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 9

Multiple Regression

While simple regression analysis is useful for many purposes, the assumption that the dependent variable Y depends on only one independent variable is very restrictive. For example, if we want to develop a model to estimate the quantity of bread demanded we can expect the latter to depend, at the very minimum, on the price of bread, on the prices of at least some substitutes and on real income.

9.1 The Basic Model

The basic linear multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots\cdots + \beta_K X_{Ki} + \epsilon_i \qquad (9.1)$$

where $i = 1 \ldots n$ and the ϵ_i are independently normally distributed with mean zero and constant variance σ^2. Actually, we can often get away with less restrictive assumptions about the ϵ_i, namely

$$E\{\epsilon_i\} = 0$$

and

$$E\{\epsilon_i \epsilon_j\} = 0, \quad i \neq j$$
$$E\{\epsilon_i \epsilon_j\} = \sigma^2, \quad i = j.$$

This says that the ϵ_i must be independently distributed with constant variance but not necessarily normally distributed. Our problem is to estimate the parameters β_k, $k = 0 \ldots K$, and σ and to establish confidence intervals and conduct appropriate statistical tests with respect to these parameters.

The n-observations on the dependent variable and the K independent variables can be represented as follows:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + \cdots\cdots + \beta_K X_{K1} + \epsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} + \cdots\cdots + \beta_K X_{K2} + \epsilon_2$$

$$Y_3 = \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} + \beta_3 X_{33} + \cdots\cdots + \beta_K X_{K3} + \epsilon_3$$

$$\cdots\cdots\cdots$$

$$\cdots\cdots\cdots$$

$$\cdots\cdots\cdots$$

$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots\cdots + \beta_K X_{Kn} + \epsilon_n$$

This appears in matrix form as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & X_{31} & \cdots & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & X_{32} & \cdots & \cdots & X_{K2} \\ 1 & X_{13} & X_{23} & X_{33} & \cdots & \cdots & X_{K3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & X_{3n} & \cdots & \cdots & X_{Kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}$$

and can be written

$$\mathbf{Y} = \mathbf{X}\mathcal{B} + \mathcal{E} \tag{9.2}$$

where $\mathbf{Y}$ is an n by 1 column vector, $\mathbf{X}$ is an n by $K+1$ matrix (i.e., a matrix with n rows and $K+1$ columns), $\mathcal{B}$ is a $K+1$ by 1 column vector and $\mathcal{E}$ is an n by 1 column vector. The first column of the matrix $\mathbf{X}$ is a column of 1's.

9.2 Estimation of the Model

Our problem is now to choose an estimate of (9.2) of the form

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e} \tag{9.3}$$

where $\mathbf{b}$ is a $K+1$ by 1 column vector of point estimates of the vector $\mathcal{B}$ and $\mathbf{e}$ is an n by 1 column vector of residuals. According to the method of least squares we choose the vector $\mathbf{b}$ so as to minimize the sum of squared residuals which appears in matrix form as

$$\begin{bmatrix} \epsilon_1 & \epsilon_2 & \epsilon_3 & \cdots & \cdots & \cdots & \cdots & \epsilon_n \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or

$$\mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} e_i^2,$$

where $\mathbf{e}'$ is the transpose of $\mathbf{e}$ and thereby consists of a row vector containing the n errors e_i. This sum of squares can be further represented as

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) \\ &= (\mathbf{Y}' - \mathbf{b}'\mathbf{X}')(\mathbf{Y} - \mathbf{Xb}) \\ &= (\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{Xb} - \mathbf{b}'\mathbf{X}'\mathbf{Y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}) \\ &= (\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{Xb} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}) \end{aligned} \tag{9.4}$$

where the second line uses the facts that the transpose of the sum of two matrices (vectors) is the sum of the transposes and the transpose of the product of two matrices (vectors) is the product of the transposes in reverse order, and the fourth line uses the fact that $\mathbf{Y}'\mathbf{Xb}$ and $\mathbf{b}'\mathbf{X}'\mathbf{Y}$ are identical

scalars—this can be seen by noting that $\mathbf{Y}'\mathbf{X}\mathbf{b}$ is

$$\begin{bmatrix} Y_1 & Y_2 & Y_3 & \cdots & Y_n \end{bmatrix} \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_K X_{K1} \\ \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_K X_{K2} \\ \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} + \beta_3 X_{33} + \cdots + \beta_K X_{K3} \\ \vdots\vdots\vdots \\ \vdots\vdots\vdots \\ \vdots\vdots\vdots \\ \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_K X_{Kn} \end{bmatrix}$$

and $\mathbf{b}'\mathbf{X}'\mathbf{Y}$ is

$$\begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \cdots & \beta_K \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & \cdots & \cdots & 1 \\ X_{11} & X_{12} & X_{13} & \cdots & \cdots & X_{1n} \\ X_{21} & X_{22} & X_{23} & \cdots & \cdots & X_{2n} \\ X_{31} & X_{32} & X_{33} & \cdots & \cdots & X_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{K1} & X_{K2} & X_{K3} & \cdots & \cdots & X_{Kn} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}$$

We now differentiate this system with respect to the vector $\mathbf{b}$ and choose that value of the vector $\hat{\mathbf{b}}$ for which $\partial \mathbf{e}'\mathbf{e}/\partial \mathbf{b} = 0$. We thus obtain

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{Y} \tag{9.5}$$

which yields

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{9.6}$$

where $(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of the matrix $\mathbf{X}'\mathbf{X}$.

The system of equations (9.5) is called the *least-squares normal equations*. In the case where there are only two independent variables plus a constant term (i.e., $K = 2$), these equations are

$$\begin{array}{ccccccc} n\,\hat{b}_0 & + & \hat{b}_1 \sum X_{1i} & + & \hat{b}_2 \sum X_{2i} & = & \sum Y_i \\ \hat{b}_0 \sum X_{1i} & + & \hat{b}_1 \sum X_{1i}^2 & + & \hat{b}_2 \sum X_{1i}X_{2i} & = & \sum X_{1i}Y_i \\ \hat{b}_0 \sum X_{2i} & + & \hat{b}_1 \sum X_{1i}X_{2i} & + & \hat{b}_2 \sum X_{2i}^2 & = & \sum X_{2i}Y_i \end{array}$$

The coefficients $\hat{b}_k$ can be obtained by actually calculating all of these sums of squares and cross products, substituting the resulting numbers into

the above system of equations, and solving that system simultaneously for the $\hat{b}_k$'s. Alternatively, the data can be expressed in matrix form (i.e., as a vector $\mathbf{Y}$ and matrix $\mathbf{X}$) and the vector $\hat{\mathbf{b}}$ obtained by applying equation (9.6) to $\mathbf{Y}$ and $\mathbf{X}$ using a standard computer linear algebra program.[1] The easiest way to obtain the $\hat{b}_k$, however, is to read the variables X_k and Y into one of the many standard statistical software packages and apply the linear-regression procedure contained in that package. This has the computer do everything—except determine what regression to run and interpret the results! Remember that a computer performs fast calculations but cannot do our thinking for us. It does exactly what it is told—whence the fundamental *gigo* principle, "garbage in → garbage out".[2]

Along with the vector of estimated regression coefficients, the standard statistical packages give the *standard deviations* (or *standard errors*) of these coefficients, the appropriate t-statistics and sometimes the P-values, the minimized sum of squared deviations of the dependent variable from the regression line, and the coefficient of determination or R^2.[3]

9.3 Confidence Intervals and Statistical Tests

To construct confidence intervals and perform statistical tests regarding the regression coefficients we need estimates of the standard deviations or standard errors of these coefficients. The matrix of variances and covariances of the regression coefficients (from which the standard statistical packages present their standard errors) is

$$\begin{bmatrix} Var\{b_0\} & Cov\{b_0b_1\} & Cov\{b_0b_2\} & \cdots\cdots & \cdots\cdots & Cov\{b_0b_K\} \\ Cov\{b_0b_1\} & Var\{b_1\} & Cov\{b_1b_2\} & \cdots\cdots & \cdots\cdots & Cov\{b_1b_K\} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ Cov\{b_0b_K\} & Cov\{b_1b_K\} & Cov\{b_2b_K\} & \cdots\cdots & \cdots\cdots & Var\{b_K\} \end{bmatrix}$$

$$= E\{(\hat{\mathbf{b}} - \mathcal{B})(\hat{\mathbf{b}} - \mathcal{B})'\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

[1] Such as, for example, MATLAB, MAPLE or OCTAVE. The first two of these are commercial programs while the latter one is freely available over the internet.

[2] Attention should also be paid to a second important principle of computing, *rtfm*. The first letter of this acronym stands for the word "read" and the last letter stands for the word "manual"!

[3] XlispStat has been used for most of the regression calculations, as well as the graphics, in this book

As in the case of simple regression the appropriate estimator for σ^2 is

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{df} = MSE$$

where $df = n - K - 1$ is the degrees of freedom and

$$\mathbf{e}'\mathbf{e} = \sum e_i^2 = SSE$$

is the minimized sum of squared deviations of Y_i from the regression line. The degrees of freedom is $n-K-1$ because we are using the data to estimate $K + 1$ parameters (for K dependent variables plus a constant term). The sum of squares 'explained' by the independent variables is

$$SSR = (\mathbf{Y} - \bar{Y})'(\mathbf{Y} - \bar{Y}) - \mathbf{e}'\mathbf{e} = \sum (Y_i - \bar{Y})^2 - \sum e_i^2 = SSTO - SSE.$$

where $\bar{Y}$ is the mean value of the dependent variable–i.e., the mean of the elements of $\mathbf{Y}$. As in the case of simple linear regression, the fraction of the variation in the dependent variable explained by the independent variables—the R^2—is equal to

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

Notice that the addition of new independent variables to the regression will always increase the R^2. To see this, think of an experiment whereby we keep adding independent variables until the total number of these variables plus the constant equals the total number of observations—this would yield an R^2 equal to unity. We can thus 'explain' more and more of the variation in the dependent variable by adding additional independent variables, paying little attention to whether the variables added are relevant determinants of the dependent variable. To obtain a more meaningful measure of how much of the variation in the dependent variable is being explained, the R^2 is frequently adjusted to compensate for the loss in the degrees of freedom associated with the inclusion of additional independent variables. This adjusted R^2, called the $\bar{R}^2$ is calculated according to the formula

$$\bar{R}^2 = 1 - \frac{n-1}{n-K-1}\frac{SSE}{SSTO}. \tag{9.7}$$

For $\bar{R}^2$ to rise as the result of the addition of another independent variable, the sum of squares of the residuals must fall sufficiently to compensate for the effect of the addition of that variable on the number of degrees of freedom.

The ratio of $(\hat{b}_k - \beta_k)$ to its standard deviation

$$t^* = \frac{\hat{b}_k - \beta_k}{\sqrt{Var\{\hat{b}_k\}}} \tag{9.8}$$

is distributed according to the t-distribution with degrees of freedom

$$df = n - K - 1.$$

The t-table at the back of any textbook in statistics can be used to establish critical values and confidence intervals. The t-values associated with the null hypothesis H_0: $\beta_k = 0$ are also given in most standard statistics computer packages. To test the null hypothesis that β_k takes a particular hypothesized value, or exceeds or falls short of a particular hypothesized value, we divide the difference between the estimated value $\hat{b}_k$ and the hypothesized value β_k by the standard error of $\hat{b}_k$, also given in most statistics computer packages. The P-values given by standard computer packages are the probabilities of observing values of the coefficients as different from zero (in either direction since the test is two-tailed) as are the respective estimated coefficients when the true value of the coefficient in question is zero. It should be noted here that, when conducting tests and setting up confidence intervals, the constant term is treated simply as another coefficient.

9.4 Testing for Significance of the Regression

Frequently we want to test whether the regression itself is significant—that is, whether the independent variables taken as a group explain any of the variation in the dependent variable. The R^2 measures this, but it is a point estimate which could be as high as it is simply because of sampling error. What we want to test is the null hypothesis

$$H_0\colon\ \beta_1 = \beta_2 = \beta_3 = \ldots\ldots = \beta_K = 0$$

against the alternative hypothesis that at least one of these coefficients is different from zero. Notice that this null-hypothesis does not require that β_0, the constant term, be zero—indeed, when there is no relationship between all K independent variables and Y_i, the constant term will be $\beta_0 = \bar{Y}$.

When we run the regression we choose the coefficients $\hat{b}_k$ that minimize the sum of squared residuals $\sum e_i^2$. If the independent variables do not contribute at all to explaining the variations in Y_i we would expect the

minimized sum of squared residuals to be the same as the sum of squared deviations of the Y_i about their mean, $\sum (Y_i - \bar{Y})^2$. That is, we would expect *SSE* to equal *SSTO*. To the extent that

$$\sum e_i^2 \leq \sum (Y_i - \bar{Y})^2$$

there is evidence that the independent variables included in the regression have some explanatory power. The trouble is, however, that *SSE* could be less than *SSTO* strictly as a result of sampling error. We must therefore test whether the observed excess of *SSTO* over *SSE* is bigger than could reasonably be expected to occur on the basis of sampling error alone.

We have already seen that a sum of squares of independently and identically distributed normal random variables divided by their variance is distributed as χ^2 with degrees of freedom equal to the number of independent squared normal deviations being summed. This means that

$$\frac{\sum e_i^2}{\sigma^2} = \chi^2(n - K - 1) \tag{9.9}$$

and

$$\frac{\sum (Y_i - \bar{Y})^2}{\sigma_y^2} = \chi^2(n - 1). \tag{9.10}$$

It can be shown (though we will not do it here) that the difference between two χ^2 variables is also distributed according to the χ^2 distribution, but with degrees of freedom equal to the difference between the degrees of freedom of the two χ^2 variables. This implies that

$$\frac{\sum e_i^2}{\sigma^2} - \frac{\sum (Y_i - \bar{Y})^2}{\sigma_y^2} = \frac{\sum e_i^2 - \sum (Y_i - \bar{Y})^2}{\sigma^2} = \chi^2(K). \tag{9.11}$$

Here $\sigma_y^2 = \sigma^2$ under the null hypothesis that adding the independent variables to the regression has no effect on the residual variance.

We have also learned earlier that the ratio of two independent χ^2 distributions divided by their respective degrees of freedom is distributed according to the F-distribution with two parameters equal to the number of degrees of freedom in the numerator and denominator respectively. Thus, using (9.9) and (9.11) we obtain

$$\frac{\sum (Y_i - \bar{Y})^2 - \sum e_i^2}{K} \div \frac{\sum e_i^2}{n - K - 1} = F(K, n - K - 1) \tag{9.12}$$

where the σ^2 variables in the denominators of (9.9) and (9.11) cancel out. If the independent variables contribute nothing to the explanation of the dependent variable we would expect $\sum (Y_i - \bar{Y})^2$ to approximately equal $\sum e_i^2$ and the calculated F-statistic to be close to zero. On the other hand, if the independent variables do explain some of the variation in the dependent variable the F-statistic will be substantially positive. The question then is whether the probability of observing a value of F as high as the one observed for this particular sample, given that the independent variables truly explain none of the variation in the dependent variable, is small enough that we can reject the null hypothesis of no effect. We choose a critical value of F based on the desired α-risk and reject the null hypothesis if the value of the F-statistic obtained from the sample exceeds this critical value.

Notice now that we can substitute

$$SSR = \sum (Y_i - \bar{Y})^2 - \sum e_i^2$$

and

$$SSE = \sum e_i^2$$

into (9.12) to obtain

$$\frac{n-K-1}{K}\frac{SSR}{SSE} = F(K, n-K-1) \tag{9.13}$$

which can be further simplified using the facts that

$$SSR = R^2 SSTO$$

and

$$SSE = (1-R^2)\, SSTO$$

to produce

$$\left[\frac{n-K-1}{K}\right]\left[\frac{R^2}{1-R^2}\right] = F(K, n-K-1). \tag{9.14}$$

We can thus calculate the F-statistic using the values for R^2, n and K without calculating the total sum of squares and the sum of squared errors.

The basic principle behind (9.12) can be generalized to test the significance of subsets of the β_k and of relationships between various β_k. The test of the significance of a regression involves a comparison of the residuals obtained from the regression and the residuals obtained from the same

regression with everything but the constant term omitted (i.e., with all coefficients but the constant term set equal to zero). We could test the joint significance of, say, two of the K independent variables, X_2 and X_3, by running the regression with these two variables omitted and comparing the residuals so obtained with the residuals from the regression with the two variables included. This is called a *test of restrictions*. The two restrictions in this example are $\beta_2 = 0$ and $\beta_3 = 0$. The null hypothesis is

$$H_0 \colon \beta_2 = \beta_3 = 0$$

against the alternative hypothesis that either β_2 or β_3 is non-zero. We call the sum of squared residuals from the regression that excludes X_2 and X_3 the *restricted residual sum of squares*, $\sum e_{iR}^2$, and the sum of squares of the residuals from the full regression the *unrestricted residual sum of squares*, $\sum e_i^2$. The question is then whether imposing the restrictions raises the residual sum of squares by a 'significant' amount—that is, by an amount which would have a probability less than α of occurring if the restrictions truly have no effect on the explanatory power of the regression. The relevant F-statistic is

$$\frac{\sum e_{iR}^2 - \sum e_i^2}{v} \div \frac{\sum e_i^2}{n-K-1} = F(v, n-K-1) \tag{9.15}$$

where v (= 2 in this example) is the number of restrictions imposed on the regression. If the resulting F-statistic is above the critical value we can reject the null hypothesis that two coefficients β_2 and β_3 are both equal to zero and accept the alternative hypothesis that at least one of them is non-zero.

The same approach can be used to test particular hypotheses about the relationship between two coefficients. Suppose we have reason to believe that β_3 should be the negative of β_2. We can test this single restriction by formulating the null hypothesis

$$H_0 \colon \beta_3 = -\beta_2$$

and testing it against the alternative hypothesis

$$H_1 \colon \beta_3 \neq -\beta_2.$$

The null hypothesis implies the regression model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots\cdots + \beta_K X_{Ki} + \epsilon_i \\ &= \beta_0 + \beta_1 X_{1i} - \beta_3 X_{2i} + \beta_3 X_{3i} + \cdots\cdots + \beta_K X_{Ki} + \epsilon_i \\ &= \beta_0 + \beta_1 X_{1i} + \beta_3 (X_{3i} - X_{2i}) + \cdots\cdots + \beta_K X_{Ki} + \epsilon_i. \end{aligned} \tag{9.16}$$

We therefore construct the new variable $(X_3 - X_2)$ and replace the two variables (X_2 and X_3) in the regression with it. The residuals from this new regression can be designated $\sum e_{iR}^2$ and inserted into (9.15) together with a value of v equal to 1, representing the single restriction, and a sample F-statistic so obtained. If this statistic exceeds the critical value of F for the appropriate degree of α-risk we reject the null hypothesis and conclude that β_3 is not equal to the negative of β_2.

9.5 Dummy Variables

The independent variables in a multiple regression need not be quantitative. For example, suppose we have some data on the salaries of managers in industry and their years of education and want to investigate whether individuals' years of education affect their salaries. We run a simple regression of salary on years of education for the data in question and obtain the following results (the standard errors of the coefficients are given in brackets and $\hat{\sigma}$ is a point estimate of σ):

Dependent Variable: Salary in \$000's

Constant	38.91	(12.88)
Years of Education	.064	(0.898)
R-Squared	.00036	
Standard Error ($\hat{\sigma}$)	8.97	
Number of Observations	8	
Degrees of Freedom	6	

The null hypothesis that years of education has a zero or negative effect on salary cannot be rejected at any reasonable level of significance given the test statistic

$$t^* = \frac{.064}{.898} = .071269.$$

When we plot the data and impose the fitted regression line on it we get the data points and the virtually horizontal regression line in Figure 9.1.

Upon examining the data, it turns out that all the data points above the nearly horizontal fitted line are for individuals who are sales managers and all the data points below the line are managers who are not in sales. Our regression should obviously contain a variable specifying whether or not the individual in the sample is a sales manager. This variable is a qualitative

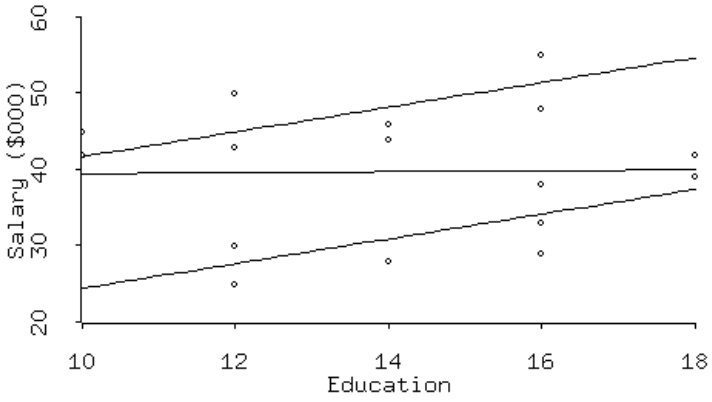

Figure 9.1: Plot and fitted lines of regression of salaries of sales managers on years of education (top line), other managers on years of education (bottom line) and all managers on years of education (middle line).

variable, usually referred to as a *dummy variable*. It consists entirely of zeros or ones—with the variable taking the value of 1 if the individual is a sales manager and 0 if the individual is not a sales manager.

Our regression model now takes the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon \tag{9.17}$$

where the variable X_1 is salary and X_2 is the dummy variable.

Consider the individual sample elements that do not represent sales managers. For these elements $X_{2i} = 0$ so the equation being fitted yields the predicted values

$$\hat{Y}_i \quad = \quad b_0 + b_1 X_{1i}. \tag{9.18}$$

For the individual sample elements that do represent sales managers, $X_{2i} = 1$ so the fitted equation becomes

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} = b_0 + b_1 X_{1i} + b_2$$

or

$$\hat{Y}_i \quad = \quad \tilde{b}_0 + b_1 X_{1i} \tag{9.19}$$

where $\tilde{b}_0 = b_0 + b_2$. Adding the dummy variable essentially allows the regression to have different constant terms for those managers who are sales managers and for those who are not sales managers. When we run this regression we get the following results:

Dependent Variable: Salary in $000's

Constant	8.254	(6.40)
Years of Education	1.62	(0.41)
Sales Manager Dummy	17.28	(2.05)
R-Squared	.845	
Standard Error ($\hat{\sigma}$)	3.66	
Number of Observations	16	
Degrees of Freedom	13	

Notice how the R^2 increases and the standard error of the regression falls when we add the dummy variable. Notice also that the test statistic for the null hypothesis that the true coefficient of the years-of-education variable is zero or less is now

$$t^* = \frac{1.62}{0.41} = 3.95$$

which has a P-value equal to .00083, so we can easily reject the null hypothesis at an α-risk of .001.

The predicted salary levels for each level of education for sales managers is given by the top upward-sloping line in Figure 9.1 and the predicted salary levels for each education level for non-sales managers is given by the lower upward-sloping line. These lines are very close to the fitted lines that would be obtained by running separate regressions for sales managers and for other managers.

We could include a second dummy variable to account for differences in the slope of the relationship between education and salary for the two groups of managers. This variable would be the product of the sales-manager dummy and the years of education—when the data element is a manager not in sales this variable would take a zero value and when the data element is a sales manager the variable would take a value equal to years of education. This dummy variable can be referred to as an *interaction* between years of education and whether the manager was sales vs. non-sales. The regression model would then be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon. \tag{9.20}$$

For data elements representing non-sales managers the predicted values will be

$$\hat{Y}_i = b_0 + b_1 X_{1i} \tag{9.21}$$

since both X_{2i} and X_{3i} will be zero for these elements. For data elements representing sales managers the predicted values will be

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 + b_3 X_{1i}$$

since for these elements $X_{3i} = X_{1i}$ and $X_{2i} = 1$, so we have

$$\hat{Y}_i = \tilde{b}_0 + \tilde{b}_1 X_{1i} \tag{9.22}$$

where $\tilde{b}_0 = b_0 + b_2$ and $\tilde{b}_1 = b_1 + b_3$.

The inclusion of dummies for both the constant term and the slope coefficient turns out to be equivalent to running two separate regressions—one for sales managers and one for other managers—except that by pooling the data and running a single regression with dummy variables included for the constant term and slope parameters we are imposing the assumption that the variance of the error term is the same in the separate regression models. Unless we have prior information about the variance of the errors there is no gain to pooling the data for the two types of managers in this case. When we include only a single dummy variable to allow, say, for differences in the constant term there is a gain from pooling the data and running a single regression provided we are prepared to force upon the model the assumption that the response of salary to years of education is the same for sales managers as for other managers. If we are not prepared to assume that the response of salary to education is the same for both groups we should run two regressions. It would still be appropriate to add two dummy variables, one for the constant term and one for the slope of salary with respect to education of sales vs. other managers, if we also have additional variables in the regression such as, for example, education of the individual manager's parents and race or religion. In this case, of course, the pooled regression will be appropriate only if we are willing to impose on the estimation the assumption that the effects of parents' education, race and religion are the same for sales managers and other managers.

9.6 Left-Out Variables

Frequently we do not have the data to include in a regression a variable that should be there. When this is the case we can often form an opinion, based on casual knowledge, about the effects of the coefficients of the included variables of leaving out a variable that should be in the regression. Suppose that the correct specification of the regression equation is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon \tag{9.23}$$

but we estimate

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon^* \tag{9.24}$$

instead.

Since in the case we are examining the regression actually estimated is a simple regression, our least-squares estimate of β_1 is

$$\hat{b}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}. \tag{9.25}$$

From the true relationship we know that

$$\begin{aligned} Y_i - \bar{Y} &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon^* - \beta_0 - \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 - \bar{\epsilon}^* \\ &= \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + \epsilon^* \end{aligned} \tag{9.26}$$

Upon substitution of this equation into (9.25), the expected value for $\hat{b}_1$ becomes

$$\begin{aligned} E\{\hat{b}_1\} = \hat{\beta}_1 &= \frac{\beta_1 \sum_{i=1}^n (X_i - \bar{X})^2 + \beta_2 \sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \beta_1 + \beta_2 \left[\frac{\sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]. \end{aligned} \tag{9.27}$$

The term in the big square brackets will be recognized as the slope coefficient of a regression of the variable X_2 on the variable X_1. Let us denote this coefficient by d_{21}. Then (9.27) becomes

$$\hat{\beta}_1 = \beta_1 + \beta_2 \, d_{21} \tag{9.28}$$

Suppose that the left-out variable is positively correlated with the included variable X_1 and positively related to the dependent variable. Then β_2 and d_{21} will both be positive and our least-squares estimate of β_1 will

be biased upward. If the left-out variable is negatively correlated with the included variable and positively related to the dependent variable, β_2 will be negative and d_{21} positive so our least-squares estimate of β_1 will be biased downward. If the left-out variable is negatively related to the dependent variable the bias will be upward when the left-out and included variables are negatively related and downward when the left-out and included variables are positively related.

9.7 Multicollinearity

Suppose a young researcher wants to estimate the demand function for money for Canada. She has learned in her intermediate macroeconomics class that the demand for real money holdings can be expressed

$$\frac{M}{P} = L(r_N, Y_R) \tag{9.29}$$

where M is the nominal money stock, P is the price level (so that M/P is the real money stock), r_N is the nominal interest rate and Y_R is the level of real income. This suggest a regression equation of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon \tag{9.30}$$

where Y is Canadian real money holdings, X_1 is the nominal interest rate and X_2 is Canadian real income. In the process of collecting her data, our researcher discovered two different measures of real income, GNP and GDP.[4] Not knowing which to use as her measure of real income, she did the easy thing and simply included both in the regression. Her regression model now becomes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon \tag{9.31}$$

where X_2 is real Canadian GDP and X_3 is real Canadian GNP. She used the Canadian 90-day commercial paper rate as a measure of the Canadian nominal interest rate.[5] All the data series were annual (as opposed to quarterly or monthly) for the years 1957 to 1996 inclusive.

[4]GDP or gross domestic product measures the level of aggregate real output produced by resources employed in the country while GNP or gross national product measures the level of aggregate real output produced by resources owned by domestic residents. To calculate GNP from GDP we have to subtract out that part of aggregate domestic output (GDP) produced by resources that are owned by foreigners and then add in the part of aggregate output abroad that is produced by resources owned by domestic residents.

[5]The 90-day commercial paper rate is the rate of interest charged on commercial paper—that is, on securities issued by major corporations for short-term borrowing—that becomes due 90 days after issue.

The researcher obtained the following regression results:

Dependent Variable: Canadian Real Money Holdings

Constant	8.50	(4.47)
90-Day Paper Rate	-2.65	(0.39)
Real GDP	-0.32	(0.50)
Real GNP	0.51	(0.53)
R-Squared	.91	
Standard Error ($\hat{\sigma}$)	6.75	
Number of Observations	40	
Degrees of Freedom	36	

Surprised that both real income coefficients were insignificant, the researcher decided to perform an F-test of the null hypothesis that both are simultaneously zero (H_0: $\beta_2 = \beta_3 = 0$). So she ran the same regression with both variables omitted, obtaining the following results:

Dependent Variable: Canadian Real Money Holdings

Constant	43.35	(8.60)
90-Day Paper Rate	1.46	(1.01)
R-Squared	.05	
Standard Error ($\hat{\sigma}$)	21.44	
Number of Observations	40	
Degrees of Freedom	38	

The mean squared errors for the respective regressions are equal to their sums of squared residuals divided by their respective degrees of freedom. Thus, the sum of squared residuals for the unrestricted regression (i.e., the one that included the two real income variables) is

$$\sum e_i^2 = df\hat{\sigma}^2 = (36)(6.75)^2 = 1640$$

and the sum of squared residuals for the restricted regression (the one that excluded the two real income variables) is

$$\sum e_{Ri}^2 = df\hat{\sigma}^2 = (38)(21.44)^2 = 17467$$

The appropriate test statistic is therefore

$$\frac{\sum e_{Ri}^2 - \sum e_i^2}{v} \div \frac{\sum e_i^2}{n-K-1} = \frac{17467 - 1640}{2} \div \frac{1640}{36}$$

$$= \frac{7913.5}{45.55} = 173.73 = F(v, n-K-1) = F(2, 36)$$

where v is the number of restrictions, equal to 2 in this case. The critical value for $F(2, 36)$ setting the α-risk at .01 is 5.18 so the researcher rejected the null hypothesis that both of the coefficients are zero.

What is happening here? Neither of the income variables is statistically significant in the regression but the two together are significant at far below the 1% level!

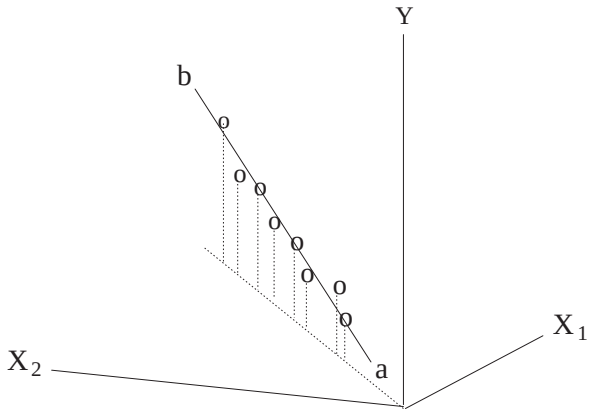

Figure 9.2: An illustration of multicolinearity of X_1 and X_2 in predicting Y.

This is an example of *multicollinearity*. The problem is that GDP and GNP are so highly correlated with each other that they are virtually the same variable. Had they been perfectly correlated, of course, the computer would not have been able to run the regression. Including two perfectly correlated variables in the regression is equivalent to including the same variable twice. This would mean that the $\mathbf{X}$ matrix would have two identical columns so that it would be non-singular and the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix would not exist. The problem here is that the two variables are not identical but

nevertheless highly correlated. This makes it impossible to determine their separate influences in the regression. The situation can be seen from Figure 9.2 for a multiple regression containing a constant term and two highly collinear independent variables X_1 and X_2. The purpose of the regression is to identify a plane in X_1, X_2, Y space that indicates how the dependent variable Y responds to changes in X_1 and X_2. When X_1 and X_2 are highly correlated, however, all the points lie very close to a ray projecting outward into X_1, X_2, Y space. It is possible to identify a relationship between X_1 and Y and between X_2 and Y but not between both X_1 and X_2 together and Y. Any estimated plane resting on the line ab in Figure 9.2 will be very unstable in the dimensions X_1, Y and X_2, Y—slightly different placements of the points in different samples will lead to planes with very different slopes in the X_1, Y and X_2, Y dimensions.

The researcher's solution to the problem in this case is easy—simply drop one of the income variables from the regression, since both are measuring the same thing, real income. Dropping real GDP, she obtains the following results:

Dependent Variable: Canadian Real Money Holdings

Constant	10.47	(3.21)
90-Day Paper Rate	-2.62	(0.38)
Real GNP	0.17	(0.01)
R-Squared	.91	
Standard Error ($\hat{\sigma}$)	6.70	
Number of Observations	40	
Degrees of Freedom	37	

Situations arise, however, in which two collinear variables really measure different things and we therefore want to identify their separate effects on the dependent variable. Suppose, for example, that we want to measure the effects of domestic and foreign real incomes and domestic relative to foreign prices on a country's balance of trade. The theoretical equation takes the form

$$B_T = B(Y_R^D, Y_R^F, P_R) \tag{9.32}$$

where Y_R^D is domestic real income, Y_R^F is foreign real income and P_R is the relative price of domestically produced goods in terms of foreign produced

goods with all prices measured in a common currency.[6] The appropriate regression model would be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon \tag{9.33}$$

where the dependent variable Y is the real balance of trade, X_1 is domestic real income Y_R^D, X_2 is foreign real income Y_R^F is foreign real income, and X_3 is the relative price of domestic goods, P_R. Since a rise in the relative price of domestic in terms of foreign goods will cause both domestic and foreign residents to switch their purchases away from domestic goods, increasing imports and reducing exports, we would expect the real balance of trade to be negatively affected, so the expected sign of β_3 is negative. An increase in domestic income might be expected to cause domestic residents to buy more foreign goods, increasing imports and reducing the real balance of trade. We would therefore expect β_1 to also be negative. An increase in foreign income, on the other hand, might be expected to cause foreigners to import more, resulting in an expansion of domestic exports and an increase in the balance of trade. The coefficient β_2 would thus expected to take on a positive sign.

When we estimate equation (9.33) for some country pairs we might find that the domestic and foreign real income variables are so highly collinear that our estimates of β_1 and β_2 will be statistically insignificant. If we drop one of the variables, the remaining real income variable acts as a measure of world real income and the response of the real balance of trade to that variable will measure the effect of a proportional rise in both domestic and foreign income on net domestic exports. Our purpose, however, is to measure the separate effects of the two income variables on the domestic real trade balance. There is no way that we can do this on the basis of the information provided by the data we are using. The only way to solve our problem is to obtain more information.

[6]The variable P_R is called the *real exchange rate*. The nominal exchange rate is the price of one country's money in terms of another country's money while the real exchange rate is the price of one country's output in terms of another country's output.

9.8 Serially Correlated Residuals

Perhaps the most important basic assumption of the linear regression model is that the errors ϵ_i are *independently* distributed. This means that the error associated with the i-th observation does not in any way depend on the error associated with the j-th observation. This assumption is frequently violated in regressions involving time series because the errors are correlated through time. As noted earlier, this situation is called *serial correlation* or *autocorrelation*. High (low) values at any point in time are associated with high (low) values in neighbouring points in time when there is positive autocorrelation.

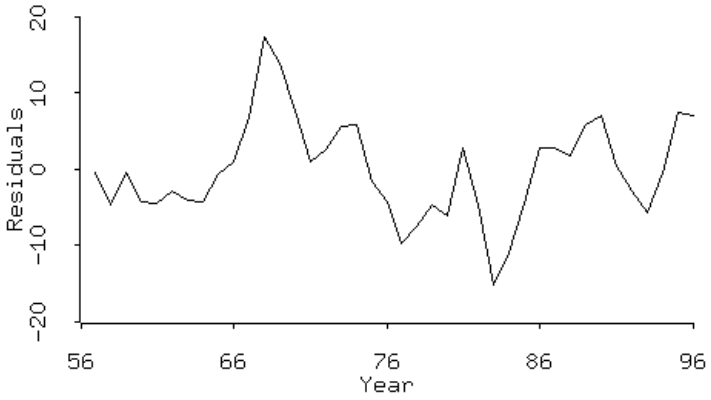

Figure 9.3: Residuals from the regression of Canadian real money holdings on the country's 90-day commercial paper rate and real GNP plotted against time.

Consider the regression of Canadian real money holdings on the 90-day commercial paper rate and real GNP reported above. The residuals from that regression are reported in Figure 9.3. It is clear from looking at the figure that these residuals are serially correlated—high values in one period are clearly associated with high values in immediately adjacent periods. To demonstrate formally that serial correlation is present, we can regress each year's residual on the residuals for several previous years. Using three lags, we obtain

Dependent Variable: Residual

Constant	.0544	(0.749122)
Residual-lag-1	1.0279	(0.171394)
Residual-lag-2	-0.4834	(0.233737)
Residual-lag-3	.0959	(0.175700)
R-Squared	.5849	
Standard Error ($\hat{\sigma}$)	4.5437	
Number of Observations	37	
Degrees of Freedom	33	

Statistically significant coefficients were obtained for one and two lags of the residuals—based on t-ratios of 6.0 and -2.06, respectively. The third lag is clearly insignificant. When the residual is correlated with the immediately previous residual, the serial correlation is called *first-order* serial correlation, when it is correlated with the residual two periods previous it is called *second-order* serial correlation, and so forth. In the above case, there is first- and second-order serial correlation in the residuals but not third-order. We do not know whether fourth-order serial correlation is present because we did not test for it—it is possible to have fourth- (or any other) order serial correlation in the residuals without having serial correlation of lower orders.

The standard procedure for detecting first-order (and only first-order) serial correlation in the residuals is to calculate the *Durbin-Watson Statistic*. This equals

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}. \tag{9.34}$$

The sampling distribution of d is a complex one. It turns out that d can take values between 0 and 4, and will differ from 2 when first-order serial correlation is present. When the first-order serial correlation is positive, d will be less than 2 and when it is negative d will be greater than 2. There is, however, a wide range of indeterminacy. In the case of positive serial correlation, one cannot clearly reject the null hypothesis of zero autocorrelation unless d is below the lower bound for the chosen level of α-risk in the table of critical values for the Durbin-Watson d statistic in the back of one's statistics textbook. And one can only accept the hypothesis of zero autocorrelation if d is above the upper bound in the table. For values of d between the lower and upper bounds we cannot draw any conclusion. For

negative serial correlation (which is present when $d > 2$) the same limits are used except we compare the numbers in the table with $4 - d$. In the regression above, the Durbin-Watson statistic is .58 which is well below the lower bound for $\alpha = .01$ and indicates positive first-order serial correlation.

What do we do when first-order serial correlation is present in the residuals? (Dealing with higher order serial correlation is beyond the technical level of the analysis here.) The answer to this question depends on why the autocorrelation is present. One possibility is that the true errors are serially correlated. This implies that the standard linear model is the incorrect one to apply to the data. An appropriate error term might be

$$\epsilon_t = \rho\,\epsilon_{t-1} + u_t$$

which implies that

$$\epsilon_t - \rho\,\epsilon_{t-1} = u_t,$$

where u_t is independently normally distributed with zero mean and variance σ^2. Assuming that the residuals actually behave in this way, we can lag the original regression equation

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \epsilon_t$$

once to yield

$$Y_{t-1} = \beta_0 + \beta_1 X_{1(t-1)} + \beta_2 X_{2(t-1)} + \epsilon_{t-1}$$

and then subtract ρ times the second equation from the first to obtain

$$Y_t - \rho\,Y_{t-1} = \beta_0 + \beta_1(X_{1t} - \rho\,X_{1(t-1)}) + \beta_2(X_{2t} - \rho\,X_{2(t-1)}) + u_t. \quad (9.35)$$

In this equation $(Y_t - \rho\,Y_{t-1})$, $(X_{1t} - \rho\,X_{1(t-1)})$ and $(X_{2t} - \rho\,X_{2(t-1)})$ are related according to the standard linear model with the independently and normally distributed error term u_t.

To estimate equation (9.35), we need an estimator of ρ. A natural way to proceed is to regress the residuals from the original regression on themselves lagged one period and use the slope coefficient as that estimator. Our regression model would be

$$e_t = \gamma + \rho\,e_{t-1} + \upsilon_t$$

where υ_t is an independent draw from the true constant-variance error term and we would expect our estimate of γ to be zero. The results from this regression are as follows:

Dependent Variable: Residual

Constant	0.140	(0.761913)
Residual-lagged	0.716	(0.118507)
R Squared:	0.497	
Standard Error ($\hat{\sigma}$)	4.7562	
Number of Observations	39	
Degrees of Freedom	37	

We can apply the resulting estimate of ρ (= 0.716) to obtain the new variables

$$\tilde{Y}_t = (Y_t - .716\, Y_{t-1})$$

$$\tilde{X}_{1t} = (X_{1t} - .716\, X_{1(t-1)})$$

and

$$\tilde{X}_{2t} = (X_{2t} - .716\, X_{2(t-1)}).$$

A new regression of the form

$$\tilde{Y}_t = \beta_0 + \beta_1 \tilde{X}_{1t} + \beta_2 \tilde{X}_{2t} + u_t$$

can then be run, yielding the following results:

Dependent Variable: Real Money Variable

Constant	1.03	(2.03)
Interest rate variable	-1.38	(0.33)
Real GNP variable	0.17	(0.02)
R-Squared	.73	
Standard Error ($\hat{\sigma}$)	4.05	
Number of Observations	39	
Degrees of Freedom	36	

It turns out that the effects of this 'correction' for serial correlation in the residuals, comparing the before and after regressions, reduces the absolute value of the slope coefficient of the interest rate variable from -2.62 to -1.38 and also reduces its standard error slightly. A sophisticated extension of this procedure is to regress the residuals of this new equation on themselves lagged and modify the estimate of ρ accordingly, doing this repeatedly until the estimates of ρ change by less than some minimal amount. When this is done, we obtain the following results:

Dependent Variable: Real Money Variable

Constant	-4.24	(24.62)
Interest rate variable	-1.09	(0.31)
Real GNP variable	0.18	(0.05)
ρ	0.928	(0.07)
R-Squared	.97	
Standard Error ($\hat{\sigma}$)	3.75	
Number of Observations	39	
Degrees of Freedom	35	

These refinements reduce further the absolute value of the slope coefficient of the interest rate variable and its standard error and raise slightly the coefficient of the real income variable and more substantially its standard error.

The 'optimal' value of ρ obtained by the above iterative method is very close to unity. In fact, a long-standing traditional approach to dealing with serial correlation in the residuals has been to take the first differences of the variables and run the regression in the form

$$Y_t - Y_{t-1} = \beta_0 + \beta_1(X_{1t} - X_{1(t-1)}) + \beta_2(X_{2t} - X_{2(t-1)}) + \vartheta_t. \qquad (9.36)$$

This assumes that

$$\vartheta_t = \epsilon_t - \epsilon_{t-1}$$

is independently and normally distributed and therefore that $\rho = 1$. When we impose this assumption on the residuals we obtain the following results:

Dependent Variable: Real Money Variable

Constant	0.496	(0.89)
Interest rate variable	-0.96	(0.32)
Real GNP variable	0.15	(0.06)
R-Squared	.22	
Standard Error ($\hat{\sigma}$)	3.73	
Number of Observations	39	
Degrees of Freedom	36	

The results differ little from those obtained when ρ was estimated iteratively.

Which coefficients are we to believe, those with no 'correction' of the residuals for serial correlation or those with a 'correction' imposed? To answer this question we must know the reason for the residuals being serially correlated. One possibility, of course, is that the residuals of the 'true' model are serially correlated. The problem with this explanation is that there is no reason in economic theory for the residuals to be serially correlated if we have correctly modeled the economic process we seek to explain. The reason why we have serial correlation in the residuals is that we have left variables that are correlated with time out of the model because we either could not measure them or could not correctly specify the underlying theory given the current state of knowledge. Obviously, the best approach is to try to better specify the model and to be sure that all variables that should be in it are included in the estimating equation. If we cannot do so our coefficients are likely to be biased for reasons outlined in section 9.6 above on left-out variables. Whether we improve things by correcting the residuals for first-order serial correlation is a question that econometricians will debate on a case-by-case basis. Clearly, however, it is inappropriate to routinely and unthinkingly impose a 'correction' on the residuals every time serial correlation is present.

9.9 Non-Linear and Interaction Models

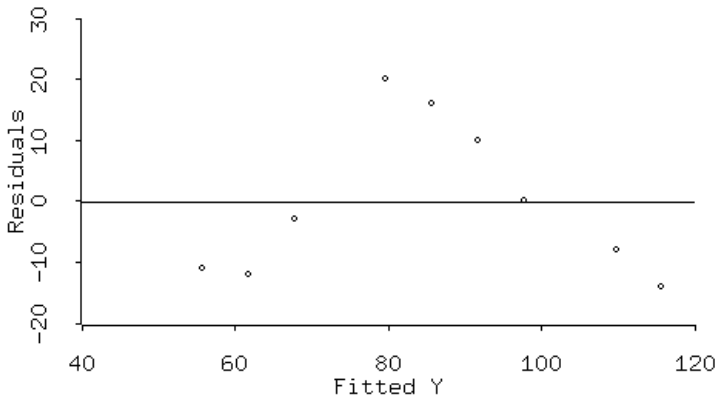

Figure 9.4: Residuals from a linear regression that suggest the underlying relationship is nonlinear.

It frequently arises that the residuals show a non-linear pattern as is illustrated in Figure 9.4. There are a number of simple ways of fitting non-linear relationships—either the dependent or independent variables or both can be transformed by inverting them or taking logarithms and using these non-linear transformations of the variables in a linear regression model. Another way is to include in the regression model squares of the independent variables along with their levels. For example, we might have

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \epsilon_t. \tag{9.37}$$

Interaction models arise when the relationship between the dependent variable and one of the independent variables depends on the level of a second independent variable. In this case, the appropriate regression model would be of the form

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{1t} X_{2t} + \epsilon_t. \tag{9.38}$$

Let us work through an example that illustrates both of these modifications to the standard linear model. It is quite common for colleges and universities to develop regression models for predicting the grade point averages (GPA's) of incoming freshmen. This evidence is subsequently used to decide which students to admit in future years. Two obvious variables that should be predictors of subsequent student performance are the verbal and mathematics scores on college entrance examinations. Data for a randomly-selected group of 40 freshmen were used to obtain the following regression results:

Dependent Variable: Freshman Grade Point Average

Constant	-1.570	(0.4937
Verbal Score (percentile)	0.026	(0.0040)
Math Score (percentile)	0.034	(0.0049)
R-Squared	.68	
Standard Error ($\hat{\sigma}$)	.402	
Number of Observations	40	
Degrees of Freedom	37	

These results indicate that students' scores on both the verbal and mathematical college entrance tests are significant positive predictors of freshman success (with t-ratios 6.3 and 6.8, respectively). An increase in a student's

verbal score by 10 percentiles will lead on average to a .26 increase in his/her GPA. For example, a student in the 70th percentile on both the verbal and mathematics sections of the entrance exam will have an expected freshman GPA of

$$-1.57 + (70)(.026) + (70)(.034) = 2.58.$$

An increase in her verbal score on the entrance exam from the 70th to the 80th percentile will increase her expected GPA by 0.26 to 2.84. And an increase in her math score from the 70th to the 80th percentile will increase her expected GPA by .34 to 2.92. An increase in both of her scores from the 70th to the 80th percentile will increase her expected GPA by .6 (= .26 + .34) to 3.18. The increase in expected GPA predicted by an increase in the percentiles achieved on the mathematics and verbal college entrance exams will be independent of the initial levels of the student's scores.

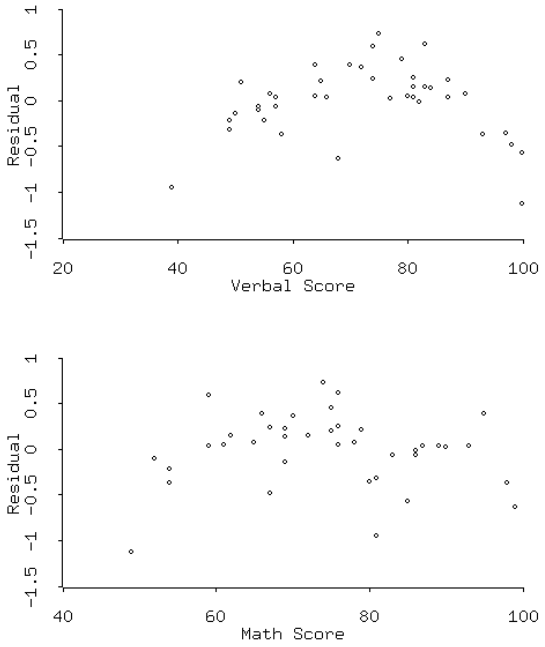

Figure 9.5: Residuals from first order regression model of grade point average on test scores.

The residuals from this regression are plotted against the two independent variables in Figure 9.5. Plotted against verbal score, they have an

inverse parabolic pattern, suggestive of non-linearity.[7] To check this out we run a second regression that includes the squared verbal and mathematics scores as additional variables together with an interactive variable consisting of the product of the verbal and math scores. The results are as follows:

Dependent Variable: Freshman Grade Point Average

Constant	-9.9167	(1.35441)
verbal score	0.1668	(0.02124)
math score	0.1376	(0.02673)
verbal score squared	-0.0011	(0.00011)
math score squared	-0.0008	(0.00016)
verb score x math score	0.0002	(0.00014)
R-Squared	.94	
Standard Error ($\hat{\sigma}$)	.187	
Number of Observations	40	
Degrees of Freedom	34	

As expected, the verbal score squared has a significant negative sign indicative of an inverse parabolic relationship (the t-statistic equals -10). The squared mathematical score is also statistically significant with a negative sign (the t-ratio equals -5). The interactive term, verbal score times math score, is not statistically significant, with a t statistic of only 1.43. The residuals from this extended regression, plotted in Figure 9.6 are very well behaved. An F-test of the null hypothesis of no effect of the squared and interactive terms yields the statistic

$$\frac{\sum e_{iR}^2 - \sum e_i^2}{3} \div \frac{\sum e_i^2}{34} = \frac{(37)(.402)^2 - (34)(.187)^2}{3} \div \frac{(34)(.187)^2}{34}$$

$$= \frac{5.98 - 1.19}{3} \div \frac{1.19}{34} = \frac{1.60}{.035} = 45.71 = F(3, 34).$$

We can reject the null hypothesis at any reasonable level of α-risk.

Notice how the addition of these second order terms (squares and cross-products) affects the response of GPA to verbal and mathematical test

[7] A parabola takes the mathematical form

$$y = ax^2 - bx - c.$$

When $a < 0$ the parabola will be inverted with the arms extending downward.

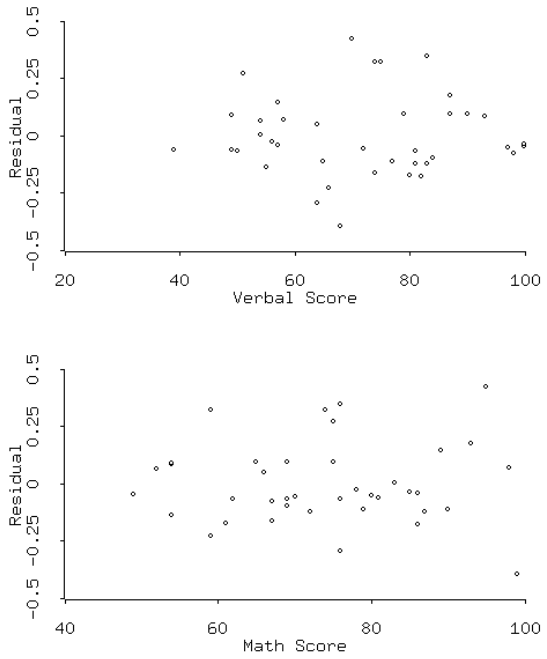

Figure 9.6: Residuals from second order regression model of grade point average on test scores.

scores. A student with scores in the 70th percentile on both the verbal and mathematical tests will have a predicted GPA of

$$-9.9167 + (.1668)(70) + (.1376)(70) - (.0011)(70)^2$$

$$-(.0008)(70)^2 + (.0002)(70)(70) = 3.04$$

which is higher than the predicted value from the regression that did not include the second order terms. Now suppose that the student's verbal test score increases to the 80th percentile. This will increase his expected GPA by

$$(.1668)(80-70) - (.0011)(80^2 - 70^2) + (.0002)(80-70)(70) = .158$$

to 3.198. An increase in the mathematical score from the 70th to the 80th percentile with his verbal score unchanged would increase his expected GPA by

$$(.1376)(80-70) - (.0008)(80^2 - 70^2) + (.0002)(70)(80-70) = .316$$

to 3.356. Given the interaction term, an increase in both the verbal and mathematical scores of the student from the 70th to the 80th percentile would increase his expected GPA by more than the sum of the two separate effects above (= .158 + .316 = .474). The increase would be

$$(.1668)(80-70) - (.0011)(80^2 - 70^2) + (.1376)(80-70) - (.0008)(80^2 - 70^2)$$

$$+(.0002)[(80-70)(70) + (70)(80-70) + (80-70)(80-70)$$

$$= .158 + .316 + (.0002)(100) = .158 + .316 + .02 = .494$$

to 3.534. Notice the difference in the levels and predicted changes in the GPA's under the second order as opposed to the first order model. Given that the interaction term is statistically insignificant, however, we might decide to make our predictions on the basis of a regression model that includes the squared terms but excludes the interaction term.

9.10 Prediction Outside the Experimental Region: Forecasting

A major purpose of regression analysis is to make predictions. Problems arise, however, when the fitted models are used to make predictions outside the range of the sample from which the regression model was estimated—i.e., outside the experimental region. The fit within sample is based on the surrounding sample points. Outside the range of the sample there is no opportunity for the fitted regression parameters to be influenced by sample observations—we simply do not know what values of the dependent variable would be associated with levels of the independent variables in this range were they to occur. As a result, the farther outside the sample range we extrapolate using the estimated model the more inaccurate we can expect those predictions to be.

Predicting outside the sample range in time series regressions is called *forecasting*. We have data on, say, the consumer price index, up to and including the current year and want to predict the level of the consumer price index next year. We develop a regression model 'explaining' past movements in the consumer price index through time and then use that model to forecast the level of the consumer price index in future periods beyond the sample used to estimate the model. To the extent that we use independent variables other than time we have to forecast the levels of those variables because their realization has not yet occurred. Errors in those forecasts will produce errors in predicting the future values of the dependent variable. These will additional to the errors that will result because we are using the regression parameters to predict values of the dependent variable outside the sample range in which those parameters were estimated.

Alternatively, we could forecast the consumer price index based on a simple regression of a range of its previous realized values against time using a model such as

$$Y_T = \beta_0 + \beta_1 T + \epsilon_t$$

where Y_T is the consumer price index at time T. This is the simplest *time-series* model we could fit to the data—time-series econometricians typically use much more sophisticated ones. The regression model is estimated for the period $T = 1, 2, \ldots N$ and then a prediction of Y for period $N + 1$ is obtained as

$$Y_{N+1} = b_0 + b_1 (N + 1).$$

Obviously, if the time-period $N + 1$ could have been used in the estimation of the model, the estimates b_0 and b_1 would be different. The further we forecast beyond period N the less the expected accuracy of our forecasts.

9.11 Exercises

1. A random sample of size $n = 20$ families is used to conduct a multiple regression analysis of how family i's annual savings S_i depends on its annual income I_i and its home-ownership status H_i. Both S_i and I_i are measured in thousands of dollars. Variable H_i is equal to 1 if family i owns its home and equal to 0 if family i rents. The regression results

Coefficient	Estimate	Standard Error
Constant — β_0	-0.320	0.620
Annual Income — β_1	0.0675	0.004
Home Ownership — β_2	0.827	0.075
Sum of Squared Errors	0.230	
Total Sum of Squares	15.725	

yield the fitted equation

$$\hat{S}_i = -0.320 + 0.0675\, I_i + 0.827\, H_i.$$

a) The value of the coefficient associated with the variable I is estimated to be 0.0675. Provide a one-sentence explanation of what this number implies about the relationship between family income and saving. Also, provide a one-sentence explanation of what the coefficient estimate 0.827 implies about the relationship between home ownership and saving.

b) Using $\alpha = .05$, conduct a test of the null hypothesis H_0: $\beta_1 = \beta_2 = 0$ versus the alternative hypothesis that at least one of β_1, β_2 is not equal to zero.

2. A shoe store owner estimated the following regression equation to explain sales as a function of the size of investment in inventories (X_1) and advertising expenditures (X_2). The sample consisted of 10 stores. All variables are measured in thousands of dollars.

$$\hat{Y} = 29.1270 + .5906\, X_1 + .4980\, X_2$$

The estimated R^2 was .92448, $\Sigma(Y_i - \bar{Y})^2 = 6,724.125$, and the standard deviations of the coefficients of X_1 and X_2 obtained from the regression were .0813 and .0567 respectively.

a) Find the sum of squared residuals and present a point estimate of the variance of the error term. (507.81, 72.54)

b) Can we conclude that sales are dependent to a significant degree on the size of stores' inventory investments?

c) Can we conclude that advertising expenditures have a significant effect on sales?

d) Can we conclude that the regression has uncovered a significant overall relationship between the two independent variables and sales?

e) What do we mean by the term 'significant' in b), c) and d) above?

3. Quality control officers at the Goodyear Tire and Rubber Company are interested in the factors that influence the performance of their Goodyear TA All Season Radial Tires. To this end, they performed a multiple regression analysis based on a random sample of 64 automobiles. Each vehicle was equipped with new tires and driven for one year. Following the test period, Goodyear experts evaluated tire wear by estimating the number of additional months for which the tire could be used. For the regression study, the dependent variable *TIRE* measures this estimated remaining lifetime in months. A totally worn out tire will report $TIRE = 0$. Independent variables selected for the study include *WEIGHT* which measures the test vehicle's weight in pounds, *CITY* which measures the number of miles driven in city traffic in thousands and *MILES* which measures the total number of miles driven (city and highway), also in thousands. The statistical software package Xlispstat reports multiple regression results and a simple regression of *TIRE* on *WEIGHT*. The standard errors of the coefficients are given in brackets.

Dependent Variable: *TIRE*

Constant	60.000	(15.000)
WEIGHT	-0.003	(0.001)
CITY	0.020	(0.008)
MILES	-0.400	(0.100)
R-Squared	.86	
Standard Error ($\hat{\sigma}$)	1.542	
Number of Observations	64	
Degrees of Freedom	60	

Dependent Variable: *TIRE*

Constant	72.000	(36.000)
WEIGHT	-0.005	(0.001)
R-Squared	.79	
Standard Error ($\hat{\sigma}$)	1.732	
Number of Observations	64	
Degrees of Freedom	62	

a) Interpret each of the estimated parameters in the multiple regression model (i.e., what does $\beta_2 = 0.020$ tell you about the relationship between city miles and tire wear?)

b) Briefly discuss why the estimated coefficient on *WEIGHT* differs between the simple and multiple regression models.

c) Perform an hypothesis test to evaluate whether the coefficient on *CITY* is significantly greater than zero. Manage the α-risk at 5%. Interpret the results of this test.

d) Test whether the estimated coefficients on *CITY* and *MILES* are jointly equal to zero. Manage the α-risk at 5%. Interpret the results of this test.

4. J. M. Keynes postulated that aggregate real consumption (RCONS) is positively related to aggregate real GNP (RGNP) in such a way that the marginal propensity to consume—the change in consumption resulting from a one-unit change in income—is less than the average propensity to consume—the ratio of consumption to income. There remains the question of whether consumption is negatively related to the rate of interest (or, which is the same thing, savings is positively related to the interest rate). The table on the next page presents some data on consumption, real GNP and interest rates in Canada, along with the LOTUS-123 regression output using these data. A dummy variable is included to test whether consumption depends on whether the exchange rate is fixed or flexible. The column PRED gives the level of consumption predicted by the regression that includes the dummy variable and the column ERROR gives the difference between the actual value of consumption and the value predicted by that regression. SQERR is the error squared and the right-most column gives the error times itself lagged.

WORKSHEET FOR REGRESSION ANALYSIS OF CANADIAN CONSUMPTION

	RCONS	RGNP	INTRATE	DUMMY	PRED	ERROR	SQERR	ERROR TIMES ERROR LAGGED
1961	105.4	161.4	3.37	0	99.7	5.7	32.3	
1962	111.1	173.4	4.38	0	105.7	5.4	29.0	30.629
1963	116.4	182.8	4.01	1	114.1	2.3	5.4	12.530
1964	122.8	196.6	4.20	1	121.9	0.9	0.8	2.023
1965	129.7	211.5	5.01	1	129.8	-0.1	0.0	-0.079
1966	136.8	228.2	6.27	1	138.4	-1.6	2.5	0.143
1967	142.9	236.3	5.84	1	143.5	-0.5	0.3	0.831
1968	150.0	248.4	6.82	1	149.6	0.4	0.2	-0.207
1969	157.1	262.0	7.84	1	156.5	0.5	0.3	0.212
1970	160.6	271.9	7.34	0	160.2	0.4	0.1	0.196
1971	169.3	288.5	4.51	0	172.4	-3.1	9.5	-1.122
1972	181.0	308.0	5.10	0	183.2	-2.2	4.8	6.763
1973	192.4	335.8	7.45	0	197.2	-4.8	22.6	10.447
1974	202.7	361.1	10.50	0	209.1	-6.4	40.6	30.310
1975	211.9	367.5	7.93	0	215.1	-3.3	10.7	20.837
1976	225.3	393.2	9.17	0	228.9	-3.7	13.6	12.036
1977	231.3	399.7	7.47	0	234.2	-2.9	8.4	10.691
1978	236.2	405.4	8.83	0	236.3	-0.1	0.0	0.294
1979	241.5	423.8	12.07	0	244.0	-2.6	6.5	0.259
1980	246.3	432.0	13.15	0	247.9	-1.6	2.5	4.012
1981	249.3	438.3	18.33	0	246.8	2.4	5.9	-3.810
1982	241.4	415.2	14.15	0	237.2	4.2	17.6	10.189
1983	250.8	427.5	9.45	0	248.6	2.3	5.1	9.495
1984	261.8	449.2	11.18	0	259.6	2.3	5.1	5.106
1985	276.1	465.9	9.56	0	270.7	5.3	28.4	12.037
1986	287.1	474.2	9.16	0	275.9	11.2	126.3	59.940
SUM	5037.1	8557.8	213.1	7.0		10.6	378.6	233.8
MEAN	193.7	329.1	8.2	0.3		0.0		
VAR	3164.17	10421.87	12.6228	0.204				

****

Regression Output:

Constant	9.12
R Squared	0.99527
No. of Observations	26

	RGNP	INTRATE	DUMMY
X Coefficient(s)	0.58	-0.90	2.53
Std Err of Coef.	0.02	0.38	2.40

Regression Output: Dummy Variable Excluded:

Constant	12.21
R Squared	0.99504
No. of Observations	26

	RGNP	INTRATE
X Coefficient(s)	0.57	-0.84
Std Err of Coef.	0.01	0.38

a) Can we conclude that consumption is positively related to income?

b) How would you test the proposition that the marginal propensity to consume equals the average propensity to consume?

c) Can we conclude that the interest rate has a negative effect on consumption?

d) Is aggregate consumption affected by whether the country was on fixed as opposed to flexible exchange rates?

e) Test whether the regression that includes all three independent variables is statistically significant.

f) Do an F-test of the proposition that consumption depends on whether the country was on a fixed or flexible exchange rate. Show that the F-statistic so obtained is equal to the square of the relevant t-statistic in the regression that includes the dummy variable.

g) Perform a crude test of whether residuals of the regression are serially correlated.