expenditure ~ Income, data = ps) :

              dfb.1_  dfb.Incm dffit   cov.r   cook.d  hat
Alaska        -2.39_*  2.52_*   2.65_*  0.55_*  2.31_*  0.21_*
Mississippi    0.07   -0.07     0.08    1.14_*  0.00    0.08
Washington DC  0.66   -0.71    -0.77_*  1.01    0.28    0.13_*
```

3.2 Diagnostic Tests

- See, Lecture-3.R and Chapter 4: Diagnostics andAlternative Methods of Regression

```
> options(prompt = "R> ", continue = "+  ", width = 64, digits = 4,
+ show.signif.stars = FALSE, useFancyQuotes = FALSE)
R> library(AER)
R> # demo("Ch-Validation", package = "AER") # you can run all the demos
R> data("Journals")
R> summary(Journals)

    title                              publisher  society
 Length:180         Elsevier             :42   no :164
 Class :character   Blackwell            :26   yes: 16
 Mode  :character   Kluwer               :16
                    Springer             :10
                    Academic Press       : 9
                    Univ of Chicago Press: 7
                    (Other)              :70
     price          pages          charpp       citations
 Min.   :  20   Min.   : 167   Min.   :1782   Min.   :  21
 1st Qu.: 134   1st Qu.: 549   1st Qu.:2715   1st Qu.:  98
 Median : 282   Median : 693   Median :3010   Median : 262
 Mean   : 418   Mean   : 828   Mean   :3233   Mean   : 647
 3rd Qu.: 541   3rd Qu.: 974   3rd Qu.:3477   3rd Qu.: 656
 Max.   :2120   Max.   :2632   Max.   :6859   Max.   :8999

  foundingyear       subs                        field
 Min.   :1844   Min.   :   2   General             :40
 1st Qu.:1963   1st Qu.:  52   Specialized         :14
 Median :1973   Median : 122   Public Finance      :12
 Mean   :1967   Mean   : 197   Development         :11
 3rd Qu.:1982   3rd Qu.: 268   Finance             :11
 Max.   :1996   Max.   :1098   Urban and Regional: 8
                               (Other)             :84

R> journals <- Journals[,c("subs", "price")]
R> journals$citeprice <- Journals$price/Journals$citations
R> journals$age <- 2000-Journals$foundingyear
R> jour_lm <- lm(log(subs) ~ log(citeprice), data = journals)
R> #summary(jour_lm)
R> # Testing for heteroskedasticity
R> bptest(jour_lm) # Breusch-Pagan test

        studentized Breusch-Pagan test

data:  jour_lm
BP = 9.803, df = 1, p-value = 0.001742

R> gqtest(jour_lm, order.by = ~citeprice, data = journals) # Goldfeld-Quandt test

        Goldfeld-Quandt test

data:  jour_lm
GQ = 1.703, df1 = 88, df2 = 88, p-value = 0.00665

R> # Testing the functional form
R> resettest(jour_lm) # RESET test
```

```
        RESET test

data:  jour_lm
RESET = 1.441, df1 = 2, df2 = 176, p-value = 0.2395

R> raintest(jour_lm, order.by = ~ age, data = journals) # Rainbow test

        Rainbow test

data:  jour_lm
Rain = 1.774, df1 = 90, df2 = 88, p-value = 0.003741

R> harvtest(jour_lm, order.by = ~ age, data = journals) # Harvey-Collier test

        Harvey-Collier test

data:  jour_lm
HC = 5.081, df = 177, p-value = 9.464e-07

R> # Testing for autocorrelation
R> data("USMacroG")
R> summary(USMacroG)

      gdp        consumption       invest       government
 Min.   :1610   Min.   :1059   Min.   : 198   Min.   : 360
 1st Qu.:2602   1st Qu.:1640   1st Qu.: 309   1st Qu.: 741
 Median :4142   Median :2715   Median : 568   Median : 952
 Mean   :4563   Mean   :2999   Mean   : 652   Mean   : 997
 3rd Qu.:6294   3rd Qu.:4235   3rd Qu.: 874   3rd Qu.:1301
 Max.   :9304   Max.   :6341   Max.   :1802   Max.   :1583

      dpi            cpi             m1           tbill
 Min.   :1178   Min.   : 70.6   Min.   : 110   Min.   : 0.81
 1st Qu.:1822   1st Qu.: 91.2   1st Qu.: 148   1st Qu.: 3.09
 Median :3133   Median :162.1   Median : 284   Median : 5.04
 Mean   :3341   Mean   :225.8   Mean   : 454   Mean   : 5.23
 3rd Qu.:4733   3rd Qu.:350.1   3rd Qu.: 764   3rd Qu.: 6.64
 Max.   :6635   Max.   :521.1   Max.   :1152   Max.   :15.09

     unemp         population    inflation
 Min.   : 2.60   Min.   :149   Min.   :-2.53
 1st Qu.: 4.40   1st Qu.:186   1st Qu.: 1.76
 Median : 5.60   Median :215   Median : 3.14
 Mean   : 5.67   Mean   :214   Mean   : 3.94
 3rd Qu.: 6.80   3rd Qu.:243   3rd Qu.: 5.59
 Max.   :10.70   Max.   :281   Max.   :16.86
                               NA's   :1
    interest
 Min.   :-11.216
 1st Qu.: -0.158
 Median :  1.513
 Mean   :  1.311
 3rd Qu.:  2.916
 Max.   : 10.626
 NA's   :1
```

```
R> library(dynlm)
R> consump1 <- dynlm(consumption ~ dpi + L(dpi), data = USMacroG)
R> #summary(consump1)
R> # Alternative way, apply the Lag operator of the Hmisc package
R> #library(Hmisc)
R> #consump0 <- lm(consumption ~ dpi + Lag(dpi), data = USMacroG)
R> #summary(consump0)
R> dwtest(consump1) # Durbin-Watson test

        Durbin-Watson test

data:  consump1
DW = 0.0866, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0

R> Box.test(residuals(consump1), type = "Ljung-Box") # Ljung-Box

        Box-Ljung test

data:  residuals(consump1)
X-squared = 176.1, df = 1, p-value < 2.2e-16

R> bgtest(consump1) # Breusch-Gofrey test

        Breusch-Godfrey test for serial correlation of order up
        to 1

data:  consump1
LM test = 193, df = 1, p-value < 2.2e-16
```

3.3 Robust Standard Errors

- See, Lecture-3.R and Chapter 4: Diagnostics andAlternative Methods of Regression

```
R> vcov(jour_lm)   # covariance matrix of parameter estimates

               (Intercept) log(citeprice)
(Intercept)      3.126e-03     -6.144e-05
log(citeprice)  -6.144e-05      1.268e-03

R> vcovHC(jour_lm) # heteroskedasticity consistent covariance matrix

               (Intercept) log(citeprice)
(Intercept)       0.003085       0.000693
log(citeprice)    0.000693       0.001188

R> vcovHAC(jour_lm)# heteroskedasticity and autocorrelationconsistent covariance matrix (Newey-West)

               (Intercept) log(citeprice)
(Intercept)      0.0026709      0.0003565
log(citeprice)   0.0003565      0.0009710

R> summary(jour_lm)

Call:
lm(formula = log(subs) ~ log(citeprice), data = journals)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.7248 -0.5361  0.0372  0.4662  1.8481

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.7662     0.0559    85.2   <2e-16
log(citeprice)  -0.5331     0.0356   -15.0   <2e-16

Residual standard error: 0.75 on 178 degrees of freedom
Multiple R-squared: 0.557,        Adjusted R-squared: 0.555
F-statistic:  224 on 1 and 178 DF,  p-value: <2e-16

R> #library(lmtest)
R> coeftest(jour_lm, vcov = vcovHC)

t test of coefficients:

               Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.7662     0.0555    85.8   <2e-16
log(citeprice)  -0.5331     0.0345   -15.5   <2e-16

R> coeftest(jour_lm, vcov = vcovHAC) # Newey-West

t test of coefficients:

               Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.7662     0.0517    92.2   <2e-16
log(citeprice)  -0.5331     0.0312   -17.1   <2e-16

R> t(sapply(c("const", "HC0", "HC1", "HC2", "HC3", "HC4"),
+  function(x) sqrt(diag(vcovHC(jour_lm, type = x)))))

      (Intercept) log(citeprice)
const     0.05591        0.03561
HC0       0.05495        0.03377
HC1       0.05526        0.03396
HC2       0.05525        0.03412
HC3       0.05555        0.03447
HC4       0.05536        0.03459
```

3.4 Testing for Normality

Let's consider the linear regression model again with, under the null hypothesis, normal errors. For a continuously observed variable, normality tests usually check for skewness (third moment) and excess kurtosis (fourth moment), because the normal distribution implies that $E\left\{\varepsilon_t^3\right\} = 0$ and $E\left\{\varepsilon_t^4 - 3\sigma^4\right\} = 0$, i.e. for a normal distribution, skewness is zero and excess kurtosis is zero. A popular test for normality is the Jargue-Bera test (see, my Probability and Statistics Review, page 13).

```
R> names(consump1)

 [1] "coefficients"  "residuals"     "effects"
 [4] "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"
[10] "call"          "terms"         "model"
[13] "index"         "frequency"     "twostage"
```

```
R> res <- consump1$residuals
R> library(fBasics)
R> normalTest(res, method = "jb")

Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 72.7384
  P VALUE:
    Asymptotic p Value: < 2.2e-16

Description:
 Wed Feb 06 10:01:35 2013 by user: Hannu

R> # or
R> normalTest(residuals(consump1), method = "jb")

Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 72.7384
  P VALUE:
    Asymptotic p Value: < 2.2e-16

Description:
 Wed Feb 06 10:01:35 2013 by user: Hannu
```