STATISTICS FOR ECONOMISTS: A BEGINNING

John E. Floyd
University of Toronto

July 2, 2010

PREFACE

The pages that follow contain the material presented in my introductory quantitative methods in economics class at the University of Toronto. They are designed to be used along with any reasonable statistics textbook. The most recent textbook for the course was James T. McClave, P. George Benson and Terry Sincich, *Statistics for Business and Economics*, Eighth Edition, Prentice Hall, 2001. The material draws upon earlier editions of that book as well as upon John Neter, William Wasserman and G. A. Whitmore, *Applied Statistics*, Fourth Edition, Allyn and Bacon, 1993, which was used previously and is now out of print. It is also consistent with Gerald Keller and Brian Warrack, *Statistics for Management and Economics*, Fifth Edition, Duxbury, 2000, which is the textbook used recently on the St. George Campus of the University of Toronto. The problems at the ends of the chapters are questions from mid-term and final exams at both the St. George and Mississauga campuses of the University of Toronto. They were set by Gordon Anderson, Lee Bailey, Greg Jump, Victor Yu and others including myself.

This manuscript should be useful for economics and business students enrolled in basic courses in statistics and, as well, for people who have studied statistics some time ago and need a review of what they are supposed to have learned. Indeed, one could learn statistics from scratch using this material alone, although those trying to do so may find the presentation somewhat compact, requiring slow and careful reading and thought as one goes along.

I would like to thank the above mentioned colleagues and, in addition, Adonis Yatchew, for helpful discussions over the years, and John Maheu for helping me clarify a number of points. I would especially like to thank Gordon Anderson, who I have bothered so frequently with questions that he deserves the status of mentor.

After the original version of this manuscript was completed, I received some detailed comments on Chapter 8 from Peter Westfall of Texas Tech University, enabling me to correct a number of errors. Such comments are much appreciated.

J. E. Floyd
July 2, 2010

Chapter 8

Simple Linear Regression

We now turn to the area of statistics that is most relevant to what economists usually do—the analysis of relationships between variables. Here we will concentrate entirely on linear relationships. For example, we might be interested in the relationship between the quantity of money demanded and the volume of transactions that people make as represented by the level of money income. Or we might be interested in the relationship between family expenditures on food and family income and family size. *Regression analysis* is used to analyse and predict the relationship between the *response* or *dependent* variable (money holdings and family expenditure on food in the above examples) and one or more *independent*, *explanatory*, or *predictor* variables. In the demand for money example the single independent variable was the level of income; in the family food expenditure example, there were two independent variables, family income and family size.

We must distinguish two types of relationships between variables. A *deterministic* relationship exists if the value of Y is uniquely determined when the value of X is specified—the relationship between the two variables is exact. For example, we might have

$$Y = \beta X$$

where β is some constant such as 10. On the other hand, there may be a relationship between two variables that involves some random component or random error. This relationship is called a *probabilistic* or *statistical* relationship. In this case we might have

$$Y = \beta X + \epsilon$$

which can be viewed as a *probabilistic model* containing two components—a *deterministic component* βX plus a *random error* ϵ. Figure 8.1 presents

an example of a deterministic straight-line relationship between X and Y along which all observed combinations of the two variables lie. An example of a probabilistic relationship is given in Figure 8.2. There is a *scatter* of observed combinations of X and Y around a straight-line functional or deterministic relationship indicating errors in the fit that result from the influence on Y of unknown factors in addition to X. For each level of X there is a probability distribution of Y. And the means of these probability distributions of Y vary in a systematic way with the level of X.

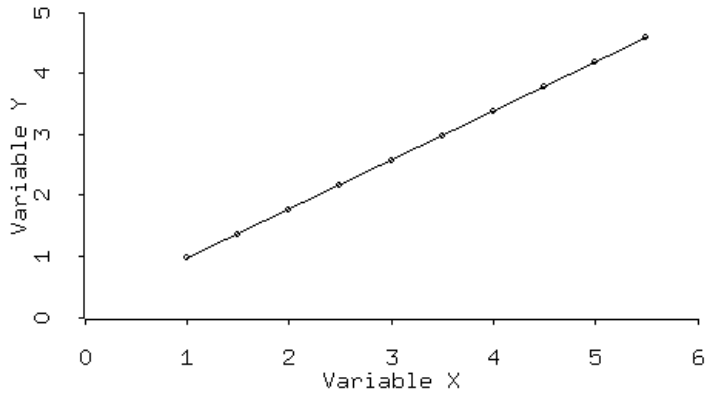

Figure 8.1: A functional or deterministic relationship between two variables X and Y.

8.1 The Simple Linear Regression Model

When the statistical relationship is linear the regression model for the observation Y_i takes the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{8.1}$$

where the functional or deterministic relationship between the variables is given by $\beta_0 + \beta_1 X_i$ and ϵ_i is the random scatter component. Y_i is the dependent variable for the ith observation, X_i is the independent variable for the ith observation, assumed to be non-random, β_0 and β_1 are parameters and the ϵ_i are the deviations of the Y_i from their predicted levels based on X_i, β_0 and β_1.

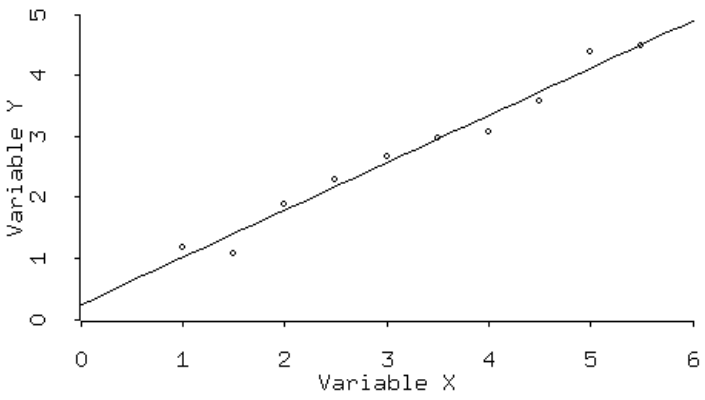

Figure 8.2: An probabilistic or statistical relationship between two variables X and Y.

The error term is assumed to have the following properties:

a) The ϵ_i are normally distributed.

b) The expected value of the error term, denoted by $E\{\epsilon_i\}$, equals zero.

c) The variance of the ϵ_i is a constant, σ^2.

d) The ϵ_i are statistically independent—that is, the covariance between ϵ_i and ϵ_j is zero.

In other words,

$$\epsilon_i = N(0, \sigma^2).$$

This normality assumption for the ϵ_i is quite appropriate in many cases. There are often many factors influencing Y other than the independent variable (or, as we shall see later, variables) in the regression model. Insofar as the effects of these variables are additive and tend to vary with a degree of mutual independence, their mean (and their sum) will tend to normality according to the central limit theorem when the number of these 'missing' factors is large. The distribution of the error term and the resulting levels of Y at various levels of X is given in Figure 8.3.

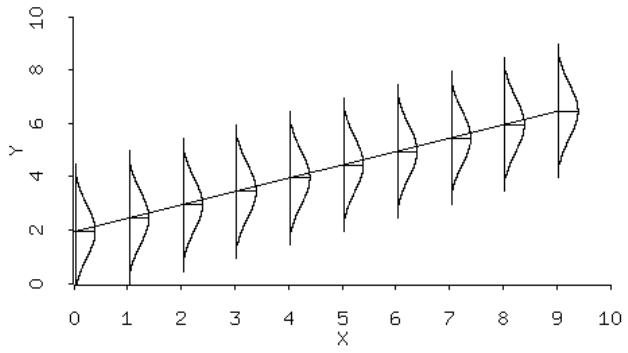

Figure 8.3: Simple regression of Y on X: The probability distribution of Y given X.

Since the error term ϵ_i is a random variable, so is the dependent variable Y_i. The expected value of Y_i equals

$$\begin{aligned} E\{Y_i\} &= E\{\beta_0 + \beta_1 X_i + \epsilon_i\} \\ &= E\{\beta_0\} + E\{\beta_1 X_i\} + E\{\epsilon_i\} \\ &= \beta_0 + \beta_1 E\{X_i\} + 0 \\ &= \beta_0 + \beta_1 X_i \end{aligned} \tag{8.2}$$

where $E\{X_i\} = X_i$ because these X_i are a series of pre-determined non-random numbers. Equation (8.2), the underlying deterministic relationship is called the *regression function*. It the *line of means* that relates the mean of Y to the value of the independent variable X. The parameter β_1 is the slope of this line and β_0 is its intercept.

The variance of Y_i given X_i equals

$$\begin{aligned} Var\{Y_i|X_i\} &= Var\{\beta_0 + \beta_1 X_i + \epsilon_i\} \\ &= Var\{\beta_0 + \beta_1 X_i\} + Var\{\epsilon_i\} \\ &= 0 + Var\{\epsilon_i\} \;=\; \sigma^2 \end{aligned} \tag{8.3}$$

where the regression function $\beta_0 + \beta_1 X_i$ is deterministic and therefore does not vary. Thus the Y_i have the same variability around their means at all X_i.

Finally, since the ϵ_i are assumed to be independent for the various observations, so are the Y_i conditional upon the X_i. Hence it follows that

$$Y_i = N(\beta_0 + \beta_1 X_i, \sigma^2).$$

8.2 Point Estimation of the Regression Parameters

Point estimates of β_0 and β_1 can be obtained using a number of alternative estimators. The most common estimation method is the *method of least squares*. This method involves choosing the estimated regression line so that the sum of the squared deviations of Y_i from the value predicted by the line is minimized. Let us denote the deviations of Y_i from the fitted regression line by e_i and our least-squares estimates of β_0 and β_1 by b_0 and b_1 respectively. Then we have

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 \tag{8.4}$$

where Q is the sum of squared deviations of the Y_i from the values predicted by the line.

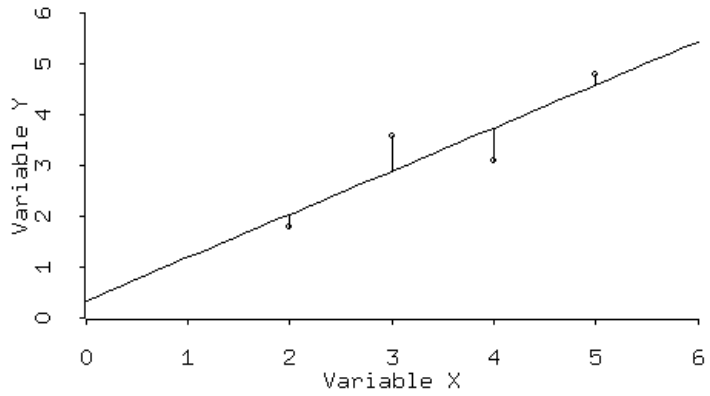

Figure 8.4: A least-squares fit minimizes the sum of the squared vertical distances of the data-points from the least-squares line.

The least-squares estimation procedure involves choosing b_0 and b_1, the intercept and slope of the line, so as to minimize Q. This minimizes the sum

of the squared lengths of the vertical lines in Figure 8.4. Expanding equaton (8.4), we have

$$\begin{aligned} Q &= \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2 \\ &= \sum_{i=1}^{n} Y_i^2 + n\, b_0^2 + \sum_{i=1}^{n} b_1^2 X_i^2 - 2\, b_0 \sum_{i=1}^{n} Y_i \\ &\qquad -2\, b_1 \sum_{i=1}^{n} Y_i X_i + 2\, b_0 b_1 \sum_{i=1}^{n} X_i \end{aligned} \tag{8.5}$$

To find the least squares minimizing values of b_0 and b_1 we differentiate Q with respect to each of these parameters and set the resulting derivatives equal to zero. This yields

$$\frac{\partial Q}{\partial b_0} = 2\, n\, b_0 - 2 \sum_{i=1}^{n} Y_i + 2\, b_1 \sum_{i=1}^{n} X_i = 0 \tag{8.6}$$

$$\frac{\partial Q}{\partial b_1} = 2\, b_1 \sum_{i=1}^{n} X_i^2 - 2 \sum_{i=1}^{n} X_i Y_i + 2\, b_0 \sum_{i=1}^{n} X_i = 0 \tag{8.7}$$

which simplify to

$$\sum_{i=1}^{n} Y_i = n\, b_0 + b_1 \sum_{i=1}^{n} X_i \tag{8.8}$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \tag{8.9}$$

These two equations can now be solved simultaneously for b_0 and b_1. Dividing (8.8) by n, rearranging to put b_0 on the left side and noting that $\sum X_i = n\bar{X}$ and $\sum Y_i = n\bar{Y}$ we obtain

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{8.10}$$

Substituting this into (8.9), we obtain

$$\sum_{i=1}^{n} X_i Y_i = \bar{Y} \sum_{i=1}^{n} X_i - b_1 \bar{X} \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2, \tag{8.11}$$

which can be rearranged to yield

$$\begin{aligned} \sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i &= b_1 \left[\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i \right] \\ \sum_{i=1}^{n} X_i Y_i - n\, \bar{Y} \bar{X} &= b_1 \left[\sum_{i=1}^{n} X_i^2 - n\, \bar{X}^2 \right] \end{aligned} \tag{8.12}$$

By expansion it can be shown that

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\,\bar{Y}\,\bar{X}$$

and

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\,\bar{X}^2$$

so that by substitution into (8.12) we obtain

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{8.13}$$

This expression can be alternatively written as

$$b_1 = \frac{\sum xy}{x^2} \tag{8.14}$$

where $x = (X_i - \bar{X})$ and $y = (Y_i - \bar{Y})$ are the deviations of the variables from their respective means and the summation is over $i = 1 \ldots n$.

The least-squares estimators b_0 and b_1 are unbiased and, as can be seen from (8.10) and (8.13), linearly dependent on the n sample values Y_i. It can be shown that least-squares estimators are more efficient—that is, have lower variance—than all other possible unbiased estimators of β_0 and β_1 that are linearly dependent on the Y_i. It can also be shown that these desirable properties do not depend upon the assumption that the ϵ_i are normally distributed.

Estimators of β_0 and β_1 can also be developed using the method of maximum likelihood (under the assumption that the ϵ_i are normally distributed). These estimators turn out to be identical with the least-squares estimators.

Calculation of the regression line is straight forward using (8.10) and (8.14). The procedure is to

a) calculate the deviations of X_i and Y_i from their respective means.

b) square the deviations of the X_i and sum them.

c) multiply the X_i deviations with the corresponding Y_i deviations and sum them.

d) plug these sums of squares and cross products into (8.14) to obtain b_1, and

e) plug this value of b_1 into (8.10) along with the means of the X_i and Y_i to obtain b_0.

The regression function $E\{Y\} = \beta_0 + \beta_1 X$ is estimated as

$$\hat{Y} = b_0 + b_1 X \tag{8.15}$$

where $\hat{Y}$ is referred to as the *predicted* value of Y. The mean response or predicted value of Y when X takes some value X_h is

$$\hat{Y}_h = b_0 + b_1 X_h.$$

The point estimate of $E\{Y_h\}$ is thus $\hat{Y}_h$, the value of the estimated regression function when $X = X_h$.

8.3 The Properties of the Residuals

To make inferences (i.e., construct confidence intervals and do statistical tests) in regression analysis we need to estimate the magnitude of the random variation in Y. We measure the scatter of the observations around the regression line by comparing the observed values Y_i with the predicted values associated with the corresponding X_i. The difference between the observed and predicted values for the ith observation is the *residual* for that observation. The residual for the ith observation is thus

$$e_i = Y_i - b_0 - b_1 X_i.$$

Note that e_i is the *estimated residual* while ϵ_i is the *true* residual or *error term* which measures the deviations of Y_i from its true mean $E\{Y\}$.

The least-squares residuals have the following properties.

a) They sum to zero — $\sum e_i = 0$.

b) The sum of the squared residuals $\sum e_i^2$ is a minimum—this follows because the method of least squares minimizes Q.

c) The sum of the weighted residuals is zero when each residual is weighted by the corresponding level of the independent variable — $\sum X_i e_i = 0$.

d) The sum of the weighted residuals is zero when each residual is weighted by the corresponding fitted value — $\sum \hat{Y}_i e_i = 0$.

8.4 The Variance of the Error Term

To conduct statistical inferences about the parameters of the regression we are going to need an estimate of the variance of the error term. An obvious way to proceed is to work with the sum of squared deviations of the observed levels of Y from the predicted levels—i.e.,

$$\sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \sum_{i=1}^{n} e_i^2.$$

It turns out that the mean or average of these squared deviations is the appropriate estimator of σ^2, provided we recognize that all n of of these deviations are not independent. Since we used the sample data to estimate two parameters, b_0 and b_1, we used up two of the n pieces of information contained in the sample. Hence, there are only $n-2$ independent squared deviations—i.e., $n-2$ degrees of freedom. Hence, in taking the average we divide by $n-2$ instead of n. An unbiased estimator of σ^2 is

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{8.16}$$

where *MSE* stands for *mean square error*. In general, a mean square is a sum of squares divided by the degrees of freedom with which it is calculated.

8.5 The Coefficient of Determination

Consider the sum of the squared deviations of the Y_i from their mean $\bar{Y}$, otherwise known as the total sum of squares and denoted by *SSTO*,

$$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

This total sum of squares can be broken down into components by adding and subtracting $\hat{Y}$ as follows:

$$\begin{aligned} SSTO &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \right]^2 \\
&= \sum_{i=1}^{n} \left[(Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \right] \\
&= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}). \quad (8.17)
\end{aligned}
$$

The term

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

equals zero, since

$$
\begin{aligned}
\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^{n} \left[(Y_i - \hat{Y}_i)\hat{Y}_i - (Y_i - \hat{Y}_i)\bar{Y} \right] \\
&= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)\hat{Y}_i - \sum_{i=1}^{n} (Y_i - \hat{Y}_i)\bar{Y} \\
&= \sum_{i=1}^{n} e_i \hat{Y}_i \; - \; \bar{Y} \sum_{i=1}^{n} e_i. \quad (8.18)
\end{aligned}
$$

From the properties a) and d) of the least-squares residuals listed on page 200 above, $\hat{Y} \sum e_i$ and $\sum e_i Y_i$ are both zero. We can thus partition the total sum of squares into the two components,

$$
\begin{aligned}
SSTO &= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\
&= \quad SSR \quad + \quad SSE. \quad (8.19)
\end{aligned}
$$

The term

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

is the sum of squares of the deviations of the observed values Y_i from the values predicted by the regression. It is the portion of the total variability of Y that remains as a residual or error after the influence of X is considered, and is referred to as the *sum of squared errors*. The term

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

is the sum of the squared deviations of the predicted values Y_i from the mean of Y. It is the portion of the total variability of Y that is explained by the regression—that is, by variations in the independent variable X. It follows that the sum of squared errors is the total sum of squares minus the portion explained by X—i.e., $SSE = SSTO - SSR$.

The *coefficient of determination*, usually referred to as the R^2, is the fraction of the total variability of the Y_i that is explained by the variability of the X_i. That is,

$$R^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO}. \tag{8.20}$$

8.6 The Correlation Coefficient Between X and Y

The correlation coefficient between two random variables, X and Y has previously been defined as

$$\rho = \frac{Cov\{XY\}}{\sqrt{Var\{X\}Var\{Y\}}}. \tag{8.21}$$

An appropriate estimator of ρ is

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}. \tag{8.22}$$

As in the case of the true correlation coefficient ρ, r can vary between minus unity and plus unity. In the present situation, however, the X_i are assumed fixed—i.e., do not vary from sample to sample—so that X is not a random variable. Nevertheless, r is still a suitable measure of the degree of association between the variable Y and the fixed levels of X. Moreover, when we square r we obtain

$$r^2 = \frac{(\sum(X_i - \bar{X})(Y_i - \bar{Y}))^2}{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2} \tag{8.23}$$

which, it turns out, can be shown to equal R^2 as defined above.

8.7 Confidence Interval for the Predicted Value of Y

Suppose we want to estimate the mean level of Y for a given level of X and establish confidence intervals for that mean level of Y. For example, an admissions officer of a U.S. college might wish to estimate the mean grade point average (GPA) of freshmen students who score 550 on the Scholastic Aptitude Test (SAT).

We have already established that the predicted value Y_h is a good point estimator of $E\{Y_h\}$. In order to obtain confidence intervals for $E\{Y_h\}$, however, we need a measure of the variance of $\hat{Y}_h$. It turns out that

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{8.24}$$

for which an appropriate estimator is

$$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{8.25}$$

where *MSE* is the mean square error, previously defined as

$$MSE = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2}.$$

The magnitude of the estimated variance $s^2\{\hat{Y}_h\}$ is affected by a number of factors:

a) It is larger the greater the variability of the residuals e_i.

b) It is larger the further the specified level of X is from the mean of X in either direction—i.e., the bigger is $(X_h - \bar{X})^2$.

c) It is smaller the greater the variability of the X_i about the mean of X.

d) It is smaller the greater the sample size n. There are two reasons for this. The greater is n, the smaller are both $1/n$ and *MSE* and, in addition, when n is larger the sum of the squared deviations of the X_i from their mean will tend to be larger.

The above points can be seen with reference to Figure 8.5. The true functional relationship is given by the thick solid line and has slope β_1 and intercept β_0. Alternative fitted regression lines are given by the upward

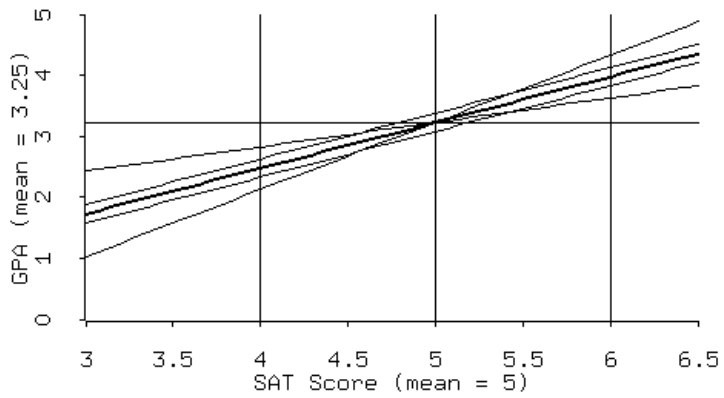

Figure 8.5: The true linear functional relationship (thick line) between Scholastic Aptitude Test (SAT) score and subsequent Grade Point Average (GPA) measured on a 5 point scale in freshman college courses, together with some possible fitted regression lines based on differing samples.

sloping thin lines. Each regression line always passes through the point $(\bar{X}\bar{Y})$ for the sample in question. Different samples of Y_i's drawn for the same set of X_i's yield different regression lines having different slopes since b_1 is a random variable. Also, different samples will yield different mean values of Y, though $\bar{X}$ will be the same because the X_i are fixed from sample to sample. This means that the level of the regression line is also a random variable as shown by the thin lines parallel to the true functional relationship—its variance at $\bar{X}$ is the variance of the error term σ^2 which is estimated by *MSE*.

The estimated variance of the predicted values of Y at $\bar{X}$, associated in the above example with a SAT score of 500, will be equal to *MSE* divided by n and will be determined entirely by the variance of the level of the line. At levels of X above the mean, say for a SAT score of 600, the variance of the predicted value of Y will be larger because there is both variance in the level of the regression line and variance of the slope of the line pivoting on $(\bar{X}\bar{Y})$. The further away one gets from the mean value of X, the bigger is the effect on the variance of the predicted Y of the variation of b_1 from sample to sample. Also, notice that the variance of the predicted Y at a SAT score of 400 will be the same as the variance of the predicted Y at a SAT

score of 600 because the effect of the sampling variation of b_1 is the same at both points (which are equidistant from $\bar{X}$) and the effect of sampling variation on the level of the regression line is the same at all X_i since it depends on $\bar{X}$ which is constant from sample to sample. We can now form the standardised statistic

$$\frac{\hat{Y}_h - E\{\hat{Y}_h\}}{s\{\hat{Y}_h\}}$$

which is distributed according to the t-distribution with $n-2$ degrees of freedom. There are two less degrees of freedom than the number of observations because we used the sample to estimate two parameters, β_0 and β_1. The confidence limits for $E\{Y_h\}$ with confidence coefficient α are thus

$$\hat{Y}_h \;\pm\; t(1-\alpha/2; n-2)\, s\{\hat{Y}_h\}.$$

This confidence interval is interpreted for repeated samples when the X_i are the same from sample to sample. Of many confidence intervals so established based on repeated samples, 100α percent will bracket $E\{Y_h\}$.

8.8 Predictions About the Level of Y

Suppose that we want to predict the grade point average of a student with a SAT score X_h equal to 600. It is important to distinguish this prediction, and the confidence interval associated with it, from predictions about the mean level of Y_h, the point estimator of which was $\hat{Y}_h$. That is, we want to predict the *level* of Y associated with a *new* observation at some X_h, not the *mean value* of Y associated with a whole sample drawn at a value of X equal to X_h. Predicting the grade point average of a randomly selected student who scored 600 on the SAT is very different from predicting what the mean grade point average of students who score 600 on the SAT will be.

If we knew the true values of the regression parameters, β_0, β_1 and σ, the procedure would be quite simple. We could simply calculate

$$E\{Y_h\} = \beta_0 + \beta_1 X_h$$

which might equal, say, 3.7. This would be the point estimate of $Y_{h(new)}$, the newly selected student's grade point average. We could then use the known value of σ to establish a confidence interval for an appropriate value of α.

But we don't know the true regression parameters and so must estimate them. The statistic $\hat{Y}_h$ is an appropriate point estimator of $Y_{h(new)}$. To get a

confidence interval we must estimate the variance of $Y_{h(new)}$. This variance is based on the variance of the difference between Y_h and $\hat{Y}_h$ together with the assumption that the new observation is selected independently of the original sample observation. This yields

$$\begin{aligned} \sigma^2\{\hat{Y}_{h(new)}\} &= \sigma^2\{Y_h - \hat{Y}_h\} \\ &= \sigma^2\{Y_h\} + \sigma^2\{\hat{Y}_h\} \\ &= \sigma^2 + \sigma^2\{\hat{Y}_h\} \end{aligned} \tag{8.26}$$

which is composed of two parts. It is the sum of

a) the variance of the mean predicted level of Y associated with the particular level of X.

b) the variance of the actual level of Y around its predicted mean level, denoted by σ^2.

In the situation above where we knew the true parameters of the regression model we could calculate $\hat{Y}_h$ exactly so that its variance was zero and the grade point average of the new student then varied only because of σ^2.

The variance of $\hat{Y}_{h(new)}$ can be estimated by

$$\begin{aligned} s^2\{\hat{Y}_{h(new)}\} &= MSE + \sigma^2\{\hat{Y}_h\} \\ &= MSE + MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \\ &= MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]. \end{aligned} \tag{8.27}$$

The calculation of the confidence interval is now a routine matter, using the fact that

$$\frac{\hat{Y}_{h(new)} - \hat{Y}_h}{s^2\{\hat{Y}_{h(new)}\}}$$

is distributed according the t-distribution with degrees of freedom equal to $n - 2$. The resulting prediction interval is, of course, much wider than the confidence interval for $E\{\hat{Y}_h\}$ because the variance of $Y_{h(new)}$ contains an additional component consisting of the variance of Y_h around $E\{\hat{Y}_h\}$.

8.9 Inferences Concerning the Slope and Intercept Parameters

In most regression analysis in economics the primary objective is to estimate β_1. The regression slope b_1 is an efficient and unbiased estimate of that parameter. To obtain confidence intervals for β_1, however, and test hypotheses about it, we need the variance of the sampling distribution of b_1. This variance, it turns out, equals

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \tag{8.28}$$

which can be estimated by the statistic

$$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2}. \tag{8.29}$$

The confidence interval for β_1 can be obtained from the fact that

$$\frac{b_1 - \beta_1}{s\{b_1\}}$$

is distributed according to the t-distribution with $n-2$ degrees of freedom. As explained in Chapter 4, the t-distribution is symmetrical and flatter than the standard-normal distribution, becoming equivalent to that distribution as the degrees of freedom become large. The confidence intervals for β_1 with confidence coefficient α are then

$$b_1 \pm t(1-\alpha/2, n-2)\, s\{b_1\}.$$

where $t(1-\alpha/2, n-2)$ is the t-statistic associated with a cumulative probability of $(1-\alpha)$ when the degrees of freedom are $(n-2)$.

Now suppose that we want to test whether there is any relationship between Y and X. If there is no relationship, β_1 will be zero. Accordingly, we set the null hypothesis as

$$H_0\colon\ \beta_1 = 0$$

and the alternative hypothesis as

$$H_1\colon\ \beta_1 \neq 0.$$

Using our sample data we calculate the standardised test statistic

$$t^* = \frac{b_1}{s\{b_1\}},$$

which is distributed according to the t-distribution with $n-2$ degrees of freedom, and compare it with the critical values of t for the appropriate degree of α-risk from the table of t-values in the back of our statistics textbook. When the standardised test statistic is in the critical range—i.e., in the range for rejecting the null hypothesis—we say that β_1 is *significantly different from zero* at the $100\,\alpha$ percent level. Also we can calculate the P-value of the test statistic t, which equals the probability that a value of b_1 as different from zero as the one observed could have occurred on the basis of pure chance.

Frequently we want to test whether or not β_1 exceeds or falls short of some particular value, say β_1^o. This can be done by setting the null and alternative hypotheses as, for example,

$$H_0 : \beta_1 \leq \beta_1^o$$

and

$$H_1 : \beta_1 > \beta_1^o,$$

expressing the standardised test statistic as

$$t^* = \frac{b_1 - \beta_1^o}{s\{b_1\}},$$

and applying the critical values from the t-table for the appropriate level of α. When the standardised test statistic is in the critical range we can say that β_1 is significantly greater than β_1^o at the $100\,\alpha$ percent level.

Occasionally, inferences concerning the intercept parameter β_0 are also of interest. The regression intercept coefficient b_0 is an unbiased and efficient estimator of β_0. To obtain confidence intervals and conduct hypotheses tests we need an estimator of the sampling variance $\sigma^2\{b_0\}$. It turns out that $b_0 = \hat{Y}_h$ where $X_h = 0$ so we can use the estimator

$$\begin{aligned} s^2\{\hat{Y}_h\} &= MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \\ = \; s^2\{b_0\} &= MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]. \end{aligned} \tag{8.30}$$

Statistical tests can now be undertaken and confidence intervals calculated using the statistic

$$\frac{b_0 - \beta_0}{s\{b_0\}}$$

which is distributed as t(n - 2).

It turns out that these tests are quite robust—that is, the actual α-risk and confidence coefficient remain close to their specified values even when the error terms in the regression model are not exactly normally distributed as long as the departure from normality is not too great.

8.10 Evaluation of the Aptness of the Model

It must now be reemphasized that the application of this regression model to practical problems involves some very critical assumptions—namely, that the true residuals are independently normally distributed with zero mean and constant variance. We can never be sure in advance that in any particular application these assumptions will be close enough to the truth to make our application of the model valid. A basic approach to investigating the aptness or applicability of the model to a particular situation is to *analyse the residuals* from the regression—$e_i = Y_i - \hat{Y}_i$.

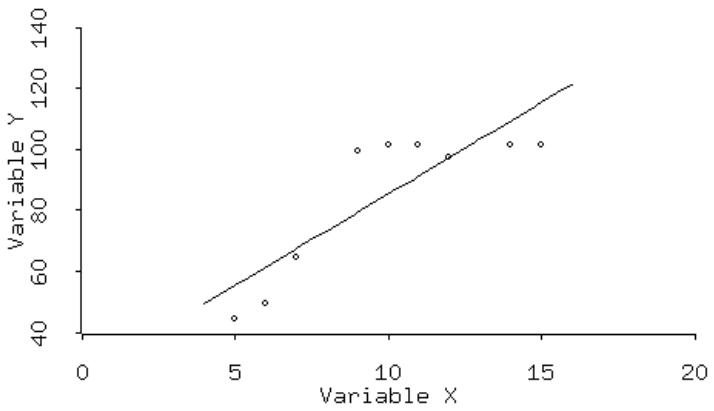

Figure 8.6: The actual and fitted values for a particular regression.

A number of important departures from the regression model may occur. First, the regression function we are trying to estimate may not be linear. We can get a good sense of whether or not this may be a problem by plotting the actual and predicted values of Y against the independent variable X, as is done in Figure 8.6, or plotting the residuals against the predicted values of Y as is done for the same regression in Figure 8.7. When the true relationship between the variables is linear the residuals will scatter

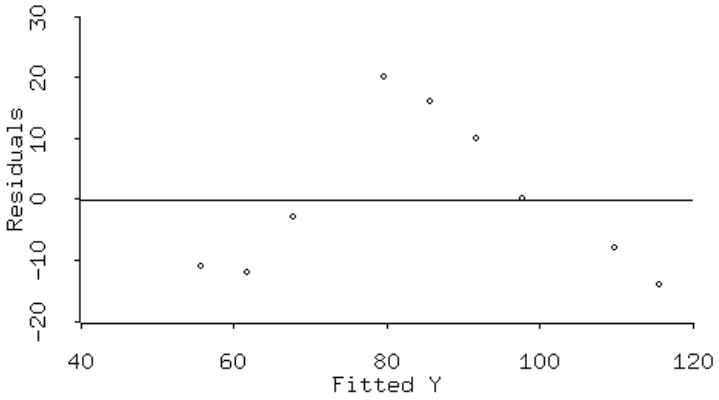

Figure 8.7: The residuals of the regression in Figure 8.6 plotted against the fitted values.

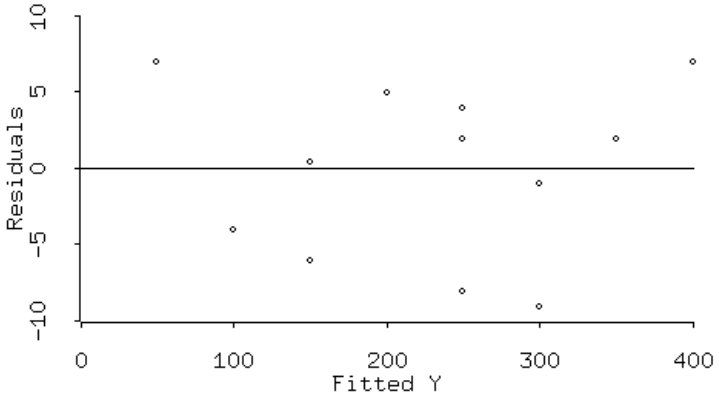

Figure 8.8: Well-behaved regression residuals plotted against the fitted values.

at random around the fitted straight line or around the zero line when plotted against the predicted values of the dependent variable. Obviously, the underlying functional relationship in Figure 8.6 is non-linear. An example of well-behaved residuals is given in Figure 8.8.

A second problem is that the variance of the e_i may not be constant

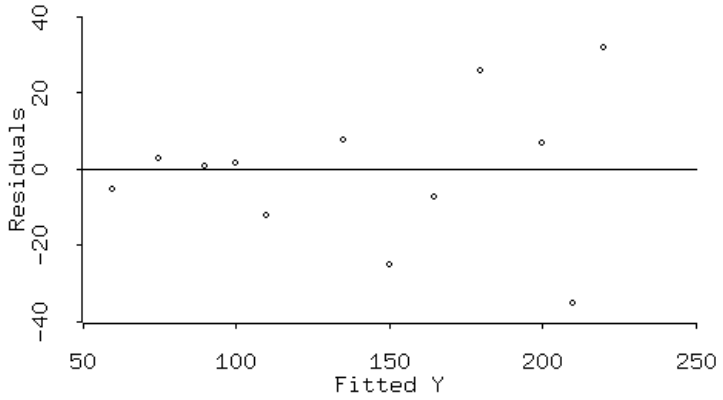

Figure 8.9: An example of heteroscedasticity—regression residuals plotted against the fitted values.

with respect to $\hat{Y}$ but may vary systematically with it. This problem is called *heteroscedasticity*. This is illustrated in Figure 8.9 where the residuals obviously increase as the predicted value of Y becomes larger.

Third, there may be lack of normality in the error terms. One way of checking the error term for normality is to standardise it by dividing it by its standard deviation—the square root of *MSE*—and checking to see whether approximately 2/3 of the errors lie within one standard deviation of zero. Alternatively, we could apply the chi-square test for normality developed in the previous chapter. Less formally, we can compare the observed frequencies of the standardised errors with the theoretically expected frequencies.

Finally, the errors may not be independent of each other. This happens frequently in time-series analysis where there is *autocorrelation* or *serial correlation* in the residuals—when the residual associated with one value of X or its predicted value of Y is high, the residual associated with the adjacent values of X or Y will also be high. This problem is discussed in detail in the next chapter.

To get around these problems it is sometimes useful to transform the variables. The residuals from estimating $Y = \beta_0 + \beta_1 X$ may be heteroscedastic, but those from estimating $log(Y) = \beta_0 + \beta_1 X$ may not be. Similarly, the relationship between $log(X)$ and $log(Y)$, or $1/X$ and $1/Y$, may be linear even though the relationship between X and Y may not be. Sometimes the residuals from the regression may not be well-behaved because, in truth, Y

depends on two variables X and Z instead of just X. By leaving Z out of the model, we are attempting to force the single variable X to explain more than it is capable of, resulting in deviations of the predicted from the actual levels of Y that reflect the influence of the absent variable Z.

8.11 Randomness of the Independent Variable

In some regression analyses it is more reasonable to treat both X and Y as random variables instead of taking the X_i as fixed from sample to sample. When X is random, the distribution of Y at a given level of X is a conditional distribution with a conditional mean and a conditional variance (i.e., conditional upon the level of X). In this case all of the results presented above for the regression model with X fixed continue to apply as long as

a) the conditional distribution of Y is normal with conditional mean $\beta_0 + \beta_1 X$ and conditional variance σ^2, and

b) the X_i are independent random variables whose probability distribution does not depend on the parameters β_0, β_1 and σ^2.

The interpretations of confidence intervals and risks of errors now refer to repeated sampling where *both* the X and Y variables change from one sample to the next. For example the confidence coefficient would now refer to the proportion of times that the interval brackets the true parameter when a large number of repeated samples of n pairs (X_i, Y_i) are taken and the confidence interval is calculated for each sample. Also, when both X and Y are random variables the correlation coefficient r is an estimator of the population correlation coefficient ρ rather than only a descriptive measure of the degree of linear relation between X and Y. And a test for $\beta_1 = 0$ is now equivalent to a test of whether or not X and Y are uncorrelated random variables.

8.12 An Example

During the first part of this century classical economics held that the real quantity of money demanded tends to be a constant fraction of real income—that is

$$\frac{M}{P} = k\, R_Y \tag{8.31}$$

where M is the nominal quantity of money held by the public, P is the general price level, R_Y is real national income and k is a constant, sometimes called the *Cambridge-k*. We want to use some data on nominal money holdings, nominal income and the consumer price index for Canada to test this idea. The data are presented in the worksheet below.

WORKSHEET FOR REGRESSION ANALYSIS OF CANADIAN DEMAND FOR MONEY

DATE	MON	GDP	CPI	RMON	RGDP	D-RMON	D-RGDP	Col. (6) Sq.	Col. (7) Sq.	(6) X (7)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1957	5.07	34.47	88.65	5.72	38.88	-7.90	-57.34	62.45	3288.30	453.17
1958	5.55	35.69	90.88	6.11	39.27	-7.51	-56.95	56.47	3243.81	428.00
1959	5.66	37.88	91.82	6.16	41.25	-7.46	-54.97	55.65	3021.77	410.09
1960	5.75	39.45	92.99	6.18	42.42	-7.44	-53.80	55.33	2894.42	400.18
1961	6.31	40.89	93.93	6.71	43.53	-6.91	-52.70	47.74	2776.85	364.09
1962	6.67	44.41	94.99	7.02	46.75	-6.60	-49.47	43.62	2447.38	326.75
1963	7.17	47.68	96.51	7.42	49.40	-6.20	-46.82	38.41	2192.23	290.18
1964	7.72	52.19	98.27	7.85	53.11	-5.77	-43.11	33.27	1858.56	248.68
1965	8.98	57.53	100.73	8.92	57.11	-4.70	-39.12	22.13	1530.04	184.02
1966	9.71	64.39	104.49	9.29	61.62	-4.33	-34.60	18.78	1197.08	149.93
1967	12.33	69.06	108.24	11.39	63.81	-2.23	-32.42	4.96	1050.80	72.22
1968	15.78	75.42	112.81	13.98	66.85	0.36	-29.37	0.13	862.57	-10.63
1969	15.40	83.03	117.74	13.08	70.52	-0.54	-25.70	0.29	660.62	13.87
1970	14.92	89.12	121.72	12.26	73.21	-1.36	-23.01	1.86	529.53	31.40
1971	16.52	97.29	125.12	13.20	77.75	-0.42	-18.47	0.18	341.06	7.81
1972	18.54	108.63	131.11	14.14	82.86	0.52	-13.37	0.27	178.68	-6.90
1973	20.61	127.37	141.07	14.61	90.29	0.99	-5.94	0.98	35.23	-5.87
1974	21.62	152.11	156.44	13.82	97.24	0.20	1.01	0.04	1.03	0.20
1975	24.06	171.54	173.44	13.87	98.91	0.25	2.68	0.06	7.20	0.67
1976	25.37	197.93	186.34	13.62	106.22	-0.01	10.00	0.00	99.92	-0.07
1977	27.44	217.88	201.35	13.63	108.21	0.00	11.99	0.00	143.71	0.05
1978	29.69	241.61	219.17	13.55	110.23	-0.08	14.01	0.01	196.35	-1.07
1979	30.97	276.10	239.23	12.94	115.41	-0.68	19.19	0.46	368.24	-12.99
1980	32.25	309.89	263.73	12.23	117.50	-1.40	21.28	1.95	452.77	-29.69
1981	33.64	356.00	296.57	11.34	120.04	-2.28	23.82	5.20	567.16	-54.31
1982	36.64	374.44	328.58	11.15	113.96	-2.47	17.73	6.11	314.47	-43.83
1983	42.32	405.72	347.58	12.17	116.73	-1.45	20.50	2.10	420.40	-29.68
1984	47.42	444.74	362.71	13.07	122.62	-0.55	26.39	0.30	696.61	-14.48
1985	62.25	477.99	377.13	16.51	126.74	2.88	30.52	8.32	931.49	88.04
1986	74.38	505.67	392.73	18.94	128.76	5.32	32.53	28.27	1058.53	172.99
1987	83.87	551.60	409.97	20.46	134.55	6.84	38.32	46.72	1468.75	261.97
1988	87.81	605.91	426.39	20.59	142.10	6.97	45.88	48.62	2105.07	319.93
1989	91.45	650.75	447.73	20.42	145.34	6.80	49.12	46.28	2412.99	334.19
1990	92.26	669.51	468.95	19.67	142.77	6.05	46.55	36.61	2166.45	281.64
1991	97.88	676.48	495.46	19.76	136.54	6.13	40.31	37.63	1625.23	247.30
1992	102.79	690.12	502.84	20.44	137.24	6.82	41.02	46.51	1682.78	279.77
1993	108.98	712.86	512.11	21.28	139.20	7.66	42.98	58.66	1847.10	329.17
1994	118.83	747.26	513.05	23.16	145.65	9.54	49.43	90.99	2443.32	471.52
1995	128.83	776.30	524.19	24.58	148.10	10.96	51.87	120.01	2690.85	568.28
SUM	1583.36	11316.84	9656.76	531.24	3752.67	-0.00	-0.00	1027.40	51809.36	6526.58
MEAN	40.60	290.18	247.61	13.62	96.22	-0.00	-0.00			

Columns (1) and (2) of the worksheet give the Canadian nominal money supply and Canadian nominal Gross Domestic Product (GDP) in billions of

current dollars. Gross Domestic Product is a measure of aggregate nominal income produced in the domestic economy. Column (3) gives the Canadian Consumer Price Index (CPI) on a base of 1963-66 = 100. The theory specifies a relationship between real money holdings and real income. Accordingly, real money holdings and real GDP are calculated in columns (4) and (5) by dividing the nominal values of these variables by the CPI and then multiplying by 100. Thus RMON and RGDP measure the Canadian real money supply and Canadian real GDP in constant 1963-66 dollars. So equation (8.31) above specifies that the numbers in column (4) should be a constant fraction of the numbers in column (5) plus a random error. So we want to run the following simple linear regression:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon \tag{8.32}$$

where Y_t is RMON (column (4)) and X_t is RGDP (column (5)). Because the observations occur through time we designate them by subscript t rather than subscript i.

To obtain a fitted line to these data we perform the calculations shown in columns (6) through (10). The columns D-RMON and D-RGDP give the deviations of RMON and RGDP from their respective means, 13.62 and 96.22, calculated at the bottom of columns (4) and (5). Column (8) gives D-RMON squared and column (9) gives D-RGDP squared. The sums at the bottom of these columns thus give

$$\sum_{t=1957}^{1995} (Y_t - \bar{Y})^2 = 1027.40$$

and

$$\sum_{t=1957}^{1995} (X_t - \bar{X})^2 = 51809.36$$

respectively. Column (10) gives the product of D-RMON and D-RGDP and the sum at the bottom gives

$$\sum_{t=1957}^{1995} (Y_t - \bar{Y})(X_t - \bar{X}) = 6526.58.$$

The estimate b_1 of β_1 can thus be calculated as

$$b_1 = \frac{\sum_{t=1957}^{1995}(Y_t - \bar{Y})(X_t - \bar{X})}{\sum_{t=1957}^{1995}(X_t - \bar{X})^2} = \frac{6526.58}{51809.36} = .126$$

and the estimate b_0 of β_0 becomes

$$b_0 = \bar{Y} - b_1\bar{X} = 13.62 - (.126)(96.22) = 1.5.$$

Next we need the R^2. This equals the square of

$$r = \frac{\sum_{t=1957}^{1995}(Y_t - \bar{Y})(X_t - \bar{X})}{\sqrt{\sum_{t=1957}^{1995}(X_t - \bar{X})^2}\sqrt{\sum_{t=1957}^{1995}(Y_t - \bar{Y})^2}} = \frac{6526.58}{\sqrt{51809.36}\sqrt{1027.40}} = .8946,$$

or $R^2 = (.8946)^2 = .8$. This means that the sum of squares explained by the regression is

$$SSR = R^2 \sum_{t=1957}^{1995} (Y_t - \bar{Y})^2 = (.8)(1027.40) = 821.92$$

and the sum of squared errors is

$$SSE = (1 - R^2) \sum_{t=1957}^{1995} (Y_t - \bar{Y})^2 = (.2)(1027.40) = 205.48.$$

The mean square error is then

$$MSE = \frac{SSE}{n-2} = \frac{205.48}{37} = 5.55.$$

To test whether there is a statistically significant relationship between real money holdings and real GNP we form a t statistic by dividing b_1 by its standard deviation. The latter equals

$$s\{b_1\} = \sqrt{\frac{MSE}{\sum_{t=1957}^{1995}(X_t - \bar{X})^2}} = \sqrt{\frac{5.55}{51809.36}} = .01035.$$

The t-statistic for the test of the null hypothesis that $\beta_1 = 0$ thus equals

$$t^* = \frac{b_1 - 0}{s\{b_1\}} = \frac{.126}{.01035} = 12.17.$$

Since this exceeds the critical value of t of 3.325 for $\alpha = .01$, the null-hypothesis of no relation between real money holdings and real income must be rejected.

To test the null-hypothesis that the constant term β_0 equals zero we obtain the standard deviation of b_0 from (8.30),

$$s^2\{b_0\} = MSE \left[\frac{1}{n} + \frac{\bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

which yields

$$s\{b_0\} = \sqrt{5.55\left[\frac{1}{39} + \frac{96.22^2}{51809.36}\right]} = \sqrt{(5.55)\left[.02564 + \frac{9258.29}{51809.36}\right]}$$

$$= \sqrt{(5.55)(.02564 + .1787)} = 1.064935.$$

The t-statistic for the test of the null hypothesis that $\beta_0 = 0$ is thus

$$t^* = \frac{b_0 - 0}{s\{b_0\}} = \frac{1.5}{1.064935} = 1.409$$

for which the P-value for a two-tailed test is .1672. We cannot reject the null hypothesis that β_0 equals zero at a reasonable significance level.

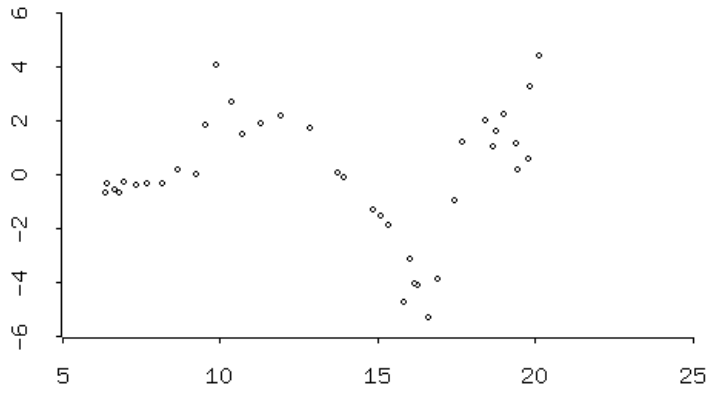

Figure 8.10: The residuals of the regression of Canadian real money holdings on Canadian real GDP, plotted against the fitted values.

The classical hypothesis that the public's real money holdings tend to be a constant fraction of their real income cannot be rejected on the basis of the data used here, because we cannot reject the hypothesis that the true relationship between RMON and RGNP is a straight line passing through the origin. Nevertheless, we must be open to the possibility that the ratio of RMON to RGNP, though perhaps independent of the level of real income, could depend on other variables not in the regression, such as the rate of interest (which equals the opportunity cost of holding money instead of

interest-bearing assets). If this were the case, we might expect the residuals from the regression to be poorly behaved. Figure (8.10) plots the residuals against the fitted values. The residuals are clearly not randomly scattered about zero. It is useful to check for serial correlation in these residuals by plotting them against time. This is done in Figure (8.11). There is obvious serial correlation in the residuals from the regression. We will address this problem again in the next chapter when we investigate the Canadian demand function for money using multiple regression.

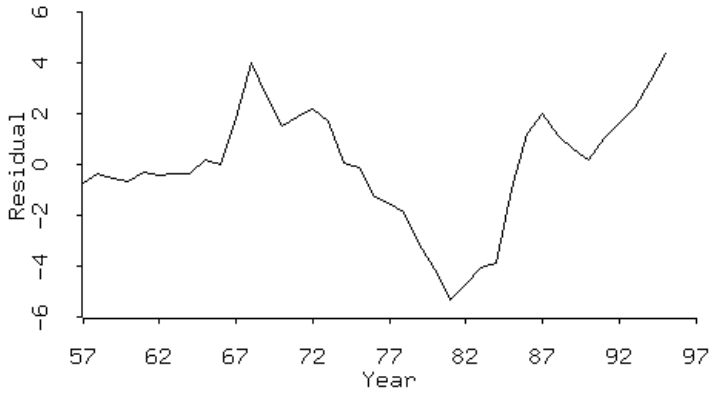

Figure 8.11: The residuals of the regression of Canadian real money holdings on Canadian real GDP, plotted against time.

8.13 Exercises

1. The following data relate to the model

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

where the X_i are assumed non-stochastic and the ϵ_i are assumed to be independently identically normally distributed with zero mean and constant variance.

i	Y_i	X_i
1	21	10
2	18	9
3	17	8
4	24	11
5	20	11
6	20	10
7	22	12
8	21	11
9	17	9
10	20	9

a) Calculate the regression estimates of α and β. (5.71, 1.43)

b) Calculate a 95% confidence interval for β. (0.56, 2.27)

2. Insect flight ability can be measured in a laboratory by attaching the insect to a nearly frictionless rotating arm by means of a very thin wire. The "tethered" insect then files in circles until exhausted. The non-stop distance flown can easily be calculated from the number of revolutions made by the arm. Shown below are measurements of this sort made on *Culex tarsalis* mosquitos of four different ages. The response variable is the average (tethered) distance flown until exhaustion for 40 females of the species.

Age, X_i (weeks)	Distance Flown, Y_i (thousands of meters)
1	12.6
2	11.6
3	6.8
4	9.2

Estimate α and β and test the hypothesis that distance flown depends upon age. Use a two-sided alternative and the 0.05 level of significance.

3. A random sample of size $n = 5$ is to be used to estimate the values of the unknown parameters of the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where the random error term ϵ_i is $N(0, \sigma^2)$. The sample values for (X_i, Y_i) are

X_i	Y_i
-2	-6
-1	-2
0	-2
1	4
2	6

a) Compute the values of the least-squares estimators for β_0 and β_1.

b) Compute the value of the least-squares estimator for σ^2 and the coefficient of determination, R^2.

c) Conduct a test of the null hypothesis H_0: $\beta_1 \leq 2.0$ versus the alternative hypothesis H_1: $\beta_1 > 2.0$ using $\alpha = .05$ and find the approximate P-value for the test.

d) Compute a 95% confidence interval for the expected value of Y when $X = 5$.

4. The District Medical Care Commission wants to find out whether the total expenditures per hospital bed for a particular item tends to vary with the number of beds in the hospital. Accordingly they collected data on number of beds for the 10 hospitals in the district (Y_i) and the total expenditures per hospital bed (X_i). Some simple calculations yielded the following magnitudes:

$$\bar{Y} = 333.0 \quad \bar{X} = 273.4$$

$$\sum_{i=1}^{10}(Y_i - \bar{Y})^2 = 10756.0$$

$$\sum_{i=1}^{10}(X_i - \bar{X})^2 = 301748.4$$

$$\sum_{i=1}^{10}(X_i - \bar{X})(Y_i - \bar{Y}) = -37498$$

Use simple regression analysis to analyse the effect of number of beds on cost of the item per bed. Can you conclude that there is a relationship between these two variables. Is that relationship positive or negative? Calculate the R^2 and the significance of the regression coefficients. Is the overall relationship between the number of hospitals in a district and total expenditures per hospital bed statistically significant at reasonable levels of α-risk?