

# Quantitative Analysis of Finance I

## ECON90033

***CORRELATION AND REGRESION ANALYSES***

***LINEAR REGRESSION:  
SPECIFICATION, ESTIMATION AND ASSESSMENT***

# CORRELATION AND REGRESION ANALYSES

- The chi-square test of independence can be used to find out whether two variables are related to each other statistically.

When they are, the relationship between them can be



Deterministic:  $Y = f(X)$

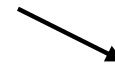
where

$Y$  is the dependent or outcome variable,

$X$  is the independent or predictor variable, and

$f$  denotes a function of  $X$ .

For example, if  $Y$  is the sales value of a product,  $X$  is the number of units sold and  $p$  is the fixed unit price, then  $Y = pX$ .



Probabilistic:  $Y = f(X) + \varepsilon$

where

$\varepsilon$  is a random variable that accounts for the difference between  $Y$  and  $f(X)$ .

For example, if  $Y$  is test mark,  $X$  is the time spent on studies, then  $Y = f(X) + \varepsilon$  (where  $\varepsilon$  is a random variable) since time is not the only factor determining test results.

- Statistics is concerned with probabilistic relationships and focuses on two questions:
  - How strong is the relationship?
  - What is the probable form of the relationship?

Accordingly, there are two types of analysis:

Correlation analysis:

What are the direction and the strength of the relationship?

Regression analysis:

How to specify and estimate the relationship?



- The direction and the strength of a linear relationship, i.e., correlation, between two quantitative variables, say  $X$  and  $Y$ , defined over the same population is provided by the population covariance ( $\sigma_{xy}$ ). Its estimator is the sample covariance ( $s_{xy}$ ).


$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)]$$

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The covariance between  $X$  and  $Y$  has the following properties:

- i. It is zero when there is no linear relationship between the two variables.
- ii. It is positive (negative) when the two variables tend to deviate in the same (opposite) direction from their respective means.
- iii. The larger the absolute value of the covariance, the stronger the linear association between the two variables.
- iv. When  $X = Y$ , the covariance is the variance of the single variable. However, while the variance is always non-negative, the covariance between  $X \neq Y$  can be positive, negative or zero.

As a measure of association, the covariance has two shortcomings:

- 
- a) It does not have natural lower and upper limits.
  - b) It depends on the scales (or units) of measurement used.

It cannot be claimed that the covariance between two variables is particularly small or large; not even that it is relatively small or large in comparison to the covariance between two other variables, unless both pairs of variables are measured on the same scales.

These problems can be overcome by standardizing the covariance with the standard deviations, i.e., by using standardized variables.



- Pearson population correlation coefficient:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = E \left[ \underbrace{\left( \frac{X - \mu_x}{\sigma_x} \right) \left( \frac{Y - \mu_y}{\sigma_y} \right)} \right]$$

Product of the standardized variables or Z variables.

It can be estimated with the Pearson sample correlation coefficient:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

The Pearson sample correlation coefficient has the following properties:

- i. The value of  $r_{xy}$  does not depend on the units of measurement.
- ii.  $-1 \leq r_{xy} \leq 1$
- iii.  $r_{xy} = \pm 1$  indicates that there is a perfect positive/negative linear relationship between the variables (all observations are on a single straight line sloping upward/downward).
- iv.  $r_{xy} = 0$  suggests that there is no linear relationship between the two variables.
- v. The sign of  $r_{xy}$  shows the nature (negative/positive) of the linear relationship and the closer  $r_{xy}$  is to  $\pm 1$ , the stronger the linear relationship between the variables.
- vi. The sampling distribution of  $r_{xy}$  is not standard, but the null hypothesis of no correlation between two normally distributed variables can be tested with a  $t$ -test based on the following statistic:

$$t_r = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \sim t_{df} \quad ; \quad df = n-2$$

## Note:

- a) The Pearson population correlation coefficient also has properties (i)-(v).
- b) The Pearson correlation coefficient requires variables that are quantitative and are measured and reported on interval or ratio scales. The  $t$ -test for  $\rho$ , however, is based on the stronger assumption that both sampled populations are normally distributed.
- c) When the  $t$ -test rejects  $H_0$ , we can conclude that there is a significant linear relationship between the two variables, but this relationship might be weak.

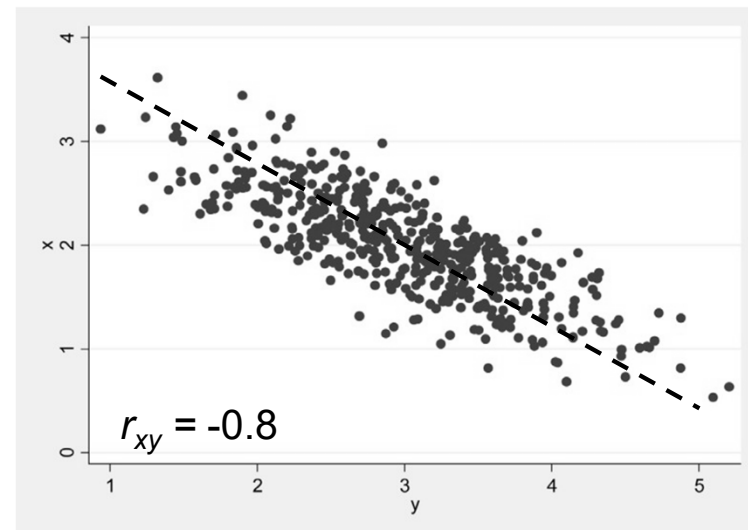
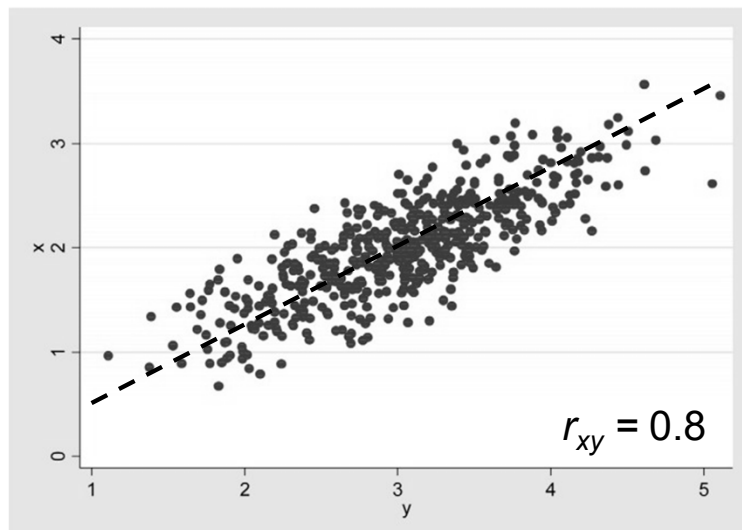
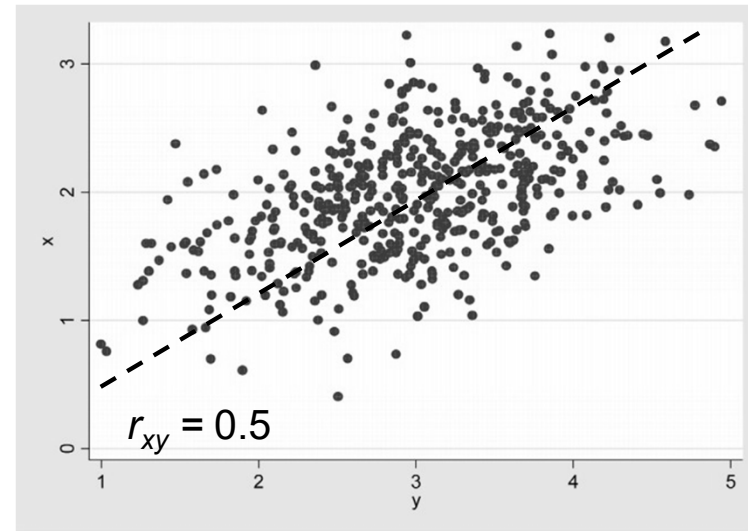
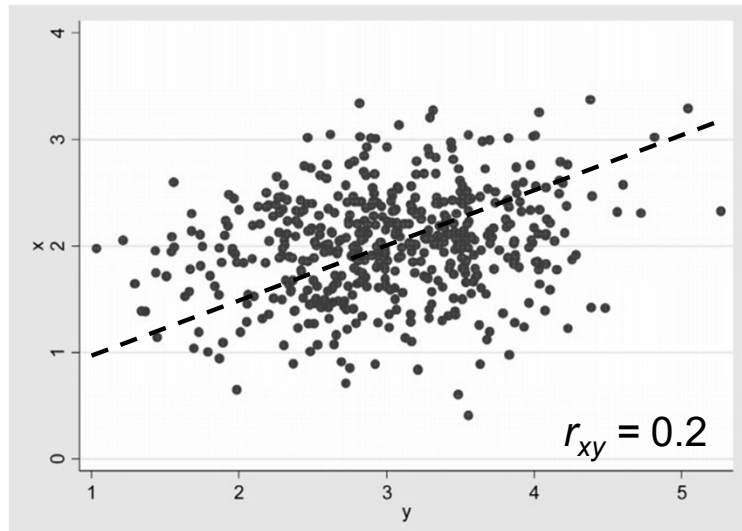
There is no strict rule for the interpretation of the numerical value of  $r_{xy}$ . However, as a rule of thumb,  $|r_{xy}| \geq 0.8$  indicates a strong linear relationship,  $0.5 \leq |r_{xy}| < 0.8$  a moderately strong linear relationship and  $|r_{xy}| < 0.5$  a weak linear relationship (see next slide).

- d) Correlation between two variables, no matter how strong it is, does not necessarily imply a causal relationship.

← Two correlated variables might be both related to a third variable.

Moreover, correlation is a symmetrical relationship, so even if two variables have a cause-effect relationship,  $r_{xy}$  does not indicate which variable causes the other to change.

The scatterplots below illustrate the correlation in samples of independent pairs of observations drawn from bivariate normal populations.



The absolute value of  $r_{xy}$  is not related to the steepness of the straight line, rather to how closely the dots are spread around it.



Note: If two variables are not related to each other at all, then they certainly do not have a linear relationship and thus are not correlated.

However, if two variables are uncorrelated, there is no linear relationship between them, but they might still be associated in some non-linear way.

← Suppose, for example, that we have the following three pairs of observations on  $X$  and  $Y = X^2$ :

X	-1	0	1
Y	1	0	1

There is a quadratic relationship between  $X$  and  $Y$ , so they are not independent of each other.

Yet, they are uncorrelated, since  $\bar{x} = 0$  ,  $\bar{y} = 2 / 3$

and

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$
$$= \frac{(-1)(1/3) + (0)(-2/3) + (1)(1/3)}{2} = 0 \longrightarrow r_{xy} = \frac{s_{xy}}{s_x s_y} = 0$$

Ex 1: (Field, pp. 117-118 and 175-177)

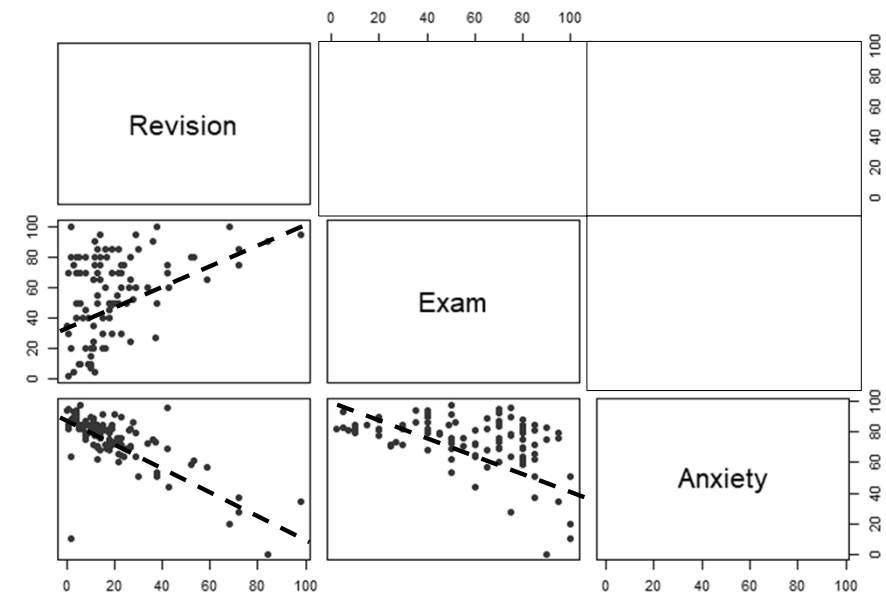
A psychologist was interested in the effects of exam stress on exam performance. She devised and validated a questionnaire to assess anxiety related to exams which produced a measure of anxiety scored out of 100.

She measured the *anxiety* of 103 randomly selected students before an exam and used the percentage marks of the students on the exam to assess their *exam* performances. She also measured the number of hours spent by each student on *revision*. These observations are in *C&R\_e1.RData*.

a) Illustrate the bivariate relationships between the three variables using a scatterplot matrix.

The six scatterplots in the off-diagonal panels illustrate each variable plotted against the other two variables, so this matrix is symmetric, and it is enough to consider only the lower or the upper triangle.

```
pairs(C_R_e1[,1:3], pch = 19, col = "red")
```



None of the best fitting straight lines (the ‘purple’ lines) appears to be horizontal but ascends or descends, implying that *revision* and *exam* tend to increase or decrease together (positive correlation), while *revision* and *anxiety*, and *exam* and *anxiety* move into opposite directions (negative correlations).

These co-movements surely have the logical directions but judging by the dispersion of points around the best fitting straight lines, only the relationship between *revision* and *anxiety* seems to be really strong.

- b) Obtain the Pearson correlation coefficients and check their significance by performing one-tail tests. Assume that the variables are normally distributed.

In general, to make the most of these tests, we have to decide whether the logical relationships between the variables are positive or negative (irrespective of the scatterplots) and set up the hypotheses accordingly.

This time we already concluded in part (a) that the relationships illustrated by the scatterplots make sense, so the alternative hypotheses are

$$H_A : \rho_{\text{Revision, Anxiety}} < 0 \quad , \quad H_A : \rho_{\text{Exam, Revision}} > 0 \quad , \quad H_A : \rho_{\text{Exam, Anxiety}} < 0$$

The correlation coefficients can be obtained and tested with the *cor.test* function of *R*.

*cor.test(Revision, Anxiety,  
alternative = "less")*

Pearson's product-moment correlation

```
data: Revision and Anxiety
t = -10.111, df = 101, [p-value < 2.2e-16]
alternative hypothesis: true correlation is less than 0
95 percent confidence interval:
 -1.0000000 -0.6176435
sample estimates:
cor
-0.7092493
```

*cor.test(Exam, Revision,  
alternative = "greater")*

Pearson's product-moment correlation

```
data: Exam and Revision
t = 4.3434, df = 101, [p-value = 1.672e-05]
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.2498621 1.0000000
sample estimates:
cor
0.3967207
```

*cor.test(Exam, Anxiety,  
alternative = "less")*

Pearson's product-moment correlation

```
data: Exam and Anxiety
t = -4.938, df = 101, [p-value = 1.564e-06]
alternative hypothesis: true correlation is less than 0
95 percent confidence interval:
 -1.0000000 -0.2995071
sample estimates:
cor
-0.4409934
```

Each sample correlation coefficient has the logical/expected sign.

Moreover, the reported  $p$ -values are practically zero, so each  $H_0$  can be rejected at any reasonable significance level.

# LINEAR REGRESSION

- You have certainly learnt about regression analysis, so I assume that you are familiar with this topic. Still, to brush up on this topic, go through these slides.
- Regression analysis is concerned with the specification and estimation of the relationship between a dependent variable and a set of independent variables.

Regression models are primarily classified according to two criteria.

- a) Based on the number of independent variables,  $k$ , regression is referred to as simple or multiple regression.

Simple regression:  $k = 1$ , i.e., there is only one independent variable (e.g., GDP is regressed on employment only).



Multiple regression:  $k \geq 2$ , i.e., there are at least two independent variables (e.g., GDP is regressed on employment and capital stock).

- b) Based on the functional form,  $f$ , regression can be linear or nonlinear in the parameters.

Linear regression:



Linear in the parameters, i.e. the dependent variable depends in a linear way on the parameters.

Non-linear regression:

Non-linear in the parameters, i.e. the dependent variable depends in some non-linear way on the parameters.

Simple linear regression: is linear in the parameters and has only one independent variable ( $k = 1$ )

Multiple linear regression: a generalization of simple linear regression, i.e., it is linear in the parameters and has more than one independent variables ( $k > 1$ )

If the model is also linear in the independent variables ( $X_1, X_2, \dots, X_k$ ), then the population multiple linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

where

$Y$ : dependent variable (quantitative, explained variable / regressand);

$X_j$ :  $j^{\text{th}}$  independent variable (quantitative, explanatory variable / regressor);

$\varepsilon$ : unobservable random error or disturbance term, which represents the net effect of all the variables other than  $X_1, X_2, \dots, X_k$  that influence  $Y$ .

It is assumed that individually these left-out variables are of minor importance, some have positive while others have negative effects on  $Y$  and on average their combined net effect is zero.

→ The conditional expected value of  $\varepsilon$  is supposed to be zero.  
In symbols:  $E(\varepsilon | X_1, X_2, \dots, X_k) = 0$ .

→ Population regression function:

$$E(Y | X_1, X_2, \dots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

the conditional expected value of  $Y$ , i.e., the mean value of the  $Y$  sub-population associated with some given set of values of the independent variables.

$\beta_0$ :  $y$ -intercept parameter – the expected value of  $Y$  when all  $X_j = 0$ .

$\beta_j$  ( $j = 1, \dots, k$ ): slope parameter of  $X_j$  – it measures the impact of a one-unit increase in  $X_j$  on the conditional expected value of  $Y$ , granted that all other independent variables in the model are kept *constant*.

← By keeping the other independent variables constant, we aim to separate the individual effect of  $X_j$  on  $E(Y | X_1, X_2, \dots, X_k)$  from the combined effect of all other included independent variables.

For example,  $\beta_1$  measures the change in  $E(Y | X_1, X_2, \dots, X_k)$  associated with a one-unit increase in  $X_1$ , holding  $X_2, \dots, X_k$  constant.

Similarly,  $\beta_2$  measures the change in  $E(Y | X_1, X_2, \dots, X_k)$  associated with a one-unit increase in  $X_2$ , holding  $X_1, X_3, \dots, X_k$  constant (etc.).

There are three possibilities:

- i. If  $\beta_j = 0$ ,  $E(Y)$  and  $X_j$  are not related to each other.
- ii. If  $\beta_j > 0$ , there is a positive linear relationship between  $E(Y)$  and  $X_j$ .
- iii. If  $\beta_j < 0$ , there is a negative linear relationship between  $E(Y)$  and  $X_j$ .



In practice, the  $\beta_0, \beta_1, \dots, \beta_k$  population parameters are unknown, the  $\varepsilon$  error term cannot be observed, and we know at most a few elements (maybe just one, often none) of each sub-population of  $Y$ .

$\beta_0, \beta_1, \dots, \beta_k$  must be estimated from a sample of corresponding observations of the independent and dependent variables:

$$(y_1, x_{11}, x_{21}, \dots, x_{k1}), (y_2, x_{12}, x_{22}, \dots, x_{k2}), \dots, (y_n, x_{1n}, x_{2n}, \dots, x_{kn})$$

Suppose we managed to do so, in one way or another, and obtained point estimates of the unknown population parameters:

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$$

→ Sample regression function:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k$$

By evaluating this function using the  $i^{\text{th}}$  set of values of the independent variables, we obtain the corresponding point estimate of  $Y$ ,  $\hat{y}_i$ , and the difference between the observed value of  $Y$  (i.e.,  $y_i$ ) and its estimate (i.e.,  $\hat{y}_i$ ) is called the residual,  $e_i$ .

$$\longrightarrow e_i = y_i - \hat{y}_i$$

The residual is the point estimate of the corresponding random error,

$$\varepsilon_i = y_i - E(y_i) \quad \text{where} \quad E(y_i) = E(Y \mid x_{1i}, x_{2i}, \dots, x_{ki})$$

and it measures the error that we commit replacing the true  $Y$  value ( $y_i$ ) with its point estimate ( $\hat{y}_i$ ).

Considering the set of  $n$  observed values of  $Y$ , we would like to keep this type of errors as small as possible.

- There are several estimation methods based on this principle that can be used to estimate the unknown parameters of a linear regression model. The most popular is the ordinary least squares (OLS) method, which aims to minimize the Sum of Squared Errors (SSE):

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \longrightarrow \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$$

For  $k = 1$  the formulas of the OLS estimators are relatively simple:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where  $s_{xy}$  is the sample covariance between  $X$  and  $Y$ , and  $s_x^2$  is the sample variance of  $X$ .

- The popularity of the OLS method is due to three factors.
  - i. Its objective function is easy to understand and is intuitively appealing.
  - ii. The OLS estimators have some attractive properties (if the model includes  $\beta_0$ ):
    - The sum of the residuals is zero;
    - The sample regression equation is satisfied by the sample means of the independent and dependent variables;
    - The estimated and observed  $Y$  values have the same mean.
    - The estimated  $Y$  values and the corresponding residuals are uncorrelated with each other and they add up to the observed  $Y$  values.
  - iii. Under the classical assumptions the OLS estimators are the ‘best’ within a wide range of estimators (see later.)

Ex 2: (Selvanathan, p. 742, ex. 17.20 and p. 813, ex. 18.8)

After several semesters without success, Pat Statsdud, a student enrolled in a statistics subject, decided to try to improve. He needed to know the secret of success for university students. After many hours of discussion with other, more successful students, Pat postulated a rather radical theory: the longer one studied, the better one's grade. Nevertheless, Pat did not want to spend too much time on studying as his ambition was to ultimately graduate with as little work as possible.

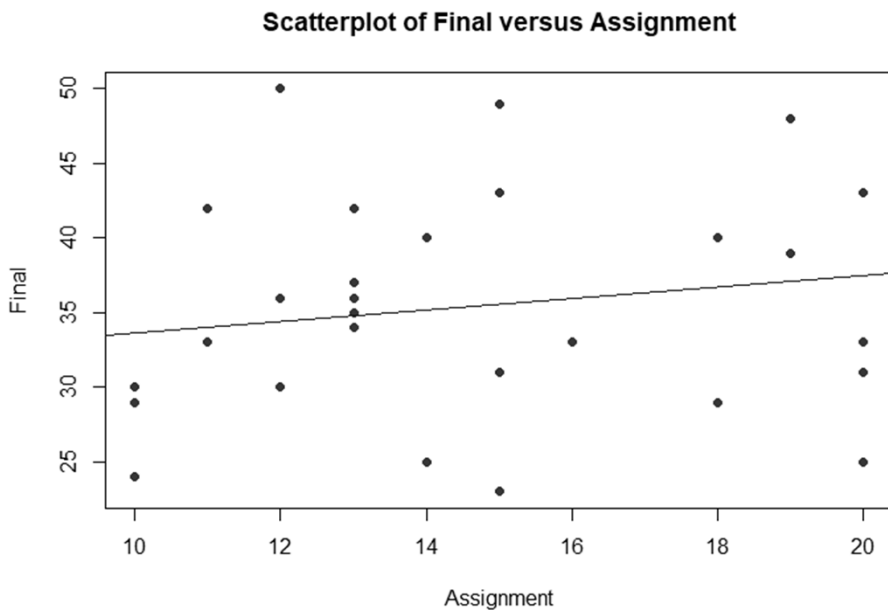
The final grade in this subject was determined in the following way: assignment 20%, mid-term test 30%, final exam 50%. Pat wished to predict the final exam mark on the basis of the assignment mark and the mid-term test mark. His marks on these were 12/20 and 14/30, respectively. Accordingly, Pat decided to undertake a multiple linear regression analysis using the final exam mark, assignment mark and mid-term test mark for 30 students who took the statistics subject last year.

→ The assignment and mid-term test marks are the independent variables ( $X_1$ ,  $X_2$ ) and the final exam mark is the dependent variable ( $Y$ ).

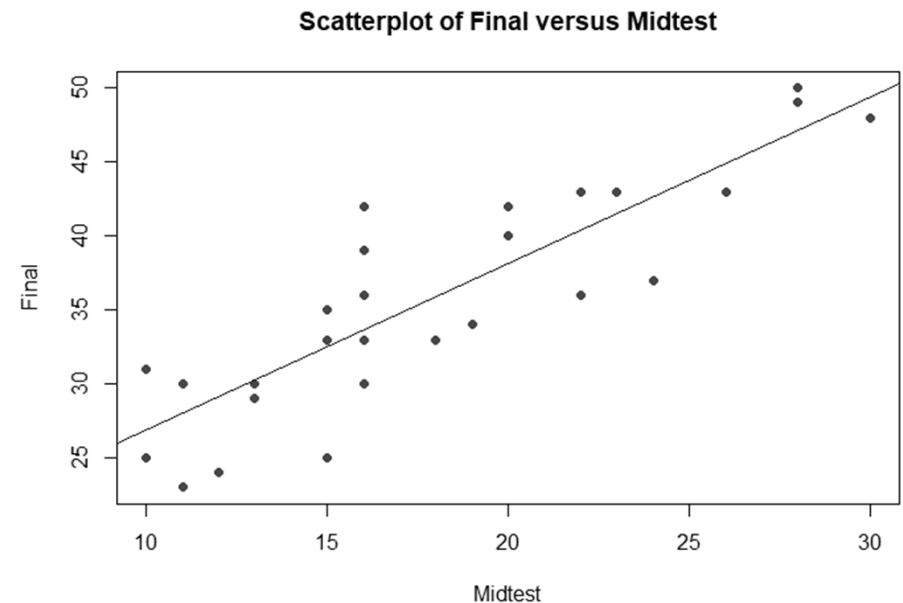
It seems logical to expect positive relationships between  $X_1$ ,  $X_2$  and  $Y$ .

a) Plot the dependent variable against each independent variable.

```
plot(Assignment, Final,  
     main = "Scatterplot of Final versus  
           Assignment",  
     col = "red", pch = 19)  
abline(lm(Final ~ Assignment), col = "blue")
```



```
plot(Midtest, Final,  
     main = "Scatterplot of Final versus  
           Assignment",  
     col = "darkgreen", pch = 19)  
abline(lm(Final ~ Midtest), col = "blue")
```



There seems to be no, or at best a weak positive linear relationship between the final and assignment marks. However, there seems to be a positive linear relationship between the mid-term test and final exam marks.

b) The straight lines on these scatterplots are the fitted straight lines obtained from two simple linear regressions. They do not substitute for a multiple linear regression, unless  $X_1$  and  $X_2$  are perfectly uncorrelated with each other, which is highly unlikely.

Regress the final exam mark ( $Y$ ) on the assignment mark ( $X_1$ ) and the mid-term test mark ( $X_2$ ) and interpret the coefficients.

```
m = lm(Final ~ Assignment + Midtest)
summary(m)

call:
lm(formula = Final ~ Assignment + Midtest)

Residuals:
    Min       1Q   Median       3Q      Max
-7.4061 -2.7218  0.2059  2.7598  8.6758

Coefficients:
(Intercept)  13.0091
Assignment     0.1940
Midtest       1.1121

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

Residual standard error: 3.752 on 27 degrees of freedom
Multiple R-squared:  0.7629,    Adjusted R-squared:  0.7453
F-statistic: 43.43 on 2 and 27 DF,  p-value: 3.647e-09
```

This sample regression equation implies that

- a. When *Assignment* = *Midtest* = 0, *Final* is expected to be 13.0.
- b. Given *Midtest*, for each additional *Assignment* mark *Final* is expected to increase by 0.194.
- c. Given *Assignment*, for each additional *Midtest* mark *Final* is expected to increase by 1.112.

→

$$\hat{y}_i = 13.009 + 0.194x_{1,i} + 1.112x_{2,i}$$

# GOODNESS OF FIT

- Suppose we have estimated a linear regression model of  $Y$  with OLS.
  - We have two estimators for  $Y$ , the sample mean,  $\bar{Y}$ , and the OLS estimator,  $\hat{Y}$ . Is it worth to use  $\hat{Y}$ , which is the more complicated and demanding estimator?

To answer this question, we need to compare the performances of the two estimators in terms of goodness of fit, i.e. how well the estimated models fit the observed  $Y$  values.

- It can be shown that the following relationship always holds:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$SST$		$SSR$		$SSE$
<i>Total sum of squares</i>	$=$	<i>Sum of squares due to regression</i>	$+$	<i>Sum of squares due to error</i>

$SST$  measures the total variation in  $Y$ , and it can be decomposed into two parts, such as,

$SSR$ : the amount of variation in  $Y$  that can be explained by the regression model, i.e., by the sample variations of the independent variables, and

$SSE$ : the amount of variation in  $Y$  that remains unexplained by the model.

Note: This decomposition is similar to the one we discussed in relation to ANOVA on week 5, slide 11 ( $SS = SST + SSE$ ).

Unfortunately, there are no uniformly accepted names and abbreviations for the sums of squares, different disciplines and even books in the same discipline often use different names and abbreviations.

We use the same names and abbreviations as the Selvanathan book, but  $SSR$  is also known as *explained sum of squares* ( $SEE$  or  $ESS$ ) or *model sum of squares* ( $MSS$ ), while  $SSE$  is also called *residual sum of squares* ( $SSR$  or  $RSS$ ) or *unexplained sum of squares* ( $USS$ ).



$$SST = SSR + SSE \longrightarrow 1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

- (Sample) coefficient of determination:

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

It measures the proportion of the total variation in  $Y$  that can be explained by the estimated regression model.

$R^2$  has the following properties:

- i.  $0 \leq R^2 \leq 1$
- ii.  $R^2 = 1$  if and only if each observation satisfies the sample regression equation without error, i.e. the fit is perfect.
- iii.  $R^2 = 0$  if and only if the model is completely useless.
- iv.  $R^2$  is a measure of goodness-of-fit, and the better the fit of the estimated model to the sample data, the closer  $R^2$  is to 1.
- v.  $R^2$  is the square of the Pearson (sample) correlation coefficient between  $Y$  and  $\hat{Y}$ .

- $R^2$  can be unrealistically high when the sample size ( $n$ ) is small relative to the number of independent variables ( $k$ ).
  - ← Every additional independent variable increases  $R^2$  even if it is insignificant, as long as the absolute value of its  $t$  value for  $H_0: \beta = 0$  is larger than one.

In order to avoid this overvaluation of the fit of the estimated regression equation,  $R^2$  has to be adjusted to  $n$  and  $k$ .

→ Adjusted (sample) coefficient of determination:

$$\bar{R}^2 = 1 - \frac{SSE / (n - k - 1)}{SST / (n - 1)} = 1 - \frac{n - 1}{n - k - 1} (1 - R^2)$$

This formula shows that when  $R^2 = 0$  (or very small)  $Adj. R^2 < 0$ .

Moreover, by rearranging this formula we obtain

$$\frac{n - 1}{n - k - 1} (1 - R^2) = 1 - \bar{R}^2 \longrightarrow 1 - R^2 < 1 - \bar{R}^2$$

Hence,  $Adj. R^2 < R^2$ , unless the fit is perfect and  $Adj. R^2 = R^2 = 1$ .

Note: On the  $R$  regression printout the unadjusted and adjusted  $R^2$  statistics are labelled as *Multiple R-squared* and *Adjusted R-squared*, respectively.

(Ex 2)

c) What conclusion can you draw from the unadjusted and adjusted  $R^2$  statistics?

From the printout, the unadjusted coefficient of determination is *Multiple R-squared* = 0.763. It shows that in this sample 76.3% of the total variation in the final exam marks can be explained by the variations in the assignment and mid-term test marks.

The adjusted coefficient of determination is *Adjusted R-squared* = 0.745.

$$\longleftarrow \bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2) = 1 - \frac{30-1}{30-2-1}(1-0.763) = 0.745$$

Hence, taking  $n$  and  $k$  into consideration, 74.5% of the total variation in the final exam marks can be explained by the two independent variables.

# THE ASSUMPTIONS OF LINEAR REGRESSION

- Similarly to the mean of a single population, say  $\mu_x$ , the unknown parameters of a population regression model ( $\beta_j, j = 0, 1, \dots, k$ ) can be estimated with point estimators and with interval estimators, and we can also use hypothesis testing to verify whether the sample at hand supports certain statements about these parameters.

To do so, first we have to study the sampling distributions of the  $\beta_j$ -hat OLS estimators.

Just as the sampling distribution of the sample mean,  $\bar{X}$ , depends on the mean, standard deviation and shape of the sampled  $X$  population, the sampling distributions of these point estimators depend on the properties of the  $\{Y_i\}$  sub-populations ( $i = 1, \dots, n$ ),

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

These properties are partly determined by the values assumed by the right-hand side variables. However, we never know these properties for sure because  $\varepsilon_i$  is an unobservable random variable and the  $x_{ij}$  values might be also randomly selected.

Hence, we need to make some assumptions and then check whether they are supported by the sample.

- Linear regression (on cross-sectional data) is based on six assumptions.

These assumptions are primarily about the conditional distributions of the  $\varepsilon$  error term, with implications for the conditional distributions of the dependent variable,  $Y$ , i.e. about

$$\varepsilon_i = \varepsilon \mid x_{1i}, x_{2i}, \dots, x_{ki} \quad \text{and} \quad y_i = Y \mid x_{1i}, x_{2i}, \dots, x_{ki}$$

*LR1:* We have a random sample of  $n > k + 1$  statistically independent but identically distributed (i.e., *iid*) sets of observations that satisfy the population regression model.

$$\longrightarrow y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

*LR2:* Each random error has zero conditional expected value, i.e.

$$E(\varepsilon_i) = 0 \quad \longrightarrow \quad E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$$

Under *LR2* the random error and the independent variables are uncorrelated, implying that

- (i)  $\varepsilon_i$  cannot be predicted from  $(x_{1i}, \dots, x_{ki})$ , and
- (ii) the effects of  $(x_{1i}, \dots, x_{ki})$  and  $\varepsilon_i$  on  $y_i$  can be separated from each other.

Without *LR2* we could not interpret  $\beta_j$  as the partial effect of  $X_j$  on the conditional expected value of  $Y$ .

*LR4*: The conditional variance of the random error is constant, i.e.

$$\boxed{Var(\varepsilon_i) = \sigma^2} \longrightarrow \boxed{Var(y_i) = \sigma^2}$$

This property is called homoskedasticity.

Otherwise, i.e., in the case of heteroskedasticity, the OLS method would fail to use all available information and hence the OLS estimators would not be efficient.

← OLS aims to minimize the sum of squared residuals treating all residuals equally important (equal weights).

However, observations drawn from populations that have relatively smaller variance are likely to be closer to their expected values and thus they are more valuable.

*LR4:* The conditional covariance between any two random errors is zero, i.e. for any  $i \neq i'$

$$E(\varepsilon_i \varepsilon_{i'}) = 0$$

—→ All pairs of random errors are uncorrelated.

*LR5:* In the sample (and hence in the population) there is no exact linear relationship among the right-hand-side variables, including the  $y$ -intercept.

—→ Each independent variable must assume at least two different values in the sample, and none of them can be perfectly substituted for by the others and the intercept.

Otherwise it would not be possible to identify the individual effects of the independent variables on the conditional expected value of the dependent variable.

- Under assumptions *LR1-LR5* and a given sample of the variables included in the model, the OLS estimators of the regression parameters have the usual ‘desirable properties’ of point estimators.

Namely (assuming that there is a  $y$ -intercept in the model):

1. The OLS estimators are unbiased.
2. The variances of the OLS estimators are directly related to  $\sigma^2$  and are inversely related to the sample variations of the independent variables and to the sample size.
  - Other things held constant, the smaller the variance of the random error and/or the more diverse and larger the sample, the smaller the variances of the OLS estimators are.
3. As the sample size increases indefinitely the variances of the OLS estimators approach zero, i.e., the OLS estimators are consistent.
4. The OLS estimators are linear functions of  $y_j, j = 1, 2, \dots, n$ , i.e., they are linear estimators.



In brief,

## Gauss-Markov Theorem: *BLUE*

Under assumptions *LR1-LR5*, the OLS estimators of the  $\beta$  regression parameters are the best in the class of all linear unbiased estimators.

- In addition to assumptions *LR1-LR5*, it is customary to make two further assumptions that facilitate statistical inference about the population regression model, especially when the sample size is relatively small.

*LR6:* The random errors are statistically independent of the independent variables.

*LR7:* The random errors are normally distributed, i.e.

$$\boxed{\varepsilon_i \sim N(0, \sigma^2)} \longrightarrow \boxed{y_i \sim N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}, \sigma^2)}$$

The random error term represents all those individually less important variables in the model that have effect on  $Y$  but are not considered explicitly in the model. According to a variant of the Central Limit Theorem, this combined effect is approximately normally distributed.

# TESTING HYPOTHESES ABOUT POPULATION PARAMETERS IN LINEAR REGRESSION

- The regression model is completely useless if none of the  $X_1, X_2, \dots, X_k$  independent variables is linearly related to  $Y$ , i.e. if each slope is zero.

$$\longrightarrow SSR = 0, SSE = SST \longrightarrow R^2 = 0$$

Otherwise, i.e., if at least one slope coefficient is different from zero,  $R^2 > 0$  and the model does have some utility. This possibility can be tested with the

*F*-test for overall significance of the regression model:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0; \longrightarrow E(Y|X_1, \dots, X_k) = \beta_0$$

$H_A$  : at least one  $\beta_i$  ( $i = 1, \dots, k$ ) is different from zero.

The test statistic is

$$F = \frac{MSR}{MSE} = \frac{SSR / k}{SSE / (n - k - 1)}$$

*MS:*  
*Mean Square*

which, under  $H_0$ , is *F*-distributed with  $df_1 = k$  and  $df_2 = n - k - 1$ .

A large  $F$  statistic indicates that  $MSR$  is large compared to  $MSE$ , and thus the regression model is useful.

→ Reject  $H_0$  and conclude that the model is significant if the calculated  $F$  ratio exceeds the critical value,  $F_\alpha$ .

The  $F$ -test statistic can be also given in terms of  $R^2$ ,

$$F = \frac{MSR}{MSE} = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$$

... and the hypotheses can be written as  $H_0 : \rho^2 = 0$  and  $H_A : \rho^2 > 0$ .

→ The  $F$ -test of overall significance can be equivalently interpreted as a test on the significance of the coefficient of determination.

(Ex 2)

d) Test the overall significance of the model with an  $F$ -test. Use  $\alpha = 0.01$ .

→  $H_0 : \beta_1 = \beta_2 = 0$  vs.  $H_A : \beta_1 \neq 0$ , or  $\beta_2 \neq 0$ , or  $\beta_1 \neq 0$  and  $\beta_2 \neq 0$

The  $F$ -test of overall significance is reported in the bottom part of the  $R$  regression printout (see slide #22) and it is reproduced on the next slide.

Residual standard error: 3.752 on 27 degrees of freedom  
Multiple R-squared: 0.7629, Adjusted R-squared: 0.7453  
F-statistic: 43.43 on 2 and 27 DF, p-value: 3.647e-09

The  $F$ -statistic is equal to 43.43 and its  $p$ -value is practically zero.

- We reject  $H_0$  and conclude at the 1% significance level that the two internal marks together have a significant influence on the final exam mark and hence the model itself is significant, i.e., it has some utility. We can equally conclude that  $R^2$  is significantly positive.
- As we discussed earlier (see slide #32), under  $LR1$ - $LR5$  the OLS estimators of the regression parameters are *BLUE*. When  $LR6$  is also satisfied, the OLS estimators are the best estimators out of all possible unbiased estimators, i.e., they are *BUE*.

Moreover, these OLS estimators are also normally distributed, i.e.,

$$\hat{\beta}_i \sim N(\beta_i, \sigma_{\hat{\beta}_i}) \longrightarrow \frac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}} \sim N(0, 1)$$

However, the standard errors of these estimators are unknown and they depend on  $\sigma$  (the standard deviation of  $\varepsilon$ ), which is also unknown.

→  $\sigma$  must be estimated from the sample, similarly to  $\beta_j$ ,  $j = 1, 2, \dots, k$ .

The estimate of  $\sigma$ , called the estimated standard error of regression, is the sample 'standard deviation' of the residuals:

$$s_{\varepsilon} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^n e_i^2}$$

On its own,  $s_{\varepsilon}$  is not really useful, but it is essential for interval estimation and hypothesis testing.

← Replacing  $\sigma$  with  $s_{\varepsilon}$ , the standardized OLS estimators become  $t$  random variables.

$$\leftarrow \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} \sim t_{df=n-k-1}$$

Given this sampling distribution, we can develop a confidence interval for and conduct a  $t$ -test about  $\beta_i$ , just as we do for  $\mu$  when  $\sigma$  is unknown.

(Ex 2)

e) Find the 95% confidence interval estimates of  $\beta_1$  and  $\beta_2$ .

Using  $t_{\alpha/2, n-k-1} = t_{0.025, 27} = 2.052$  from the  $t$  table and the point estimates and estimated standard errors from the  $R$  printout,

Coefficients:	$\hat{\beta}_i$	$s_{\hat{\beta}_i}$			
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	13.0091	3.5278	3.688	0.00101	**
Assignment	0.1940	0.2004	0.968	0.34172	
Midtest	1.1121	0.1219	9.120	9.87e-10	***

$$\hat{\beta}_1 \pm t_{\alpha/2, n-k-1} s_{\hat{\beta}_1} = 0.194 \pm 2.052 \times 0.200 = (-0.216 ; 0.604)$$

$$\hat{\beta}_2 \pm t_{\alpha/2, n-k-1} s_{\hat{\beta}_2} = 1.112 \pm 2.052 \times 0.122 = (0.862 ; 1.362)$$

These confidence intervals can be computed in  $R$  with the *confint* function:

	2.5 %	97.5 %
(Intercept)	5.7706640	20.2475022
Assignment	-0.2172731	0.6052515
Midtest	0.8618710	1.3622892

→ With 95% confidence, with every additional assignment mark the final exam mark is expected to change by about -0.217 to 0.605, and with every additional mid-semester test mark the final exam mark is expected to change by about 0.862 to 1.362 ('ceteris paribus').

f) Can Pat infer that the assignment mark and the mid-semester test mark are both linearly related to the final exam mark? Use  $\alpha = 0.05$ .

The question implies  $H_0 : \beta_i = 0$  and  $H_A : \beta_i \neq 0$  and the critical values are  $\pm t_{\alpha/2, n-k-1} = \pm t_{0.025, 27} = \pm 2.052$ .

$$t_{1,obs} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{0.194}{0.200} = 0.970$$

It is in between the two critical values, so  $H_0$  is maintained.

$$t_{2,obs} = \frac{\hat{\beta}_2}{s_{\hat{\beta}_2}} = \frac{1.112}{0.122} = 9.115$$

It is larger than the upper critical value, so  $H_0$  is rejected.

Or, from the *R* printout:

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.0091    3.5278    3.688  0.00101 **
Assignment    0.1940    0.2004    0.968  0.34172
Midtest      1.1121    0.1219    9.120 9.87e-10 ***
  
```

The two-tail  $p$ -value, i.e.,  $Pr(> |t|)$ , of *Assignment* is  $0.3417 > 0.05$ , so at the 5% significance level the first  $H_0$  is maintained. However, the two-tail  $p$ -value of *Midtest* is practically zero, so the second  $H_0$  can be rejected in favour of  $H_A$  at any reasonable significance level.

→ At the 5% significance level the mid-semester test mark is linearly related to the final exam mark, but the assignment mark is not.

g) Can Pat infer from the results that each of the independent variables, i.e., the assignment mark and the mid-semester test mark, is positively related to the final exam mark? Use  $\alpha = 0.05$ .

We need to perform right-tail  $t$ -tests on the slope parameters ( $i = 1, 2$ ) with  $H_0 : \beta_i = 0$  and  $H_A : \beta_i > 0$ .

Both slope estimates are positive and half of the reported  $p$ -value for the assignment mark is about 0.17 while that of the mid-semester test mark is zero.

→ There is a significantly positive linear relationship between the mid-semester test mark and the final exam mark, but the assignment mark does not appear to be positively related to the final exam mark.

Note: Although the assignment mark proved to be insignificant, it would be a mistake to drop it from the model automatically because statistical insignificance and theoretical irrelevance are not the same.

← A variable which is insignificant in a sample regression model might still be an important independent variable in the population regression model and could become significant when the model is re-estimated using a different sample.



# WHAT SHOULD YOU KNOW?

- To explain the differences between simple and multiple regressions.
- To estimate a linear regression model with  $R$  and to evaluate the results.
- The unadjusted and adjusted coefficients of determination.
- The assumptions of the linear regression model.
- To perform the  $F$ -test of overall significance and the  $t$ -test on individual regression parameters.

*For further information see, for example, Heij et al.: Ch 1-3.*