

Chapter 10

Simple Linear Regression

10.1 Introduction

Regression analysis may be defined as the analysis of relationships among variables. Regression is perhaps the most frequently used analytical tool in the physical and social sciences, in this, and the next two chapters, the focus is on the analysis of economic data or *econometrics*. Econometrics is the branch of economics concerned with the *empirical estimation* of economic relationships. The "Metric" part of the word signifies measurement; and econometrics is basically concerned with measuring economic relationships. Econometrics has been defined as the "...*application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results.*"¹

Econometrics utilises economic theory, as embodied in an **econometric model**; facts, as summarised by **relevant data**; and statistical theory, as refined into **econometric techniques**, in order to measure and test empirically certain relationships among economic variables, thereby giving empirical content to economic reasoning.

The Econometric Approach

The two basic ingredients of any econometric study are **theory** and **facts**. The principle task of econometrics is that of combining these two ingredients, using statistical techniques to estimate economic relationships.

Theory is one of the basic ingredients in any econometric study, but it must be developed in a usable form. The most usable form is typically that of a **model**, in particular an econometric model. The model summarises the theory relevant to the system under consideration, and it is the most convenient way of summarising this theory for empirical measurement and testing. An important aspect of econometrics and an essential part of any study is the **specification** of the model.

The other basic ingredient is a set of **facts**, referring to events in the real world relating to the phenomena under investigation. These facts lead to a set of data, representing observations of relevant facts. Usually the data needs to be refined, or "massaged" in a variety of ways before it is suitable for use. This refinement includes various

¹ Gerhard Tintner, *Methodology of Mathematical Economics and Econometrics*, Chicago University Press, 1968, p.74.

adjustments, such as seasonal or cyclical adjustments, extrapolation, interpolation, merging of different data sources, and the use of other information to adjust the data.

Combining theory with the refined data set requires the use of a set of **econometric techniques**. These are extensions of classical methods of statistics, particularly statistical inference (the use of sample information to infer certain characteristics of a population).

The result of the process is an estimated econometric model, in which certain magnitudes, known as **parameters**, are estimated on the basis of relevant data. The estimated model provides a way of measuring and testing relationships suggested by economic theory.

The econometric approach thus combines theory and facts in a particular way. From the viewpoint of theory, econometrics can be considered the application of "real-world" data to economic theory. Conversely, from the viewpoint of facts, econometrics can be considered a systematic way of studying economic history.

The three principle purposes of econometrics are illustrated in Figure 10.1. They are:

- structural analysis
- forecasting
- policy evaluation

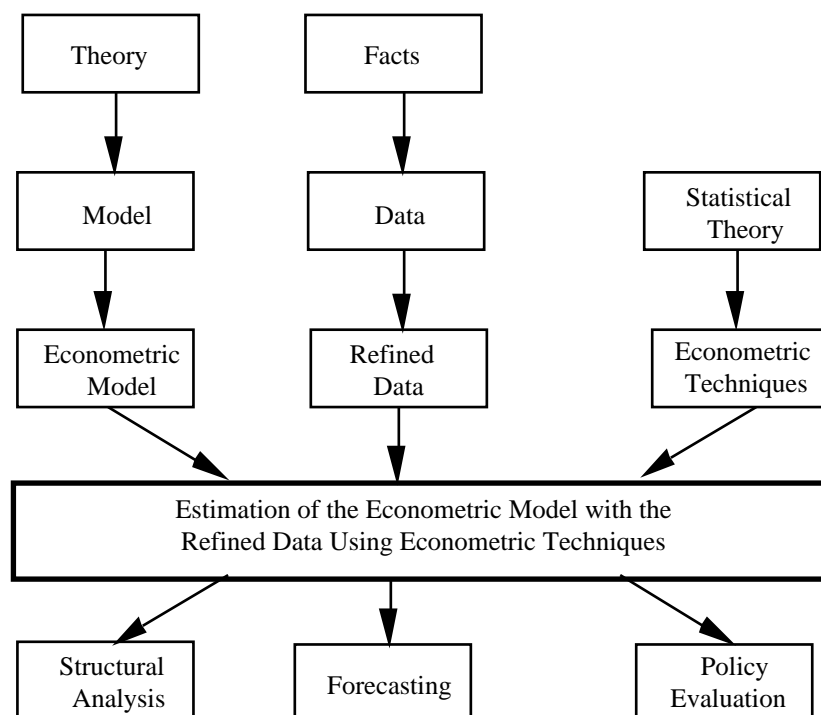


Figure 10.1 The Econometric Approach

Structural Analysis is the use of an estimated econometric model for the quantitative measurement of economic relationships. It represents what could be considered the "scientific" purpose of econometrics - that of understanding real-world phenomena by quantitatively measuring, testing and validating economic relationships. One result of the analysis may be a "feedback " influence on theory. For example, a measured relationship between the rate of inflation and the rate of unemployment, the Phillips curve, has led to various developments in the theory of unemployment. A second example could be the effect of property attributes on its price. The relationship

between property price and its area, the number of bedrooms, its location, etc., may be examined using an econometric model.

Forecasting is the use of an estimated econometric model to predict quantitative values of certain variables outside the sample of data actually observed. Forecasts may be the basis for action; for example, the employment of additional workers and the purchase of more raw materials in a firm may be based on a forecast that sales will increase in the next month. The decision to erect an office building may be based on forecasted rents.

Policy evaluation is the use of an estimated econometric model to choose between alternative policies. One approach is to introduce explicitly an objective function to be maximised by choice of policies and to regard the estimated model as a constraint in the optimisation process. An alternative approach might be to simulate different policies and make conditional forecasts of the future values of relevant variables for each policy. The selection of a most desired alternative among the various possible "candidate futures" would indicate which policy should be pursued. In either case, the selection of a particular policy, combined with the effects of those outside events that have an influence on the system, leads to specific outcomes. The outcomes, in turn, lead to another "feedback relationship" connecting policy evaluation with the facts, as shown in Figure 10.1.

These three principle purposes of econometrics are closely related. The structure determined by structural analysis is used in forecasting using an econometric model, while policy evaluation using an econometric model is a type of conditional forecast.

A Simple Model

A **model**, by definition, is any representation of an actual phenomenon such as an actual system or process. The actual phenomenon is represented by the model in order to explain it, to predict it, and to control it - purposes corresponding to the three purposes of econometrics discussed above: structural analysis, forecasting and policy evaluation.

Any model represents a compromise between reality and manageability. It must be a "reasonable" representation of the real world system and in that sense be "realistic" in incorporating the main elements of the phenomenon being represented. On the other hand, it must be manageable in that it yields certain insights or conclusions not obtainable from direct observations of the real-world system.

Striking the proper balance between realism and manageability is the essence of good modelling. A "good" model is both realistic and manageable. It specifies the interrelationships among the parts of a system in a way that is sufficiently detailed and explicit to ensure that the study of the model leads to insights concerning the real-world system. At the same time, however, it specifies them in a way that is sufficiently simplified and manageable to ensure that the model can be readily analysed and conclusions reached concerning the real-world system.

An **economic model** is a set of assumptions that approximately describes the behaviour of an economy (or sector of an economy). An **econometric model** consists of the following:

- (i) A set of behavioural equations derived from the economic model. These equations involve some observed variables and some "disturbances" (which are a catchall for all the variables considered as irrelevant for the purpose of this model as well as for all unforeseen events).

- (ii) A statement of whether there are errors of observation in the observed variables.
- (iii) A specification of the probability distribution of the "disturbances" and errors of measurement.

With these specifications we can proceed to test the empirical validity of the economic model and use it to make forecasts or use it in policy analysis.

As a general rule the first models of a phenomenon are quite simple, emphasising manageability at the cost of not treating reality in great detail. Suppose the phenomenon we are concerned with is the firm's sales of a particular product. Suppose that our theory tells us that sales and advertising are positively connected. We might express this theory in mathematical notation as follows:

$$Y = \alpha + \beta X + \epsilon \quad (\text{the behavioural equation})$$

where Y = sales of the firm's product
 X = advertising expenditure
 α = population parameters
 ϵ = a random error term that is not explained by the model.

The behavioural equation is assumed to represent the true underlying relationship between Y and X , also referred to as the population relationship. It is customary to use Greek letters to represent the population parameters α and β .

The parameters α and β are constants to be estimated. α is the magnitude of sales in the absence of advertising and β is the rate of change in sales associated with each unit of advertising. We expect both α and β to be positive numbers greater than zero. Sales, Y , are dependent on advertising, X . X is the independent or explanatory variable.

To estimate α and β , the population parameters, we would require all sales and all advertising for the particular product, in other words, the entire population. In most cases this is either impossible or impracticable. Consequently, we are usually content with **sampling** the population, that is, selecting at random certain members of the population.

The sample data may be "cross-section" or "time-series". Data for one time period for sales and advertising for sixty different products are cross-section. Alternatively, we might have data for the sales and advertising relating to one product over a period of time – time-series.

If the data is not available in published records produced by governments or private organisations, knowledge of survey and sampling techniques will be required. If the data problem is resolved we can proceed with the estimation of the equation using **regression** analysis. Regression will produce estimates of α and β in the following form:

$$Y = \hat{\alpha} + \hat{\beta}X$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimators of α and β respectively. Parameter estimates in the statistics or econometric literature are typically represented with a *hat* over the symbol for the parameter. A frequently used representation is to use the English alphabet equivalent, in the present example **a** and **b**, this is the representation used in this text.

Verification of an econometric model involves determining the "statistical significance" of the regression equation or model. One obvious type of verification is to examine the signs of **a** and **b**. If the sign of **b** was negative then we would seriously

question our theory or the data used to estimate the model. Alternatively, an error in model specification might have occurred. For example, the firm's sales might be more appropriately depicted by an equation of the form

$$Y = \beta_0 + \beta_1 X^2$$

In addition to whether or not β_1 is of the expected sign, we are interested in knowing whether β_1 is significantly different from zero. If not, then there is no statistically significant relationship between the dependent and independent variable. If there is no relationship then the equation may be depicted as in Figure 10.2.

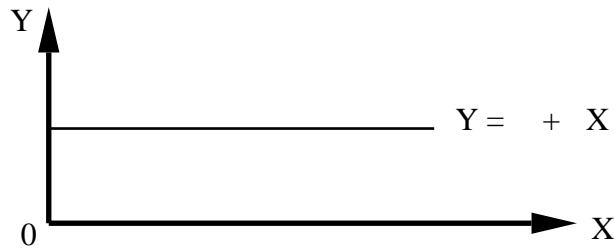


Figure 10.2 $Y = \beta_0 + \beta_1 X$ with $\beta_1 = 0$

Another aspect of verifying the model is how well the model "fits" the data. One measure of this is the **coefficient of determination**.

Forecasting: If we find there is a statistically significant relationship between Y and X, the model might serve for purposes of projection or prediction. Preparing a forecast and determining its accuracy are important aspects of econometric work.

The above methodology, which is illustrated in Figure 10.3, might lead us to reject or revise an existing theory, or develop a new theory. However, when a theory is rejected by statistical evidence, it only indicates one particular sample, one particular methodology, and one particular instance for which the theory was in error. Theory should only be modified when it fails in repeated statistical tests using alternative samples and methodologies.

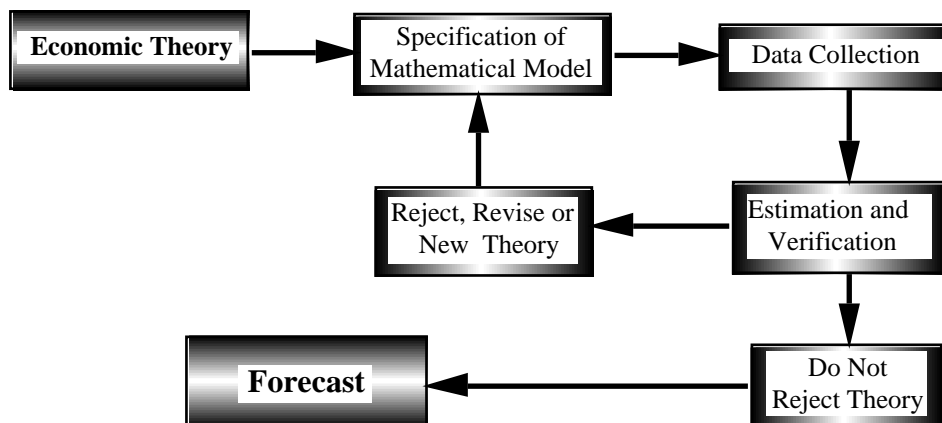


Figure 10.3 Methodology of Econometrics

10.2 Bivariate Regression

An econometric model consisting of one equation may be called a regression model, since the term regression² means a dependence relation. But, just as econometric

²The term *regression* is synonymous with Francis Galton, an amateur mathematician, who in 1875

models are a subclass of economic models, regression models are a subclass of statistical models. Thus, a regression model is a statistical dependence relation. The term regression may be interpreted as a dependence on the average. Regression analysis represents a statistical assessment of a dependent relationship. Methods available for the analysis are numerous, although the least-squares method is the most frequently used.

The Least-Squares Method

*The method of least squares is the automobile of modern statistical analysis; despite its limitations, occasional accidents, and incidental pollution, it and its numerous variations, extensions and related conveyances carry the bulk of statistical analysis, and are known and valued by all.*³

The least-squares method is in the nature of curve fitting. Suppose there are two variables X and Y whose linear dependence relation is given by

$$Y_i = \alpha + \beta X_i + \epsilon_i \quad (10.1)$$

in the population with ϵ_i being a random disturbance (also referred to as "a random error term"). The task is to find the estimates of the parameters α and β that will satisfy some given condition, in the case of least squares this condition is the *minimisation of the sum of the squared errors*. Under very general conditions the least-squares method is one of the best methods available for the estimation of the parameters α and β . This method satisfies a number of desirable properties discussed in detail below.

Equation (10.1) may be rearranged to make the random disturbance the subject of the expression as follows,

$$\epsilon_i = Y_i - \alpha - \beta X_i$$

The least-squares method minimises the sum of squared errors, estimation in the presence of error. Thus the approach embodies the concept of error (in the statistical sense) in the *true* model and it is in fact the existence of the error that enables the estimation of the model's parameters. If the random disturbance did not exist then the outcome would be a simple deterministic relationship between Y and X. In this case there would be no need for a statistical model.

Given that n pairs of observations are collected on (X,Y), that is a sample of size n, the least-squares method determines the values of the parameters, α and β to,

$$\text{Minimise } S = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \alpha - \beta X_i)^2 \quad (10.2)$$

To minimise the function S, represented by equation (10.2), obtain the partial derivative of S with respect to the parameters α and β and equate them to zero.

$$\begin{aligned} \frac{\partial S}{\partial \alpha} &= 2 \sum_{i=1}^n (Y_i - \alpha - \beta X_i)(-1) = 0 \\ \frac{\partial S}{\partial \beta} &= 2 \sum_{i=1}^n (Y_i - \alpha - \beta X_i)(-X_i) = 0 \end{aligned} \quad (10.3)$$

Equating the first derivative to zero ensures that a maximum, a minimum or turning point has been located. To verify which of these has been located it is necessary to

discovered regression to the mean after conducting experiments with parents and their offspring. He found that offspring had a tendency to 'regress to the mean' or average ancestral type.

³S.M. Stigler, "Gauss and the Invention of Least Squares", The Annuals of Statistics, Vol.9, No.3, 1981, pp.465-474.

take the second derivative of S.

$$\frac{\partial^2 S}{\partial a^2} = 2n \quad \text{and} \quad \frac{\partial^2 S}{\partial b^2} = 2 \sum X_i^2$$

In each case the second derivative is positive, $n > 0$ and $\sum X_i^2 > 0$, therefore a minimum has been located. Rearranging (10.3) as follows,

$$\begin{aligned} Y_i &= a + bX_i \\ X_i Y_i &= aX_i + bX_i^2 \end{aligned} \quad \begin{array}{l} \text{Normal} \\ \text{equations} \end{array} \quad (10.3)'$$

This produces two equations, the *normal equation*, which may be used to solve for the two unknown population parameters a and b . In practice sample data is used to estimate these population parameters. Their estimates are denoted by \hat{a} and \hat{b} , respectively.

Solving expression (10.3)' for the estimates of a and b , which we represent by \hat{a} and \hat{b} , we obtain

$$\hat{b} = \frac{n(\sum X_i Y_i) - (\sum X_i)(\sum Y_i)}{n(\sum X_i^2) - (\sum X_i)^2} \quad (10.4)$$

This may also be expressed in the form:

$$\hat{b} = \frac{(\sum Y_i - \bar{Y})(\sum X_i - \bar{X})}{(\sum X_i - \bar{X})^2} = \frac{S_{yx}}{S_{xx}} \quad (10.4)'$$

$$\hat{b} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} = \frac{\text{COV}(Y,X)}{\text{Var}(X)} \quad (10.4)''$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X} \quad (10.5)$$

The estimated values, \hat{a} and \hat{b} , may now be used to obtain the fitted values of Y_i , which are represented as \hat{Y}_i . For a given value of X_i the corresponding fitted value of Y_i , may be represented by equation (10.6),

$$E(Y_i/X_i) = \hat{Y}_i = \hat{a} + \hat{b}X_i \quad (10.6)$$

The expected value of Y_i associated with a given value of X_i is the conditional mean of Y_i , this is the fitted least-squares regression line, \hat{Y}_i , illustrated in Figure 10.4.

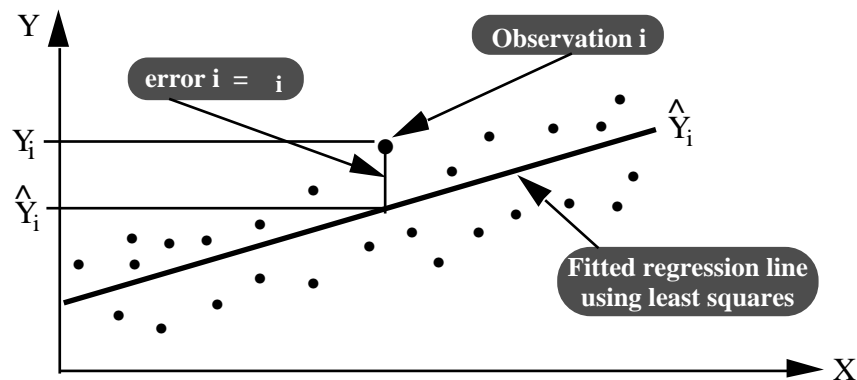


Figure 10.4 Deviations from Fitted Regression Line

10.3 The Gauss-Markov Theorem and OLS Assumptions

Theoretical Properties of Estimators – The *Gauss-Markov Theorem*

The population parameters, β_0 and β_1 , are unknown and sample data is used to obtain their estimates. The properties of these estimators are therefore of concern. The desirable properties of estimators are:

- (i) Unbiased
- (ii) Efficiency
- (iii) Consistency

Bias

If \mathbf{b} is the estimator of β then $\text{BIAS} = E(\mathbf{b}) - \beta$

Where $E(\mathbf{b})$ is the expected value of \mathbf{b} . The population is repeatedly sampled, at least theoretically, to obtain estimates of the true population parameter, β , from each sample. This produces a sampling distribution from which may be determined the expected value of \mathbf{b} .

$$E(\mathbf{b}) = \frac{1}{k} \sum_{i=1}^k \mathbf{b}_i, \text{ for } k \text{ samples} \quad (10.7)$$

Hence, bias is the difference between the average of all the estimators, \mathbf{b}_i , and the true population parameter, β . If $E(\mathbf{b}) - \beta = 0$ then \mathbf{b} is an unbiased estimator of β .

Efficiency

An estimator may be unbiased but not as "efficient" as another estimator. Efficiency is defined by the variance (VAR) of the sampling distribution of the estimator:

$$\text{VAR}(\bar{X}) = \frac{\sigma^2}{k} = \frac{(\bar{X} - \mu)^2}{k} \quad (10.8)$$

Where \bar{X} = mean of each sample, μ = population mean,
 k = number of samples in the sampling distribution.

The most efficient estimator is that estimator with the smallest variance.

Consistency

As the sample size gets larger the variance of the sampling distribution becomes smaller and smaller. This property is known as "consistency".

Estimators that possess these properties are said to be **BLUE** (best linear unbiased estimators).

Basic Assumptions Of The Least-Squares Model

The least-squares method requires that the following assumptions apply to all observations:

- (i) Normality: ϵ_i is normally distributed with a mean of zero.

$$\text{i.e. } E(\epsilon_i) = \frac{1}{n} \sum_{i=1}^n \epsilon_i = 0 \quad (10.9)$$

$$\text{and } E(Y_i) = E[\beta_0 + \beta_1 X_i + \epsilon_i] = \beta_0 + \beta_1 X_i$$

- (ii) ϵ_i has constant variance — **Homoscedasticity**

$$\begin{aligned}\text{VAR}(\epsilon_i) &= E[\epsilon_i - E(\epsilon_i)]^2 & (10.10) \\ &= E(\epsilon_i^2) = \sigma^2 \quad \text{since } E(\epsilon_i) = 0\end{aligned}$$

- (iii) ϵ_i has zero covariance — **no Autocorrelation**

$$\begin{aligned}\text{COV}(\epsilon_i, \epsilon_j) &= E[\epsilon_i - E(\epsilon_i)][\epsilon_j - E(\epsilon_j)] \\ &= E[\epsilon_i \epsilon_j] = 0, \quad i \neq j\end{aligned} \quad (10.11)$$

The error term associated with observation i is not correlated with the error term associated with observation j , this is referred to as serial independence.

- (iv) Nonstochastic X :

The observations on the independent variables, X_i , have values that are fixed in repeated samples, and for any sample size

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ is a finite non-zero number.} \quad (10.12)$$

- (v) Zero covariance between ϵ_i and X_i

$$\begin{aligned}\text{COV}(\epsilon_i, X_i) &= E[\epsilon_i - E(\epsilon_i)][X_i - E(X_i)] \\ &= E[\epsilon_i][X_i - \mu] \quad \text{since } E(\epsilon_i) = 0 \text{ and } E(X_i) = \mu \\ &= E[\epsilon_i X_i] - \mu E[\epsilon_i] \\ &= E[\epsilon_i X_i] \\ &= X_i E[\epsilon_i] \quad \text{since by assumption } X_i \text{ is nonstochastic.}\end{aligned}$$

$$\text{COV}(\epsilon_i, X_i) = 0 \quad (10.13)$$

- (vi) The model is linear in parameters. While the parameters cannot be non-linear, this does not preclude the presence of non-linear variables.

- (vii) The regression model is correctly specified.

This assumption requires that the variables included in the model are correct, what is the functional form of the model and what are the probabilistic assumptions made about Y_i , X_i , and ϵ_i entering the model.

The stochastic nature of the regression model implies that for every value of X there is a whole probability distribution of values of Y . This means that the values of Y can never be forecast exactly. The uncertainty concerning Y arises due to the presence of the stochastic disturbance ϵ_i which, being random, imparts randomness to Y .

Consider the example of a production function for a firm and assume that output depends in some specified way on the quantity of labour input in accordance with the engineer's blueprint. Such a production function may apply in the short run when the quantities of other inputs are fixed. But, in general, the same quantity of labour will lead to different quantities of output because of variations in weather, human performance, frequency of machine breakdowns, and many other factors. Output, which is the dependent variable in this case, will depend not only on the quantity of labour input, which is the explanatory variable, but also on a large number of random causes, which we summarise in the form of the stochastic disturbance. The probability distribution of Y and its characteristics are then determined by the values of X and the probability distribution of ϵ_i .

If the *blueprint* relation between output and labour were completely and correctly specified, then we would measure the value of β from the observations on X and Y after each production run. In reality this is almost never the case. In fact, we consider ourselves lucky when we know even the mathematical form of the relation without knowing the parameters. Typically, the mathematical form of the relation has to be assumed and the values of the parameters are estimated from observations on X and Y . Using the estimated values of the parameters, we can then "estimate" the values of the stochastic disturbance for each pair of values of X and Y .

We can now see that the full specification of the regression model includes not only the form of the regression equation (10.1) but also a specification of the probability distribution of the disturbance and a statement indicating how the values of the explanatory variable are determined. This information is provided by the **basic assumptions**.

The model (10.1) including the foregoing assumptions represents the so-called **classical normal linear regression model**, which provides a point of departure for most of the work in econometric theory.

Variation About the Regression Line

Now that the properties of the least-squares estimators for the regression coefficients have been considered, we need to consider how well the estimated line fits the sample data. To answer this question two related concepts are examined:

- (i) the standard error of the estimate, and
- (ii) the coefficient of (multiple) determination.

Standard Error of the Estimate (S_e)

The standard error of the estimate provides a measure of the average distance, of each observation, from the fitted line, that is, how the data is dispersed about the conditional mean, \hat{Y}_i .

$$S_e = \sqrt{\frac{(Y_i - \hat{Y}_i)^2}{n - (k+1)}} = \sqrt{\frac{e_i^2}{n-2}} \quad (10.14)$$

The sum of the squared errors is divided by $n-2$ since two degrees of freedom from our data are used up, having estimated \mathbf{a} and \mathbf{b} previously. The representation used for the number of estimated parameters is $k+1$, where k is the number of independent variables in the regression equation and 1 is added for the constant term.

This measure highlights those observations that are a long way from the fitted line as each error $(Y_i - \hat{Y}_i)$ is squared. It is desirable that the standard error be as "small as possible". By this is meant that it is small as a proportion of the observed values of Y or small when compared to this measure for competing regression models.

The Coefficient of Determination (R^2)

The coefficient of determination is a measure of "goodness of fit", how closely the fitted line is to the observations. This measure may be understood by reference to Figure 10.5.

$Y_i - \bar{Y}$ = Total deviation of an observation from the mean

$\hat{Y}_i - \bar{Y}$ = Deviation explained by the regression line

$Y_i - \hat{Y}_i$ = Deviation not explained by the regression line

$$\begin{array}{l} (Y_i - \bar{Y})^2 \\ \text{Total sum} \\ \text{of squares} \\ \text{SST} \end{array} = \begin{array}{l} (\hat{Y}_i - \bar{Y})^2 \\ \text{Regression sum} \\ \text{of squares} \\ \text{SSR} \end{array} + \begin{array}{l} (Y_i - \hat{Y}_i)^2 \\ \text{Error sum} \\ \text{of squares} \\ \text{SSE} \end{array}$$

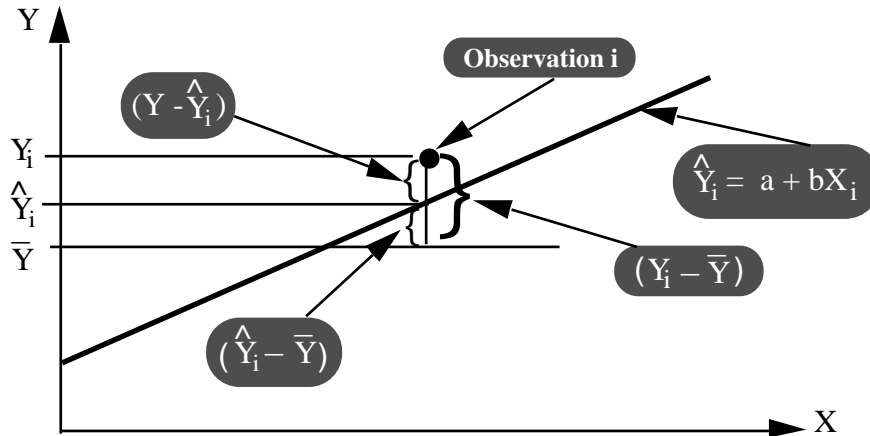


Figure 10.5 Variation About the Regression Line

Total variation in Y consists of the variation due to X (which is explained by the regression equation) and the variation due to the random component.

$$\text{SST} = \text{SSR} + \text{SSE}, \quad \text{or} \quad 1 = \frac{\text{SSR} + \text{SSE}}{\text{SST}}$$

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

$$R^2 = \frac{(\hat{Y}_i - \bar{Y})^2}{(Y_i - \bar{Y})^2} = 1 - \frac{(Y_i - \hat{Y}_i)^2}{(Y_i - \bar{Y})^2} \quad (10.15)$$

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

An R^2 close to 1 suggests that the estimated regression line fits the data well. If R^2 is close to 0 then none of the variation in Y is explained by the regression. The goal of regression analysis is not to achieve an $R^2 = 1$, indeed such a result would be inconsistent with the least-squares approach of estimation in the presence of error. While a *high* R^2 is desirable in a general sense, its magnitude will be determined by the data in question. If the data contains a good deal of randomness then the R^2 will not be close to 1, this result is typical for returns data or studies examining data containing a high degree of volatility. Hedonic price models used for valuation purposes are more likely to produce high R^2 .

Hypothesis Testing

Another technique for evaluating the explanatory power of the regression equation is the **F test**. The F test is a joint test for the model, that is, all of the parameters are jointly considered. In the case of the simple regression model an appropriate hypothesis might be:

$$\begin{aligned} H_0: & \beta = 0 && \text{(no relationship between X and Y)} \\ H_1: & \text{either } \beta > 0 \text{ or } \beta < 0 \text{ or both not zero} && \text{(there is a relationship)} \end{aligned}$$

If the F ratio, from the regression equation, is less than the critical F value (from tables) then accept H_0 and reject H_1 . That is, if $F < F_{\alpha, k, n - (k+1)}$ accept H_0 .

It is customary to state the null hypothesis, H_0 , in the manner stated above i.e. "no relationship exists between X and Y".

The F statistic is related to R^2 as follows

$$F = \frac{SSR/k}{SSE/[n - (k + 1)]} \quad (10.16)$$

or

$$F = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} \quad (10.16)'$$

The critical value for **F** is obtained from the F Distribution table using k degrees of freedom for the numerator and $n - (k+1)$ degrees of freedom for the denominator and the chosen value of α . The computed F is compared with the critical F (obtained from the F distribution tables). The computed F must be greater than the critical F if H_0 is to be rejected.

Test of Significance for the Regression Coefficients (t-test)

The t-test examines if each of the explanatory variables, X_i , (in a simple linear regression model there is only a single explanatory variable) are statistically significant in explaining the dependent variable Y. The t-ratios for the coefficient estimate, **b**, may be determined as follows:

$$t_{n - (k+1)} = \frac{b - \beta}{S_b} \quad \text{and} \quad S_b = \frac{S_e}{\sqrt{\sum X_i^2 - n\bar{X}^2}} \quad (10.17)$$

where

- b = the regression coefficient
- β = the hypothesised true regression parameter
- S_b = the standard error of the regression coefficient
- S_e = the standard error of the estimate
- n = sample size
- k = number of independent variables (X)
- $k+1$ = number of estimated regression coefficients, in this case **a** and **b**.

Test the hypothesis that $\beta = 0$

$$\begin{aligned} H_0: & \beta = 0 \quad \text{(no relationship between X and Y)} \\ H_1: & \beta \neq 0 \quad \text{(there is a relationship between X and Y)} \end{aligned}$$

The null hypothesis states that there is no relationship between X and Y therefore, the true population parameter, β , is zero. This may be illustrated with the aid of Figure 10.6 below.

Suppose we wish to test the hypothesis that no relationship exists between X and Y at a 5% level of significance— i.e. the **area** of the critical region in the diagram = 0.05, or the **area** of each tail is $\alpha/2 = 0.025$, since the test is a two-tailed test – H_0 is rejected if the computed t-ratio is greater than or less than the critical values.

Using a table for the normal distribution, assuming our sample size is at least 30, we need to look up a **z-value** corresponding to an area of:

$$0.5 - 0.025 = 0.4750$$

This corresponds to a z-value of 1.96. If our t-statistic is within the range -1.96 or $+1.96$ then we would accept the null hypothesis that no relationship exists between X and Y, otherwise, we accept the alternative hypothesis that there is a relationship.

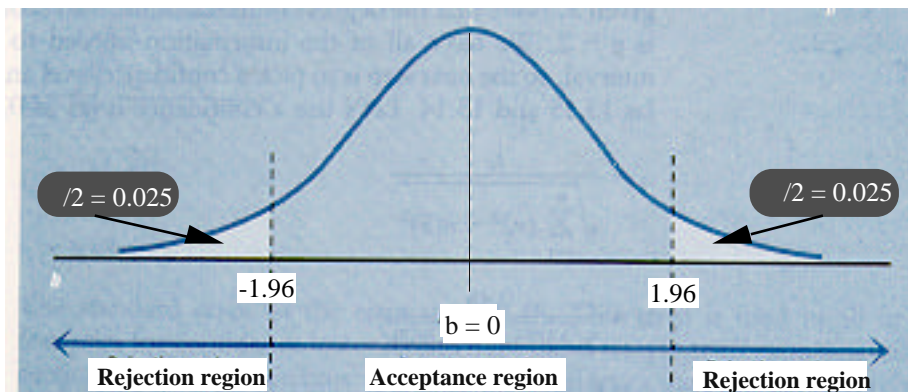


Figure 10.6. t-test for Regression Coefficient

10.4 Illustrative Example of Simple Regression

A hypothetical data set, Table 10.1, containing 11 observations for Y and X, is used to illustrate the simple linear regression model. The model attempts to explore the relationship between the sale price of residential properties and their areas, with area being the explanatory variable.

Sale Price	Area								
Y '000	X '000	YX	XX	\hat{Y}	$(X - \bar{X})^2$	$(Y - \hat{Y})^2$	$(Y - \bar{Y})^2$	$(\hat{Y} - \bar{Y})^2$	
5	4	20	16	4.84	25.92	0.03	34.92	36.85	
7	6	42	36	7.22	9.55	0.05	15.28	13.58	
11	9	99	81	10.80	0.01	0.04	0.01	0.01	
8	6	48	36	7.22	9.55	0.60	8.46	13.58	
7	6	42	36	7.22	9.55	0.05	15.28	13.58	
9	7	63	49	8.42	4.37	0.34	3.64	6.22	
12	11	132	121	13.19	3.64	1.41	1.19	5.18	
11	10	110	100	11.99	0.83	0.99	0.01	1.17	
14	12	168	144	14.38	8.46	0.14	9.55	12.03	
17	14	238	196	16.76	24.10	0.06	37.10	34.26	
19	15	285	225	17.95	34.92	1.09	65.46	49.64	
120	100	1247	1040	120.00	130.91	4.79	190.91	186.12	

Table 10.1 Data For Sale Price and Area

From equations (10.4) and (10.5) the estimates for a and b are:

$$b = \frac{n(\sum X_i Y_i) - (\sum X_i)(\sum Y_i)}{n(\sum X_i^2) - (\sum X_i)^2} = \frac{11(1247) - (100)(120)}{11(1040) - (100)^2} = 1.1924$$

$$a = \bar{Y} - b\bar{X} = \frac{120}{11} - 1.1924\left(\frac{100}{11}\right) \\ = 10.909 - 1.1924(9.0909) = 0.069$$

$$SST = \sum (Y_i - \bar{Y})^2 = \text{Total variation in Y} \quad 190.91$$

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2 = \text{Variation explained by regression model} \quad 186.12$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2 = \text{Unexplained variation in Y} \quad 4.79$$

$$S_e = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n - (k+1)}} = \sqrt{\frac{e_i^2}{n-2}} = \sqrt{\frac{4.79}{11-2}} = 0.7295$$

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{186.12}{190.91} = 0.975$$

Standard errors

$$S_a = S_e \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}} = 0.7295 \sqrt{\frac{1}{11} + \frac{(9.0909)^2}{130.91}} = 0.62$$

$$S_b = \frac{S_e}{\sqrt{\sum X_i^2 - n\bar{X}^2}} = \frac{0.7295}{\sqrt{1040 - 11(9.0909)^2}} = 0.0638$$

t-ratios

$$t_a = \frac{a}{S_a} = \frac{0.069}{0.62} = 0.111$$

$$t_b = \frac{b}{S_b} = \frac{1.1924}{0.0465} = 18.70$$

F-test

$$F = \frac{SSR/k}{SSE/[n - (k + 1)]} = \frac{186.12/1}{4.79/(11 - 2)} = 349.70$$

Presentation of Regression Results

Results are usually presented in a concise form, enabling the reader to get a complete picture at a glance. It is customary to place both the standard errors and t-ratios for the coefficients directly beneath the coefficient estimates. The following format is satisfactory for a single equation.

$$\hat{Y}_i = a + bX_i$$

Std. Err	(S_a)	(S_b)	$n =$, $df =$
t - ratio	(t_a)	(t_b)	$S_e =$, $R^2 =$, $F =$

Regression			
Results:	$\hat{Y}_i = 0.069 + 1.1924X_i$		
Std. Err	(0.620)	(0.0638)	$S_e = 0.7295$, $n = 11$, $df = 9$
t - ratio	(0.111)	(18.70)	$R^2 = 0.975$, $F = 349.70$

Interpretation of Results

The constant term, **a**, is the estimated value of Y_i when $X_i = 0$, i.e. when a property has no area. It is not realistic to think of a property without area, therefore this parameter has no economic interpretation. However, this does not mean the constant term should be excluded from the model. As discussed below, when a t-test for this coefficient is carried out, it will be shown that it is not significant in the model. It is important to retain the constant term in the regression so that the other coefficient, **b**, may be correctly estimated. If **a** is forced to be zero then the regression line is forced through the origin and the slope of the line, represented by the parameter **b**, will be incorrect. This is illustrated in Figure 10.7.

The coefficient of X, **b**, is the slope of the regression line. The value is positive, consistent with the belief that area contributes positively to price. The specific value of **b** indicates that for each additional unit of area (measured in thousands of square feet) sale price will increase by 1.1924 (i.e. \$1,192.4). The marginal contribution of area to price is 1.1924, this is, the rate of change in Y as X changes by 1 unit.

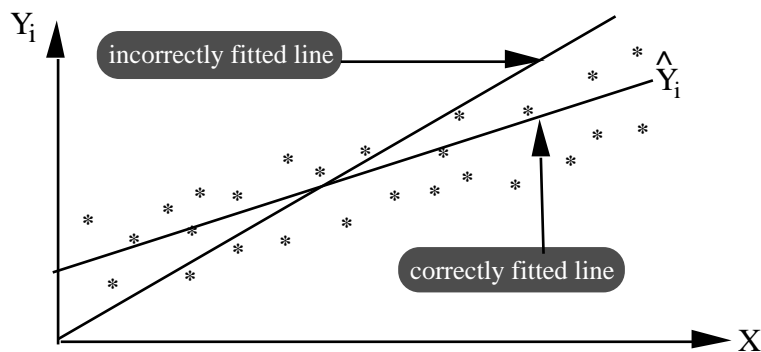


Figure 10.7: Regression Line Forced Through The Origin

Signs of the Coefficients

The coefficients are very important and provide an essential link between the theoretical model and the estimated equation. In the present example the signs for both **a** and **b** are positive. If, for example, the sign of **b** was negative then the model would be suggesting that as area increases property price decreases. This would not be consistent with expectations and it would be necessary to revise the model if in fact the sign was negative.

Standard error and Coefficient of Determination

The standard error of the estimate is relatively low indicating that the observations around the fitted regression line are not too widely dispersed. To verify this we need to consider the R^2 , which is an absolute measure. Since $R^2 = 0.975$, which is close to

If we conclude that the regression line fits the data well. Notice that R^2 is a measure of *goodness of fit* and it should not, on its own, be considered as a measure of explanatory power of the model.

If the model is theoretically based, that is, formulated from economic theory, and the estimated model conforms to the theoretical one, then we can be more confident of using R^2 as a measure of the explanatory power of the model. An R^2 close to 1, or close to 0, does not provide you with much information about the model. An R^2 close to zero, for example, could mean that there is a lot of *noise* (random variation) in the data and the model may well be an accurate representation of the problem being analysed. Due to the nature of the ordinary least squares fitting technique, which is based on the minimisation of the sum of squared errors, the value of R^2 should always be less than 1.

t-test

To establish whether or not the coefficients are significant it is necessary to conduct a t-test. The degrees of freedom are given by $n - (k+1) = 11 - 2 = 9$. A two tailed test is appropriate if the hypothesis for the population parameter is:

$$\begin{aligned} H_0: & \quad = 0 && \text{(no relationship exists between X and Y)} \\ H_1: & \quad \neq 0 && \text{(there is a relationship between X and Y)} \end{aligned}$$

The test statistic is constructed as follows:

$$t_b = \frac{b - 0}{S_b} = \frac{1.1924 - 0}{0.0638} = 18.70$$

Inspection of the t-tables for 9 degrees of freedom and a combined tail area of 0.05 gives a critical t-ratio of 2.262. The computed value of $t_b = 18.70$, hence we reject H_0 and accept H_1 , there is a relationship between X and Y. The computed t-ratio for **b** is very large indicating that the variable X is highly significant.

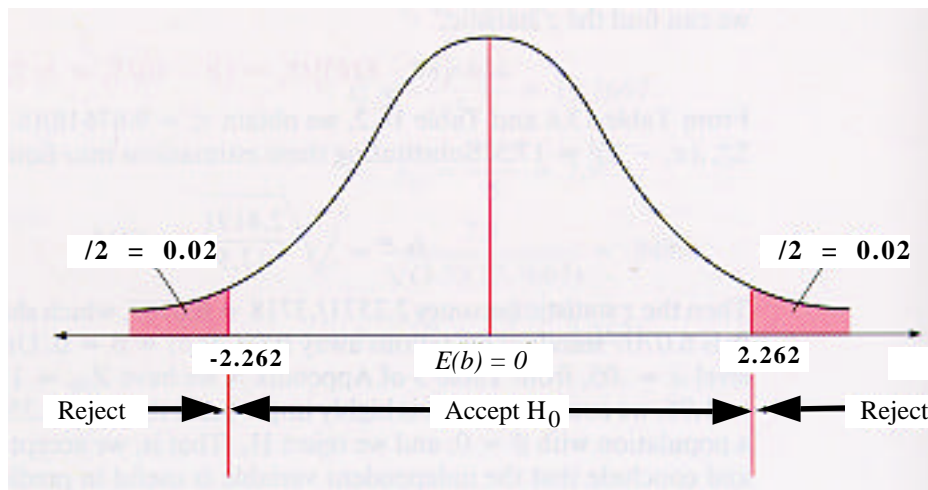


Figure 10.8: Hypothesis Test for Regression Coefficients

For the present example it is more appropriate to specify a one-tailed test, since the expectation is that $b > 0$. In this case the appropriate hypothesis would be:

$$\begin{aligned} H_0: & \quad \leq 0 && \text{(the true parameter is 0 or negative)} \\ H_1: & \quad > 0 && \text{(the true parameter is positive)} \end{aligned}$$

Since $t_b = 18.70$ H_0 is rejected in favour of H_1 .

If the same test is carried for the constant term, a , we find that the computed t-ratio of 0.111 falls in the acceptance region for H_0 . This outcome was foreshadowed previously when it was pointed out that a 's role in this model was to ensure that the equation is fitted correctly. It has no theoretical role, an expectation that is borne out by the empirical result.

F-test

The F-test is a joint test to determine if the model is significant. The calculated F statistic is compared with the critical F from the F-distribution table. A level of significance of 0.01 (a right tail area of $\alpha = 0.01$) is assumed.

Critical $F_{\alpha, k, n - (k+1)} = F_{0.01, 1, 9} = 10.6$. The hypothesis is:

$$H_0: \beta_1 = \beta_2 = 0 \quad \left(\begin{array}{l} \text{all coefficients are zero - the} \\ \text{model does not explain price} \end{array} \right)$$

$$H_1: \text{at least one of } \beta_1 \text{ or } \beta_2 \text{ not zero} \quad \left(\begin{array}{l} \text{the model does explain price} \end{array} \right)$$

The computed F statistic is, $F = 349.70$, which falls in the rejection region, refer Figure 10.9. Since the computed F is significantly greater than the critical F, $349.7 > 10.6$, we conclude that the equation is significant, reject H_0 and accept H_1 .

The model of the relationship between sale price and area meets all the statistical and theoretical requirements that are necessary so we would be inclined to accept it as a satisfactory model. Alternative functional forms of the model may be considered and compared with the results obtained for the present model in order to find a "best" model. Some of these models are discussed in a later chapter when functional form is examined in greater detail.

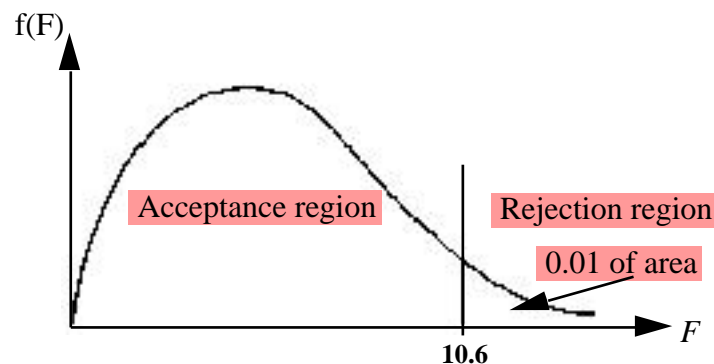


Figure 10.9: F-test for the Regression Equation

The modelling process, that is, endeavouring to specify a mathematical function that captures the underlying relationship between the variables of interest, takes a good deal of time and experience to master. Once the model has been specified and the data collected, estimating model parameters is a simple task using a computer. However, identifying the relevant variables, obtaining data for these variables and correctly specifying the underlying relationship that captures the data generating process presents the greatest challenge.

10.5 Regression Analysis Using Excel

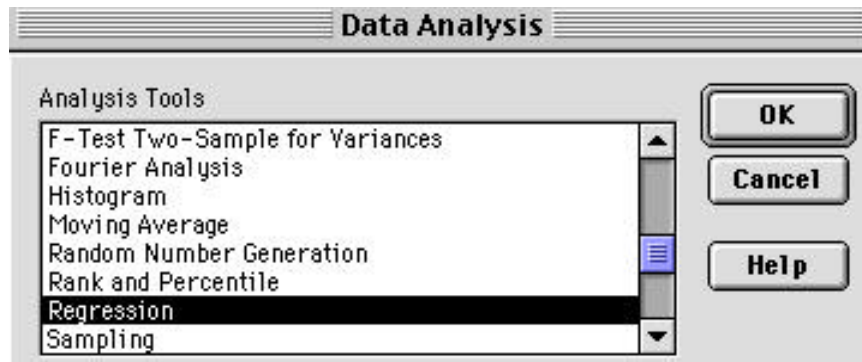
Excel's regression add-in may be used to estimate the model and may also be used to examine several questions about the model. This is demonstrated using the data from the previous example. To ensure that the Analysis ToolPak is installed in the version

of Excel you are using, refer to Section 2.3 of Chapter 2.

Having entered the data in the spreadsheet, from the **Tools** pull down menu, select **Data Analysis** and choose **Regression**.

Figure 10.10
Excel
Add-Ins...

Tools for
statistical
analysis.



Complete the regression dialog box as indicated in Figure 10.11. Click **OK** and the requested regression output will be placed in a new worksheet. On this occasion the diagnostic tools provided by Excel, analysis of the **Residuals** and the **Normal Probability** plots, have been selected. Each section of the output report will now be explained.

The **Input** and **Output Options** sections (required to create the standard regression output shown in Figure 10.12) of the regression dialog box must be completed. The **Residuals** and **Normal probability** sections are optional.

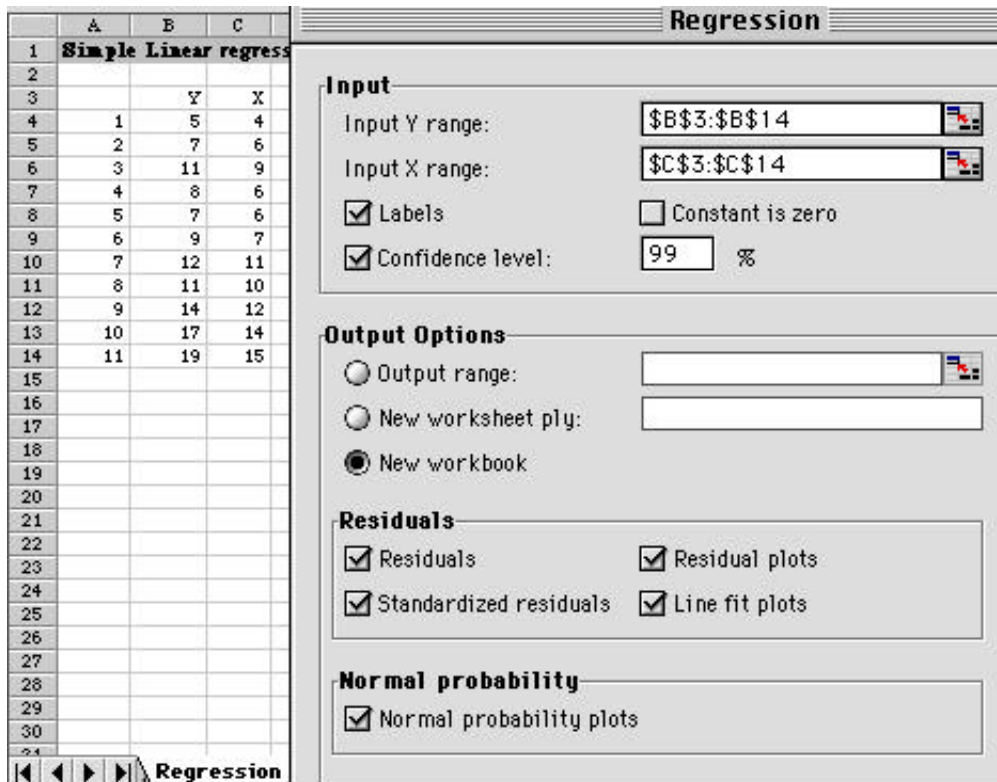


Figure 10.11 Completing the Regression Dialog Box

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.987							
R Square	0.975							
Adjusted R Square	0.972							
Standard Error	0.730							
Observations	11							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	186.117	186.117	349.525	1.6445E-08			
Residual	9	4.792	0.532					
Total	10	190.909						
	<i>Coefficients</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 99.0%</i>	<i>Upper 99.0%</i>
Intercept	0.06944	0.62014	0.11198	0.91330	-1.33341	1.47230	-1.94591	2.08480
X	1.19236	0.06378	18.69559	0.00000	1.04809	1.33664	0.98509	1.39963

Figure 10.12 Standard Regression Output for Valuation Model

Explanation of Regression Output

Regression Statistics

The *Multiple R* is the correlation between the dependent variable Y and the fitted line, \hat{Y} , and equal to $\sqrt{R^2}$. It is the correlation between the dependent variable, Y , and the linear combination of the independent variables, (in the present example a single independent variable) represented by \hat{Y} .

R Square is the ratio of the variation explained by the regression equation to the total variation, this defined in the previous section.

The *Adjusted R Square* is the original *R Square* adjusted for the loss in degrees of freedom. For a given sample size, as the number of explanatory variables included in the model increase, the available degrees of freedom decrease. This reduction in degrees of freedom is reflected in the *Adjusted R Square*. This is the appropriate statistic to report when multiple regression is used and is discussed in the next chapter.

The *Standard Error* (of the estimate) represents the average deviation of an observation from the regression line, see equation (10.14) in Section 10.3. The number of observations is the sample size used for the analysis.

ANOVA Table

The analysis of variance (ANOVA) is concerned with variance decomposition which is an alternative way of viewing regression. The total variance in Y , the squared deviation of Y from its mean, may be decomposed into two separate components, the variance explained by the regression model and the error or unexplained variance. This was represented in the Section 10.3 as:

$$\begin{aligned} (Y_i - \bar{Y})^2 &= (\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 \\ SST &= SSR + SSE \end{aligned} \quad (10.18)$$

The *F* statistic is the ratio of the explained variation to the unexplained variation, see equation (10.16) earlier. The column headed *SS* refers to the *sums of squares* associated with each component of the variance: Regression–*SSR*, Residual–*SSE* and Total–*SST*.

The *MS* column is formed by dividing the sums of squares by their respective degrees of freedom: *SSR* has k degrees of freedom (in the present example $k = 1$), *SSE* has $n -$

$(k+1)$ degrees of freedom ($11-2 = 9$), and SST has degrees of freedom equal to the sum of the degrees of freedom for SSR and SSE, $k + n-(k+1) = n-1$.

The F statistics is obtained from the MS value for the Regression (Regression Mean Square) divided by the MS value for the Residual (Residual Mean Square)—this is identical to equation (10.16).

The *significance F* is the probability of accepting the null hypothesis. Since the value is very small Excel provides it in E-notation, 1.6445E-08, the decimal place is moved 8 places to the left.

Coefficients, etc.

Excel provides a table listing the coefficients, standard errors, t-ratios p-values and confidence intervals. The coefficients are the estimated parameters of the regression model, their standard errors and t-ratios are also presented—these are identical to the values obtained in Section 10.4.

The p-value indicates the level of significance of the estimated parameter. If this value is less than the chosen level of significance, the decision rule is to reject H_0 . The choice of significance level depends on the trade off between Type I error, the cost of incorrectly rejecting H_0 , and Type II error, incorrectly accepting H_0 . If the cost of Type I error is relatively high, then a small value of α is chosen to increase the chance of accepting the null.

A 95% confidence interval is automatically provided by Excel. If a confidence level other than 95% is specified Excel will provide the confidence bands in the rightmost columns of this table. From Figure 10.11 it will be observed that a 99% confidence interval has been specified, this is provided in columns H and I of Figure 10.12. The critical value of t at the 99% confidence limit with 9 degrees of freedom is 3.25.

$$99\% \text{ confidence interval for } a : \quad \pm t(S_a) = 0.069 \pm 3.25(0.62) \\ -1.946 \quad 2.084$$

$$99\% \text{ confidence interval for } b : \quad \pm t(S_b) = 1.1924 \pm 3.25(0.0638) \\ 0.985 \quad 1.399$$

	A	B	C	D	E	F	G
22	RESIDUAL OUTPUT				PROBABILITY OUTPUT		
23							
24	<i>Observation</i>	<i>Predicted Y</i>	<i>Residuals</i>	<i>Standardized Residuals</i>		<i>Percentile</i>	<i>Y</i>
25	1	4.8389	0.1611	0.2327		4.5455	5
26	2	7.2236	-0.2236	-0.3230		13.6364	7
27	3	10.8007	0.1993	0.2879		22.7273	7
28	4	7.2236	0.7764	1.1215		31.8182	8
29	5	7.2236	-0.2236	-0.3230		40.9091	9
30	6	8.4160	0.5840	0.8436		50.0000	11
31	7	13.1854	-1.1854	-1.7124		59.0909	11
32	8	11.9931	-0.9931	-1.4345		68.1818	12
33	9	14.3778	-0.3778	-0.5457		77.2727	14
34	10	16.7625	0.2375	0.3431		86.3636	17
35	11	17.9549	1.0451	1.5097		95.4545	19

Figure 10.13 Excel's Residual and Probability Output Reports

The *Residual Output* contains the fitted (or predicted) values of Y, the residuals and the standardised residuals. The predicted values of Y are obtained from the regression equation by substituting each X_i in equation (10.189) to obtain the fitted values.

$$\hat{Y}_i = 0.069 + 1.1924X_i \quad (10.19)$$

The residuals are obtained by taking the difference between the observed values of Y_i

and the fitted values: $Y_i - \hat{Y}_i$.

The *Standard Residuals* are formed by dividing each residual by the residuals standard deviation, obtained using equation (10.20).

$$\text{Standard Residual} = \frac{e_i}{\sqrt{\frac{\sum_{i=1}^n (e_i - \bar{e})^2}{(n-1)}}} = \frac{e_i}{\sqrt{\frac{\sum_{i=1}^n (e_i^2)}{(n-1)}}} \quad (10.20)$$

The *Probability Output* section of the report is used to form the Normal Probability Plot. The observed values of Y are arranged in ascending order and the sample percentiles are calculated, the two variables are then plotted, Figure 10.14.

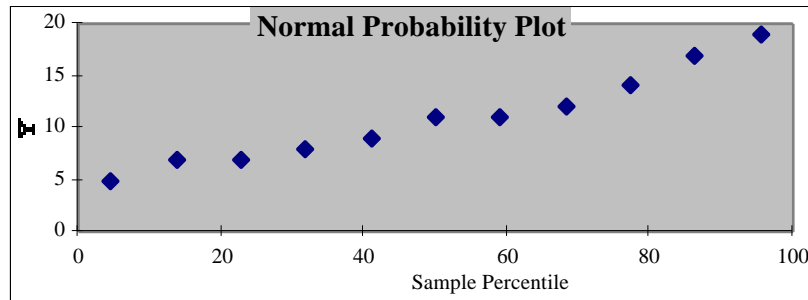


Figure 10.14 Normal Probability Plot for Y

If the plot is a straight line, or very close to a straight line, then the observed values of Y are said to be drawn from a normal distribution. One of the classical least squares assumptions is that the residuals from the model are normal, this assumption is examined below. Two additional charts are available as part of the output, they are provided as Figures 10.15 and 10.16.

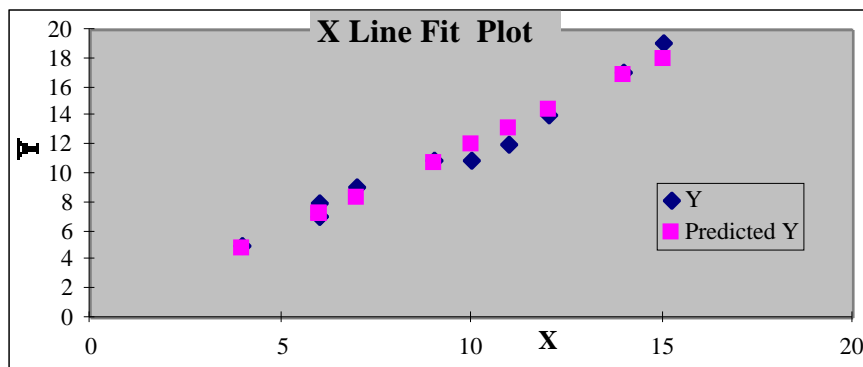


Figure 10.15 Plot of Actual and Predicted Y Values

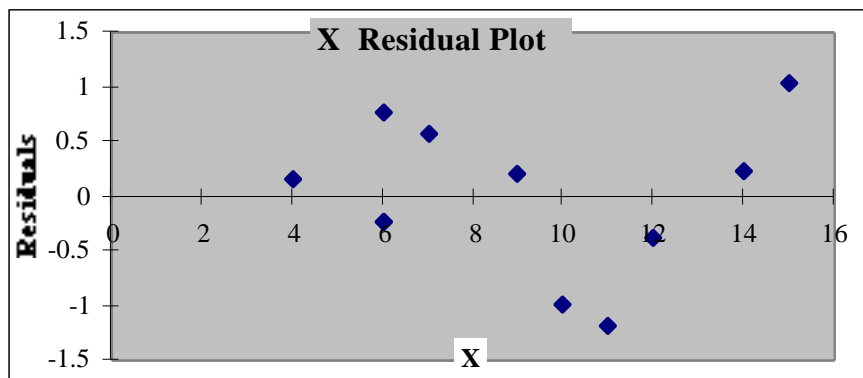


Figure 10.16 Plot of Residuals from the Model against the X Values


Figure 10.16 concerns another model assumption, namely that the errors are uncorrelated with X , see assumption (v), equation (10.13). Figure 10.16 would suggest that there is unlikely to be a relationship between the model residuals, e_i and X_i , since the points are randomly scatter about the zero line.

Constructing a Normal Probability Plot for the Errors

The residuals are ordered from lowest to highest (worksheet cells **D4 : D14**) and compared with the expected values based on the standard normal distribution. The expected values are obtained using the NORMSINV function in Excel.

Cell	Formula/Value	Copy to cells
A4	1	
A5	=1+A4	A6 : A14
B4	=A4/12	B5 : B14
C4	=NORMSINV(B4)	C5 : C14

Table 10.2 Calculation Worksheet for Normal Probability Plot

Invoke the Chart Wizard and plot the Z-values on the vertical axis and the ordered residuals on the horizontal axis—use  XY (Scatter) , this is shown in Figure 10.17.

Insert a straight line in the chart using Excel's **trendline** feature. Select the scatter plot by clicking on any one of the points. From the **Chart** pull down menu select **Add Trendline**, choose the **Linear** trend and click the **OK** button. The points are very close to the line, hence the residuals may be considered normal.

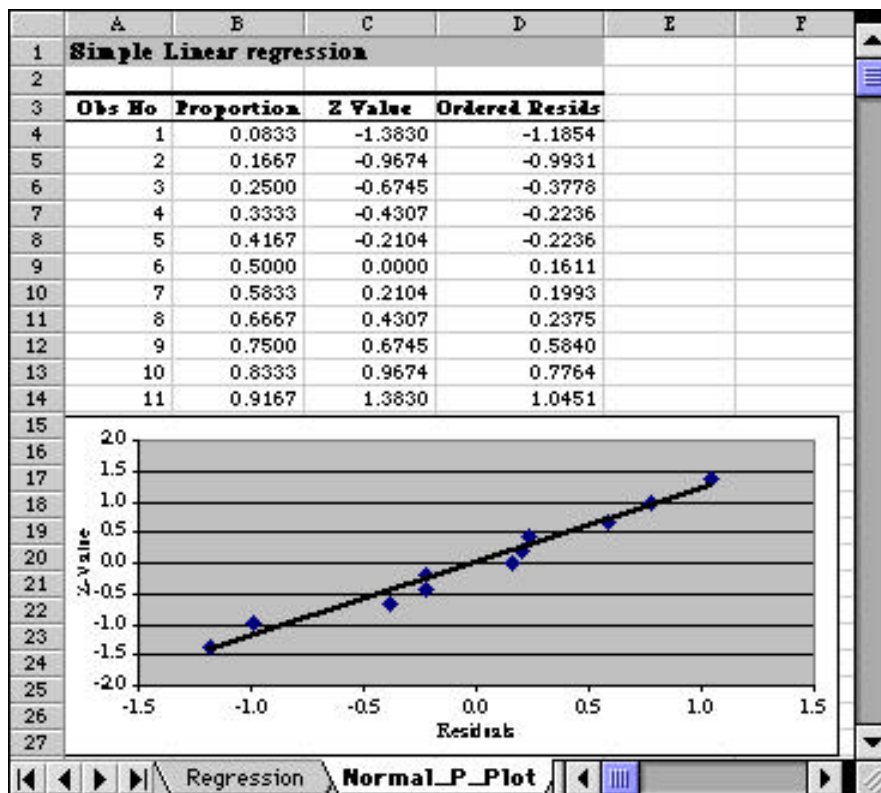


Figure 10.17 Normal Probability Plot for Regression Residuals

The simple correlation coefficient between the ordered residuals and their expected

values under normality is 0.985. This value is compared with the critical value from the tabulated values to be found at the end of the book. For $n = 11$ and $\alpha = 0.05$ the critical value is 0.923, therefore accept the hypothesis that the errors are normal.

Regression and Related Functions in Excel

Excel's built-in functions may be used to obtain several of the key regression statistics and make predictions based on linear regression. These functions will now be discussed and applied to the property price data contained in Table 10.1. The data has been entered in cells B4 : C14.

Number of observations	=COUNT(B4:B14)
Mean of Y	=AVERAGE(B4:B14)
Mean of X	=AVERAGE(C4:C14)
Intercept	=INTERCEPT(B4:B14,C4:C14)
Slope	=SLOPE(B4:B14,C4:C14)
R Square	=RSQ(B4:B14,C4:C14)
Standard Error of Estimate	=STEYX(B4:B14,C4:C14)
Correlation Coefficient	=CORREL(B4:B14,C4:C14)
SST	=VAR(B4:B14)*(COUNT(B4:B14)-1)
SSR (array formula)	=SUM(((INTERCEPT(B4:B14,C4:C14) +(SLOPE(B4:B14,C4:C14)*C4:C14)) -AVERAGE(B4:B14))^2)
SSE	=B26-B27
Forecast (X = 5) (array formula)	=FORECAST(5,B4:B14,C4:C14)

Functions may be embedded in each other to carry out calculations. For example, calculation of SSR requires a total of four functions and is entered as an array formula.⁴

The TREND function predicts Y values based on the regression equation formed from the data and the new X values provided. TREND is an array function, its arguments are:

=TREND(known_y's, known_x's, new_x's, const)

const is a logical argument, it is set equal to 1 if a constant term is required in the equation and set to zero if the constant is not required.

Select the cell range E10 : E14 and enter the formula

=TREND(B4:B14,C4:C14,D10:D14,1)

Hold the **Shift** and **Ctrl** keys down and press the **Enter** key.

The LINEST function is also an array function, its arguments are:

=LINEST(known_y's, known_x's, const, stats)

⁴Array formulae are entered by holding down the **Shift** and **Ctrl** keys and then hit the **Enter** key. Further information on array formulae may be found in Chapter 2, Section 2.7.

const is a logical argument, it is set equal to 1 if a constant term is required in the equation and set to zero if the constant is not required.

stats is a logical value specifying whether to return additional regression statistics. If additional statistics are required this value is set to 1.

LINEST returns the regression output in the following order:

b	a
S_b	S_a
R Square	S_e
F	df
SSR	SSE

To obtain this output, select a block of cells 5-rows by 2-columns. The LINEST function may also be used for multiple regression, the topic examined in the next chapter.

A closely related function is LOGEST which is used to estimate the regression parameters for an exponential curve. Its arguments are identical to the LINEST function.

	A	B	C	D	E
1	Simple Linear regression				
2					
3			Y	X	New X's Predicted Y's
4	1	5	4		
5	2	7	6		
6	3	11	9		
7	4	8	6		
8	5	7	6		
9	6	9	7	TREND Function	
10	7	12	11	6.5	7.8198
11	8	11	10	8	9.6083
12	9	14	12	12	14.3778
13	10	17	14	14	16.7625
14	11	19	15	16	19.1472
15					
16	Number of observations	11	LINEST Function		
17	Mean of Y	10.9091	b	a	
18	Mean of X	9.0909	S_b	S_a	
19	Intercept	0.0694	R Square	S_e	
20	Slope	1.1924	F	df	
21	R Square	0.9749	SSR	SSE	
22	Standard Error of Estimate	0.7297			
23	Correlation Coefficient	0.9874	1.1924	0.0694	
24	SST	190.9091	0.0638	0.6201	
25	SSR	186.1167	0.9749	0.7297	
26	SSE	4.7924	349.5251	9.0000	
27	Forecast (X = 5)	6.0313	186.1167	4.7924	

Figure 10.18 Using Regression Functions

Minimising SSE Using Solver



Figure 10.19 Solver Dialog Box to Minimise SSE

The *Least Squares* approach to estimating the regression parameters may be explored at a more intuitive level by those who lack an appreciation for calculus or wish to obtain an alternative view of this procedure. The relationship described by equation (10.6) is restated here as (10.21).

$$\text{Minimise } \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (Y_i - a - bX_i)^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (10.21)$$

For some arbitrary values of a and b the sum of squared errors may be calculated. Solver may then be employed to minimise this value by changing the values of a and b . The coefficients a and b are each assigned initial values of 2, cells C2 and C3, and then used to obtain the squared errors, cells C6 : C16. The function =SUM(C6 : C16) is placed in cell C18, the cell containing the value to be minimised. The completed solver parameters dialog box is shown in Figure 10.19. Click the **Solve** button and obtain the correct values for a and b .

10.6 The Capital Asset Pricing Model

An important application of regression analysis is the estimation of a security's *beta*. In chapter 5, section 5.8, a brief introduction to modern portfolio theory was provided. Risk and return for a particular asset were considered and the benefits of diversification illustrated. A more efficient method of evaluating an asset's systematic risk⁵ is by means of the Capital Asset Pricing Model (CAPM). The formal derivation of the CAPM is discussed in many corporate finance texts.⁶

For a two-asset portfolio return and risk may be represented by equations (10.22) and (10.23) respectively.

$$r_p = w_1r_1 + w_2r_2 \quad (10.22)$$

$$r_p = \sum_{i=1}^n w_i r_i \quad \{\text{for } n \text{ assets}\} \quad (10.22)'$$

⁵ Systematic risk is that risk that cannot be diversified away by forming portfolios of assets.

⁶ A more complete treatment of the CAPM and its derivation may be found in E.J. Elton & M.J. Gruber, 1995, *Modern Portfolio Theory and Investment Analysis*, fifth edition, John Wiley & Sons.

$$\sigma_p^2 = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2w_1w_2 \sigma_{12} \quad (10.23)$$

$$\sigma_p^2 = \sum_{i=1}^n w_i^2 \sigma_i^2 + \sum_{i=1}^n \sum_{j=1, j \neq i}^n w_i w_j \sigma_{ij} \quad \{\text{for } n \text{ assets}\} \quad (10.23)'$$

where σ_i^2 = variance of returns for asset i , $i = 1, 2$.
 σ_{12} = covariance of returns for assets 1 and 2.

w_i = proportion of investors funds allocated to asset i , $i = 1, 2$.
 $w_1 + w_2 = 1$ for 2 assets and $\sum_{i=1}^n w_i = 1$ for an n asset portfolio.

The number of variance terms in equation (10.23)' is n and the number of covariance terms is $n(n-1)/2$. Therefore, when the number of assets, n , is very large the number of covariance terms grows rapidly. If $n = 15$, the number of variance terms is 15 and the number of covariance terms is $15(15-1)/2 = 105$. The CAPM approach does not require this number of calculations.

To gain an intuitive understanding of the CAPM we begin by making a distinction between diversifiable and non-diversifiable risk. Diversifiable risk, sometimes referred to as idiosyncratic or unsystematic risk, can be eliminated through diversification. Diversifiable risk is that risk which is associated with a particular company or industry. A single risky investment exposes the investor to greater risk than if a diversification strategy is adopted. If a portfolio consists of a large number of assets that are uncorrelated with each other then unsystematic risk will be eliminated. Since this form of risk can be eliminated through diversification the market does not reward those investors who take it on – the return for taking on diversifiable risk is zero, the only risk worth considering is therefore systematic risk.

Systematic risk, or market risk, cannot be eliminated. This is due to the wider economic factors affecting the general economy. For example, the recession of the early 1990s has affected most industries, therefore, investing in several different industries will not remove the effects of the recession.

The underlying motivation of the CAPM is that there is a linear relationship between risk and return. To see this more clearly consider a situation where an investor has a portfolio, called m , consisting of a number of assets, the return on this portfolio is r_m and the variance is σ_m^2 . Assume now that there is a risk-free asset, such as a government bond, and the investor is able to borrow freely at the risk-free rate, r_f .⁷ The investor forms a new portfolio consisting of some combination of the original portfolio, m , and the risk free asset. The expected return on this new portfolio is given by

$$r_p = w_m r_m + (1 - w_m) r_f \quad (10.24)$$

where w_m is the proportion of funds invested in portfolio m , and $(1 - w_m)$ is the proportion invested in the risk free asset. The variance of the new portfolio is given by

$$\sigma_p^2 = w_m^2 \sigma_m^2 + (1 - w_m)^2 \sigma_f^2 + 2w_m w_f \sigma_{mf} \quad (10.25)$$

where σ_{mf} is the covariance between the expected return for portfolio m and the risk-free asset. By definition, the risk-free asset has zero variance and is uncorrelated with any other asset. Hence, $\sigma_f^2 = 0$, and $\sigma_{mf} = 0$ so the variance of the new portfolio is simply

$$\sigma_p^2 = w_m^2 \sigma_m^2 \quad (10.26)$$

⁷ In practice the existence of a risk-free asset is difficult to establish. Even a government bond with a guaranteed return does not protect against inflation or changes in the rate of interest.

or equivalently

$$r_p = w_m r_m + (1 - w_m) r_f \quad (10.26)'$$

Rearranging (10.26)' the following relationship is obtained for w_m

$$w_m = \frac{r_p - r_f}{r_m - r_f} \quad \text{and} \quad (1 - w_m) = 1 - \frac{r_p - r_f}{r_m - r_f}$$

Replacing w_m and $(1 - w_m)$ in equation (10.24) with these expressions produces

$$r_p = \frac{r_p - r_f}{r_m - r_f} r_m + \left[1 - \frac{r_p - r_f}{r_m - r_f} \right] r_f \quad (10.27)$$

Collecting terms this equation may be restated as

$$r_p = r_f + \left[\frac{r_m - r_f}{r_m - r_f} \right] (r_p - r_f) \quad (10.28)$$

giving a linear relationship between risk and return. The equation, representing the *ex-ante* view of returns, is a straight line with intercept r_f and slope $(r_m - r_f) / (r_m - r_f)$, this is illustrated in Figure 10.20. Notice that as the amount of risk increases the return increases to compensate for taking on the additional risk.

Before regression is used to estimate the parameters of equation (10.28) it is necessary to convert it to an estimable form. Consider a small portfolio, consisting of a single asset i , and a large well diversified portfolio, m , consisting of all the assets in the market. Rewriting equation (10.28) for asset i and the market portfolio m in the form of (10.29)

$$r_i = r_f + (r_m - r_f) \frac{i}{m} \quad (10.29)$$

or equivalently

$$(r_i - r_f) = (r_m - r_f) \frac{i}{m} \quad (10.29)'$$

The term $(r_i - r_f)$ is the risk premium for asset i , that is, the return associated with asset i in excess of the risk-free return, $(r_m - r_f)$ is the market risk premium, the return achieved by investing in the market portfolio over and above the risk-free return.

From (10.29)' it may be seen that the risk premium for asset i is equal to the market risk premium adjusted by the factor i / m . This factor of proportionality, i / m , indicates the dependence of asset i 's return on the market return. The proportionality factor is related to the *beta* of asset i , which is defined below.

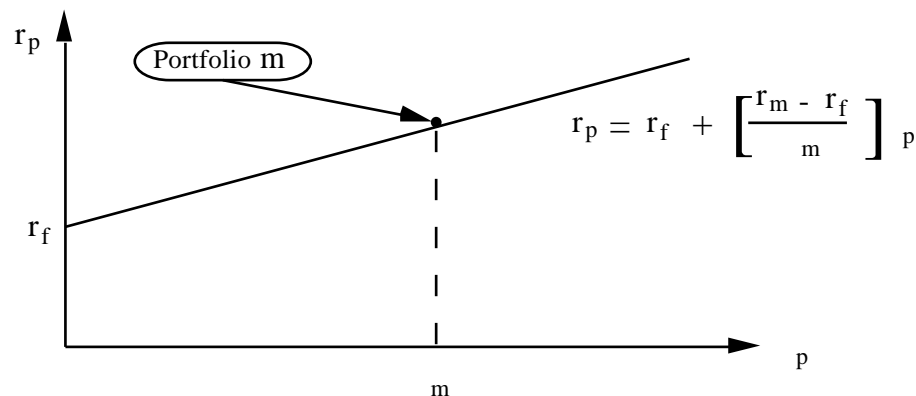


Figure 10.20 The Linear Relationship between Risk and Return

The general form of the bivariate regression model, specified in equation (10.1), is

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

The least squares estimates of the parameters, α and β , are obtained from,

$$\hat{\beta} = \frac{(Y_i - \bar{Y})(X_i - \bar{X})}{(X_i - \bar{X})^2} = \frac{S_{yx}}{S_{xx}} = \frac{\text{COV}(Y_i, X_i)}{\text{VAR}(X_i)} \quad (10.30)$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$$

Equation (10.29)' may be rewritten as a regression equation by adding the constant term α , the coefficient of the independent variable, β , and the error term, ϵ_i .

$$(r_i - r_f) = \alpha + \beta(r_m - r_f) + \epsilon_i \quad (10.31)$$

It is assumed that this equation satisfies the assumptions described in section 10.3. Using least squares the estimate of β is given by

$$\text{Beta} = \beta = \frac{\text{COV}[(r_i - r_f), (r_m - r_f)]}{\text{VAR}(r_m - r_f)} = \frac{\text{im}}{\text{m}^2} \quad (10.32)$$

Equation (10.31) may be used to estimate the beta for any asset – beta can also be obtained from the ratio of the asset's covariance with the market, im , to the market variance, m^2 .

Equation (10.32) indicates that when an asset's returns are perfectly positively correlated with the returns from investing in a market portfolio, then β will equal 1 ($\text{im} = \text{m}^2$). Beta is therefore a measure of relative risk, the betas for all assets are compared with 1 to determine their systematic risk relative to the market. Thus β is a measure of the riskiness of an asset relative to the market portfolio. A beta greater than 1 indicates that investing in this asset is more risky than investing in a market portfolio – the sensitivity of asset i to the prevailing economic conditions is greater than that of the general market.

Asset i might represent shares in a particular company and a change in a macro economic variable, such as a reduction in interest rates, will have a greater impact on this company's share price than for the average of companies across all industries. High risk companies, like mining companies, typically have betas greater than 1 while low risk companies, such as public utilities, are inclined to have betas less than 1.

The market portfolio used is generally some index, representing a large number of assets, such as the All Ordinaries (for Australia) or the Dow Jones (for the US).

Estimating an Asset's Beta

To estimate the beta for a particular asset returns data are required for the asset in question, a market index, and the returns on a risk-free asset. Ten years of monthly returns (US data) for the electronics company Tandy, from January 1978 to December 1987 are used for the present example.⁸ Data for the risk-free rate is the return on 30-day US Treasury bills. The market index is a value-weighted composite monthly market return based on transactions from the New York Stock Exchange and the American Exchange over the same ten year time span – this data is taken from the Centre for Research on Security Prices (CRSP).

The estimation of β and α is usually carried out using time series data. It is assumed

⁸The data used for the analysis is included at the end of the exercises. A spreadsheet containing this data is available at the Web site for this text: <http://www.bf.rmit.edu.au/quant>

that the beta of an asset remains fairly constant over time. The data set is used to verify this contention by breaking the data into two five-year periods, estimating two separate regression equations and then compare the results. The estimated regression equations for each time period are provided in Table 10.3.

January 1978 to December 1982	January 1983 to December 1987
$(r_i - r_f) = 0.031 + 1.03(r_m - r_f)$ t - ratios (1.99) (5.09)	$(r_i - r_f) = -0.009 + 1.035(r_m - r_f)$ t - ratios (-0.84) (5.44)
$R^2 = 0.309, S_e = 0.119, n = 60$ $F = 25.93, DW = 1.95,$	$R^2 = 0.338, S_e = 0.087, n = 60$ $F = 29.61, DW = 1.96,$

Table 10.3 Comparison of Security Beta Across two Time Periods

Both equations generate a value of beta that is consistent and of the same magnitude suggesting that this coefficient has remained fairly constant across the two time periods. A formal test to establish if beta from the first time period, β_{1i} , is significantly different from the beta in the second time period, β_{2i} , is given by (10.33)

$$t = \frac{(\beta_{1i} - \beta_{2i})}{SE(\beta_{1i} - \beta_{2i})} = 0.0145 \quad (10.33)$$

This test indicates that the coefficient, β_i , is not significantly different across the two time periods. Notice that this test does not say anything about the equations associated with each time period being different, simply that the coefficient β_i is not significantly different between the two time periods.

Our hypothesis for the CAPM would suggest that the constant term should be insignificant and that should be significant. An examination of the t-ratios indicates that while is highly significant, is also significant at the 5% level in the first regression. This is not as expected, it may be that the constant term is capturing some factor that is affecting Tandy's risk premium. An examination of the events occurring during the two time periods and how they affected Tandy's share prices might provide some useful insights at this point.

Another point to note about these regressions is the R^2 , it is very low. This is not too surprising when we realise that returns data has a lot of variability causing the random error term to play a dominant role in the model. The value of R^2 is typically of this order of magnitude for the CAPM.

10.7 Regression Using Matrix Algebra

Matrix algebra is a very powerful tool in regression analysis and is used extensively for multiple regression. To provide a preview of the next chapter we will use the example from the previous section to introduce the application of matrices in regression.

The normal equations, given by (10.3) above, may be represented in matrix form as,

$$\begin{matrix} Y_i \\ X_i Y_i \end{matrix} = \begin{matrix} n & X_i \\ X_i & X_i^2 \end{matrix} \begin{matrix} a \\ b \end{matrix} \quad (10.34)$$

This may also be expressed in the form $X'Y = (X'X)b$

To solve for **a** and **b**, the vector of unknown coefficients, **b**, the inverse of the (X'X) matrix is obtained, represented as (X'X)⁻¹. The solution to this system of equations is given by

$$\underline{b} = (X'X)^{-1}X'Y \quad (10.35)$$

$$(X'X) = \begin{matrix} n & X_i \\ X_i & X_i^2 \end{matrix} = \begin{matrix} 11 & 100 \\ 100 & 1040 \end{matrix}$$

$$X'Y = \begin{matrix} Y_i \\ X_i Y_i \end{matrix} = \begin{matrix} 120 \\ 1247 \end{matrix}$$

$$\begin{aligned} \text{Determinant} = |(X'X)| &= n X_i^2 - (X_i)(X_i) \\ &= n(X_i - \bar{X})^2 = 11(130.91) = 1440.01 \end{aligned}$$

$$\text{Cofactor } (X'X) = \begin{matrix} X_i^2 - X_i & -100 \\ -X_i & n \end{matrix} = \begin{matrix} 1040 & -100 \\ -100 & 11 \end{matrix}$$

The (X'X) matrix is symmetric and its inverse is also symmetric. The cofactor and adjoint matrices are identical for symmetric matrices.

$$\text{Inverse} = (X'X)^{-1} = \frac{1}{n(X_i - \bar{X})^2} \begin{matrix} X_i^2 - X_i & -X_i \\ -X_i & n \end{matrix}$$

Substituting the values previously calculated for this matrix,

$$(X'X)^{-1} = \frac{1}{1440.01} \begin{matrix} 1040 & -100 \\ -100 & 11 \end{matrix}$$

$$\begin{aligned} \begin{bmatrix} a \\ b \end{bmatrix} &= (X'X)^{-1}X'Y \\ &= \frac{1}{1440.01} \begin{matrix} 1040 & -100 \\ -100 & 11 \end{matrix} \begin{matrix} 120 \\ 1247 \end{matrix} \\ &= \frac{1}{1440.01} \begin{matrix} 100 & 0.069 \\ 1717 & 1.1924 \end{matrix} \end{aligned}$$

The variance-covariance matrix = S_e²(X'X)⁻¹

$$S_e^2(X'X)^{-1} = (0.7295)^2 \begin{matrix} \frac{1040}{1440.01} & \frac{-100}{1440.01} \\ \frac{-100}{1440.01} & \frac{11}{1440.01} \end{matrix}$$

From the diagonal elements we obtain the standard errors for S_a and S_b.

$$S_a^2 = (0.7295)^2 \left[\frac{1040}{1440.01} \right] = 0.3843 \text{ and } S_a = \sqrt{0.3843} = 0.62$$

$$S_b^2 = (0.7295)^2 \left[\frac{11}{1440.01} \right] = 0.00407 \text{ and } S_b = \sqrt{0.00407} = 0.0638$$

The t-ratios may then be computed in the usual way:

$$t_a = \frac{a}{S_a} = \frac{0.069}{0.62} = 0.111$$

$$t_b = \frac{b}{S_b} = \frac{1.1924}{0.0465} = 18.70$$

The use of matrix algebra in regression analysis is discussed more formally in the next chapter where the multiple regression model is introduced.

Appendix 10 The Sums of Squares

The deviation of Y_i from its mean consists of two parts; the deviation of the regression line from the mean of Y_i and the deviation of Y_i from the regression line.

Deviation from mean of Y: $Y_i - \bar{Y}$

Deviation of fitted line from mean of Y: $\hat{Y}_i - \bar{Y}$

Deviation of Y from fitted line: $Y_i - \hat{Y}_i$

$$(Y_i - \hat{Y}_i) = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y}) \quad (\text{A10.1})$$

The sum of squares of the deviations is represented by equation (A10.2).

$$(Y_i - \hat{Y}_i)^2 = (Y_i - \bar{Y})^2 - (\hat{Y}_i - \bar{Y})^2 \quad (\text{A10.2})$$

SSE
SST
SSR

The proof of this relationship may be established by squaring both sides of A10.1.

$$\begin{aligned} (Y_i - \hat{Y}_i)^2 &= \left[(Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y}) \right]^2 \\ &= (Y_i - \bar{Y})^2 - 2(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2 \end{aligned}$$

$$\begin{aligned} -2(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) &= -2(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})(\hat{Y}_i - \bar{Y}) \\ &= -2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) - 2(\hat{Y}_i - \bar{Y})^2 \end{aligned}$$

$$\begin{aligned} (Y_i - \hat{Y}_i)^2 &= (Y_i - \bar{Y})^2 - 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) - 2(\hat{Y}_i - \bar{Y})^2 + (\hat{Y}_i - \bar{Y})^2 \\ &= (Y_i - \bar{Y})^2 - 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})^2 \end{aligned}$$

Adding the summation operator

$$(Y_i - \hat{Y}_i)^2 = (Y_i - \bar{Y})^2 - 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})^2 \quad (\text{A10.3})$$

Since $\hat{Y}_i = a + bX_i$ and $a = \bar{Y} - b\bar{X}$

Thus $\hat{Y}_i = \bar{Y} + b(X_i - \bar{X})$

Also, $b = \frac{(Y_i - \bar{Y})(X_i - \bar{X})}{(X_i - \bar{X})^2}$

Substituting for \hat{Y}_i in the middle part of equation A10.3:

$$\begin{aligned} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \left[Y_i - (\bar{Y} + b(X_i - \bar{X})) \right] \left[\bar{Y} + b(X_i - \bar{X}) - \bar{Y} \right] \\ &= \left[Y_i - \bar{Y} - b(X_i - \bar{X}) \right] \left[b(X_i - \bar{X}) \right] \\ &= \left[(Y_i - \bar{Y})b(X_i - \bar{X}) - b^2(X_i - \bar{X})^2 \right] \\ &= b \left[(Y_i - \bar{Y})(X_i - \bar{X}) - b (X_i - \bar{X})^2 \right] \end{aligned}$$

Substituting for b inside the brackets

$$\begin{aligned} &= b \left[(Y_i - \bar{Y})(X_i - \bar{X}) - \frac{(Y_i - \bar{Y})(X_i - \bar{X})}{(X_i - \bar{X})^2} (X_i - \bar{X})^2 \right] \\ &= b \left[(Y_i - \bar{Y})(X_i - \bar{X}) - (Y_i - \bar{Y})(X_i - \bar{X}) \right] = 0 \end{aligned}$$

Therefore, $(Y_i - \hat{Y}_i)^2 = (Y_i - \bar{Y})^2 - (\hat{Y}_i - \bar{Y})^2$

That is, $SSE = SST - SSR$

References

- Berk, K.N., & Carey, P., 1998, *Data Analysis with Microsoft Excel*, Duxbury Press. Chapter 8.
- Berndt, E.R., 1991, *The Practice of Econometrics Classical and Contemporary*, Addison-Wesley Publishing Company. Chapter 2.
- Gujarati, D.M., 1996, *Basic Econometrics*, Third edition, McGraw-Hill International Editions. Chapters 1, 2, 3 and 4.
- Lee, C.F., 1993, *Statistics for Business and Financial Economics*, D.C. Heath and Company. Chapters 13 and 14.
- Levine, D.M., Berenson, M.L. & Stephan, D., 1997, *Statistics for Managers Using Microsoft Excel*, Prentice-Hall, Inc. Chapter 11.
- Lombardo, R., 1998, *Property Data Analysis - A Primer*, Property Studies Education Unit, RMIT, Melbourne, Chapters 11 and 12.

- Lusht, K.M., 1997, *Real Estate Valuation, Principles and Applications*, Richard D. Irwin. Chapter 9
- Middleton, M.R., 1997, *Data Analysis Using Microsoft Excel*, Duxbury Press. Chapter 14.
- Neter, J., Kutner, M.H., Nachtsheim, C.J, & Wasserman, W, 1996, *Applied Linear Statistical Models*, Fourth Edition, Irwin. Chapter 1 to 5.
- Ramanathan, R., 1998, *Introductory Econometrics With Applications*, The Dryden Press, Harcourt Brace College Publishers. Chapter 3.
- Thomas, R.L., 1997, *Modern Econometrics an Introduction*, Addison-Wesley. Chapters 4 and 6.

Exercises

1.
 - (i) What is regression analysis?
 - (ii) In regression analysis what is the estimating equation?
 - (iii) What is the purpose of correlation analysis?
 - (iv) Define direct and inverse relationships.
 - (v) To what does the term causal relationship refer?
 - (vi) Explain the difference between linear and curvilinear relationships.
 - (vii) Explain why and how we construct a scatter diagram.
 - (viii) What is multiple regression analysis?

2. Given the equation $Y = 1 + 2x$
 - (i) Complete the following table by determining the value of Y for each given value of X.
 X: 0 1 2 3 4 5 6 7
 Y:
 - (ii) Plot each pair of points (X,Y) on graph paper and draw the straight line representing the equation in (i).

3. Determine the slope and the Y intercept for each of the following equations:

(i) $Y = -2 + X$	(ii) $3Y = 12 + 1.5X$
(iii) $Y = \frac{-X}{2}$	(iv) $-Y = X - 2$

4. Use the results obtained in question 3 to graph the equations obtained in (i) through (iv).

5. On the graphs drawn for question 4, show by plotting new lines what changes in location occurs if the equations in 3(i) to (iv) are changed as follows:
 - (i) $Y = -2 - X$ instead of $Y = -2 + X$
 - (ii) $3Y = 12 + 3X$ instead of $3Y = 12 + 1.5X$
 - (iii) $Y = \frac{-X}{2} + 4$ instead of $Y = \frac{-X}{2}$
 - (vi) $-Y = X + 4$ instead of $-Y = X - 2$

6. For each of the following sets of data, determine the estimated regression equation: $Y = a + bX$.

(i) $\bar{X} = 10, \bar{Y} = 20, XY = 3000, X^2 = 2000, n = 10$

(ii) $\bar{X} = 10, \bar{Y} = 20, XY = 1000, X^2 = 2000, n = 10$

(iii) $\bar{X} = 50, \bar{Y} = 10, XY = 30000, X^2 = 135000, n = 50$

7. A government economist wishes to establish the relationship between annual family income X and savings Y . A sample of $n = 100$ families has been randomly chosen from various income levels between \$5,000 and \$20,000. A thorough investigation of these families has been made, and the following intermediate calculations have been obtained (X and Y are measured in thousands of dollars):

$$X = \$1239, Y = \$79, XY = 1613, X^2 = 17322, Y^2 = 393$$

(i) Determine the equation for the estimated regression line.

(ii) State the meaning of the slope b and Y intercept a .

(iii) Calculate the standard error of the estimate, $S_{y\hat{x}}$, and the standard error of the regression coefficient, S_b . Does a comparison between these two indicate that the regression line may be a useful tool for predicting family savings? Why or why not?

8. Given the following data set:

X: 52 57 56 36 47 49 60 56 52 42 56 52 57 57 36 42 60 49 42 47

Y: 62 50 57 38 55 43 59 57 55 47 55 56 57 48 36 50 63 47 44 49

(i) Compute the regression coefficients a and b and formulate the regression equation for estimating Y from X .

(ii) Draw a scatter diagram by plotting the 20 observations and draw the regression line obtained in (i) on the graph.

(iii) Compute the standard error of the estimate, $S_{y\hat{x}}$, and the standard error of the coefficient, S_b .

(iv) Test the hypothesis $H_0: b = 0$ against the alternative hypothesis $H_1: b \neq 0$, using a 5% level of significance. Use a normal probability plot to determine if the residuals from the model are normally distributed.

9. Show that the regression line, $\hat{Y} = a + bX$, always passes through the point, (\bar{X}, \bar{Y}) .

10. Prove that $\bar{Y} = \bar{\hat{Y}}$.

11. Prove that $(Y_i - c)^2$ is a minimum if $c = \bar{Y}$.

12. Explain why an estimating equation is valid over only the range of values used for its development.

13. Explain the difference between the coefficient of determination and the coefficient of correlation.

14. Why should we be cautious in using past data to predict future trends?
15. Why must we not attribute causality in a relationship even when there is strong correlation between the variables or events?
16. In the mid to late 1970s and late 1980s, the prices of single-family homes rose faster than the rate of inflation. As a result, many investors directed their funds to the housing market as a hedge against inflation. One way an investor can assess the value of a specific house is to compare it to the sale prices of similar houses that have recently been sold. Another popular approach involves the use of a regression analysis to model the relationship between price and the variables that influence price. Independent variables that could be utilised are total living area, number of rooms, number of baths, age of property, and so on. Of these factors total living area provides the most information for determining the worth of a house. The table following lists the final selling price and total living area for a sample of 30 residential properties in the same geographic area of Melbourne that were sold during the last 6 months of 1994. The regression printout using a straight-line model to relate price to area for these data is given also.

ANOVA

	df	SS	MS	F	Significance F
Regression	1	3.27E+10	3.27E+10	99.71	9.94E-11
Residual	28	9.19E+09	3.28E+08		
Total	29	4.19E+10			

	Coefficients	Std Error	t Stat	P-value
Intercept	19455.51	15569.49	1.249	0.221793542
AREA	1812.84	181.55	9.98	9.93898E-11

$R^2 = 0.7807$, $Adj.R^2 = 0.773$, Standard Error = 18114.68, n = 30

	AREA	PRICE		AREA	PRICE
1	50.00	79000	16	85.00	147600
2	56.12	93000	17	110.00	195000
3	65.00	103000	18	110.00	195000
4	61.00	107500	19	85.00	165000
5	57.58	132500	20	75.60	180000
6	53.00	120000	21	106.00	210000
7	91.00	190000	22	106.00	215000
8	85.00	187000	23	107.00	215000
9	85.00	189950	24	94.20	209000
10	85.00	182500	25	78.12	179000
11	112.00	214950	26	78.12	179000
12	85.00	189000	27	78.12	179000
13	86.00	182330	28	78.12	172000
14	85.00	184950	29	94.20	184000
15	112.00	210000	30	59.9	151000

- (i) Construct a scattergram for the data.
- (ii) Find the least squares line and plot it on your scattergram.
- (iii) Find r^2 and interpret its value in the context of the problem.
- (iv) Do the data provide evidence that living area contributes information for predicting the price of a home? Use $\alpha = .05$.
- (v) Find a 95% confidence interval for β_1 . Does your confidence interval support the conclusion you reached in part (iv)? Explain. (Note that the standard error of β_1 is given on the printout in the column headed **Standard Error** and in the row corresponding to AREA.)

- (vi) Find the observed significance level for the test in part (iv), and interpret its value.
 - (vii) Estimate the mean selling price for homes with a total living area of 76.8. Use a 95% confidence interval.
 - (viii) Do the errors from this model satisfy the least squares assumption of normality?
17. Using the returns data below from two different industries, divide your sample into the first half (January 1978-December 1982) and the second half (January 1983-December 1987) and choose the half with which you will work.
- (i) Using your computer regression software, the 60 observations you have chosen, and equation (9.22), estimate by ordinary least squares the parameters β_0 and β_1 for the firm in each of these two industries. Do the estimates of β_0 correspond well with your prior intuition or beliefs? Why or why not?
 - (ii) For one of these companies, make a time plot of the historical company risk premium, the company risk premium predicted by the regression model, and the associated residuals. Are there any episodes or dates that appear to correspond with unusually large residuals? If so, attempt to interpret them.
 - (iii) For each of the companies, test the null hypothesis that $\beta_1 = 0$ against the alternative hypothesis that $\beta_1 \neq 0$, using a significance level of 95%. Would rejection of this null hypothesis imply that the CAPM has been invalidated? Why or why not?
 - (iv) For each company, construct a 95% confidence interval for β_1 . Then test the null hypothesis that the company's risk is the same as the average risk over the entire market, that is, test that $\beta_1 = 1$ against the alternative hypothesis that $\beta_1 \neq 1$. Did you find any surprises?
 - (v) For each of the two companies, compute the proportion of total risk that is market risk, also called systematic and nondiversifiable. William F. Sharpe⁹ states that "Uncertainty about the overall market . . . accounts for only 30% of the uncertainty about the prospects for a typical stock." Does evidence from the two companies you have chosen correspond to Sharpe's typical stock? Why or why not? What is the proportion of total risk that is specific and diversifiable? Do these proportions surprise you? Why?
 - (vi) In your sample, do large estimates of β_1 correspond with higher R^2 values? Would you expect this always to be the case? Why or why not?

Returns data for three companies in different industries are provided in the following table.

Electronics	Tandy	Tandy
Airlines	Delta	delta
Computers	Digital Equipment Company	dec
Risk free rate	30-day Treasury bills	T-bill
Market portfolio	Market	market

⁹ Sharpe, W.F., (1985) Investments, Third Edition, Englewood Cliffs, N.J., Prentice Hall, page 167.

	Month.	Tandy Γ_i	DEC Γ_i	DELTA Γ_i	T-bill Γ_f	Market Γ_m
1	1978.1	-0.075	-0.100	-0.028	0.00487	-0.045
2	2	-0.004	-0.063	-0.033	0.00494	0.010
3	3	0.124	0.010	0.070	0.00526	0.050
4	4	0.055	0.165	0.150	0.00491	0.063
5	5	0.176	0.038	-0.031	0.00513	0.067
6	6	-0.014	-0.021	0.023	0.00527	0.007
7	7	0.194	0.107	0.185	0.00528	0.071
8	8	0.222	-0.017	-0.021	0.00607	0.079
9	9	-0.100	-0.037	-0.081	0.00645	0.002
10	10	-0.206	-0.077	-0.153	0.00685	-0.189
11	11	0.086	0.064	0.055	0.00719	0.084
12	12	0.085	0.117	-0.023	0.00690	0.015
13	1979.1	-0.046	-0.012	-0.054	0.00761	0.058
14	2	-0.135	-0.066	-0.060	0.00761	0.011
15	3	0.122	0.088	0.098	0.00769	0.123
16	4	-0.094	0.005	-0.056	0.00764	0.026
17	5	-0.148	-0.028	0.063	0.00772	0.014
18	6	0.096	0.059	-0.006	0.00715	0.075
19	7	0.006	0.009	0.075	0.00728	-0.013
20	8	0.250	0.140	0.021	0.00789	0.095
21	9	-0.005	-0.027	-0.026	0.00802	0.039
22	10	-0.037	-0.010	-0.147	0.00913	-0.097
23	11	0.170	0.095	0.063	0.00819	0.116
24	12	0.037	0.018	0.020	0.00747	0.086
25	1980.1	0.032	0.058	0.022	0.00883	0.124
26	2	0.143	0.034	-0.093	0.01073	0.112
27	3	-0.105	-0.182	-0.031	0.01181	-0.243
28	4	-0.038	0.047	-0.018	0.00753	0.080
29	5	0.256	0.016	0.144	0.00630	0.062
30	6	0.041	0.021	0.010	0.00503	0.086
31	7	0.446	0.183	0.283	0.00602	0.065
32	8	0.167	0.081	-0.056	0.00731	0.025
33	9	0.157	0.045	-0.053	0.00860	0.015
34	10	-0.015	-0.028	0.046	0.00895	0.006
35	11	0.212	0.056	0.220	0.01137	0.092
36	12	0.022	0.035	0.040	0.00977	-0.056
37	1981.1	-0.139	-0.089	0.112	0.01092	-0.014
38	2	0.082	0.006	0.031	0.01096	-0.009
39	3	0.299	0.075	0.024	0.01025	0.067
40	4	0.092	0.075	0.062	0.01084	-0.008
41	5	0.136	0.107	0.105	0.01255	0.064
42	6	-0.167	-0.112	-0.114	0.01128	-0.003
43	7	0.032	-0.014	-0.094	0.01154	-0.033
44	8	-0.063	-0.065	-0.072	0.01169	-0.031
45	9	-0.008	-0.019	-0.013	0.01054	-0.164
46	10	0.241	0.102	-0.072	0.01003	0.062
47	11	-0.037	-0.065	-0.032	0.00816	0.069
48	12	-0.046	-0.060	-0.062	0.00740	-0.039
49	1982.1	0.059	0.027	0.056	0.00949	-0.079
50	2	-0.101	-0.049	0.145	0.00946	-0.101
51	3	-0.051	-0.104	0.038	0.01067	-0.028
52	4	0.053	0.054	-0.025	0.00972	0.041
53	5	-0.163	-0.056	0.042	0.00908	0.003
54	6	0.023	-0.073	0.106	0.00914	-0.078
55	7	0.050	-0.055	-0.118	0.00714	-0.006
56	8	0.017	0.273	0.055	0.00503	0.122
57	9	-0.026	-0.061	-0.139	0.00563	0.008
58	10	0.455	0.133	0.171	0.00620	0.136
59	11	0.273	0.175	0.289	0.00614	0.049
60	12	-0.042	-0.052	0.093	0.00648	0.014

61	1983.1	0.091	0.225	0.040	0.00646	0.065
62	2	0.032	-0.010	0.027	0.00599	0.028
63	3	-0.004	0.034	-0.016	0.00686	0.043
64	4	0.084	-0.060	-0.043	0.00652	0.097
65	5	-0.010	-0.052	-0.045	0.00649	0.080
66	6	-0.168	0.075	0.012	0.00673	0.048
67	7	-0.123	-0.142	-0.259	0.00714	-0.017
68	8	-0.048	0.007	0.080	0.00668	-0.034
69	9	-0.083	-0.005	0.041	0.00702	0.000
70	10	-0.058	-0.364	0.039	0.00678	-0.082
71	11	0.082	0.065	0.120	0.00683	0.066
72	12	0.095	0.034	-0.028	0.00693	-0.012
73	1984.1	-0.190	0.208	-0.013	0.00712	-0.029
74	2	-0.100	-0.024	-0.117	0.00672	-0.030
75	3	-0.008	0.057	0.065	0.00763	0.003
76	4	0.120	0.053	-0.085	0.00741	-0.003
77	5	-0.231	-0.071	-0.070	0.00627	-0.058
78	6	-0.037	-0.043	-0.012	0.00748	0.005
79	7	0.029	-0.009	0.045	0.00771	-0.058
80	8	0.079	0.159	0.040	0.00852	0.146
81	9	-0.100	-0.025	0.008	0.00830	0.000
82	10	-0.096	0.093	0.161	0.00688	-0.035
83	11	0.027	0.006	-0.026	0.00602	-0.019
84	12	0.005	0.070	0.156	0.00612	-0.001
85	1985.1	0.170	0.084	-0.010	0.00606	0.097
86	2	0.119	-0.067	0.087	0.00586	0.012
87	3	0.094	-0.071	-0.003	0.00650	0.008
88	4	-0.133	-0.050	-0.123	0.00601	-0.010
89	5	0.091	0.057	0.179	0.00512	0.019
90	6	0.087	-0.101	0.021	0.00536	-0.003
91	7	-0.119	0.080	0.008	0.00562	0.012
92	8	0.063	0.032	-0.066	0.00545	0.005
93	9	-0.011	0.036	-0.112	0.00571	-0.055
94	10	0.098	0.040	-0.083	0.00577	0.026
95	11	0.021	0.073	0.020	0.00540	0.059
96	12	0.098	0.095	0.030	0.00479	0.013
97	1986.1	-0.040	0.162	0.122	0.00548	-0.009
98	2	0.096	0.093	-0.055	0.00523	0.049
99	3	-0.047	-0.063	0.076	0.00508	0.048
100	4	-0.058	0.119	0.059	0.00444	-0.009
101	5	0.094	0.037	-0.043	0.00469	0.049
102	6	-0.092	-0.063	-0.070	0.00478	0.004
103	7	-0.078	0.066	0.018	0.00458	-0.076
104	8	0.018	0.105	0.018	0.00343	0.049
105	9	-0.108	-0.110	0.026	0.00416	-0.047
106	10	0.242	0.103	0.134	0.00418	0.018
107	11	0.094	0.048	-0.018	0.00420	0.000
108	12	-0.023	0.008	-0.010	0.00382	-0.005
109	1987.1	0.130	0.385	0.161	0.00454	0.148
110	2	0.174	0.056	0.133	0.00437	0.065
111	3	-0.118	0.061	-0.129	0.00423	0.037
112	4	-0.119	0.055	-0.121	0.00207	-0.025
113	5	-0.026	-0.082	0.151	0.00438	0.004
114	6	0.045	0.041	0.014	0.00402	0.038
115	7	0.087	0.000	0.043	0.00455	0.055
116	8	0.027	0.157	-0.037	0.00460	0.015
117	9	0.088	0.001	-0.067	0.00520	-0.015
118	10	-0.246	-0.281	-0.260	0.00358	-0.260
119	11	-0.190	-0.127	-0.137	0.00288	-0.070
120	12	0.040	0.134	0.121	0.00277	0.073

