CHAPTER 8

THE PROBLEM OF CORRELATED ERRORS

8.1 INTRODUCTION: AUTOCORRELATION

One of the standard assumptions in the regression model is that the error terms ε_i and ε_j, associated with the ith and jth observations, are uncorrelated. Correlation in the error terms suggests that there is additional information in the data that has not been exploited in the current model. When the observations have a *natural* sequential order, the correlation is referred to as *autocorrelation*.

Autocorrelation may occur for several reasons. Adjacent residuals tend to be similar in both temporal and spatial dimensions. Successive residuals in economic time series tend to be positively correlated. Large positive errors are followed by other positive errors, and large negative errors are followed by other negative errors. Observations sampled from adjacent experimental plots or areas tend to have residuals that are correlated since they are affected by similar external conditions.

The symptoms of autocorrelation may also appear as the result of a variable having been omitted from the right-hand side of the regression equation. If successive values of the omitted variable are correlated, the errors from the estimated model will appear to be correlated. When the variable is added to the equation, the apparent problem of autocorrelation disappears. The presence of autocorrelation has several effects on the analysis. These are summarized as follows:

Regression Analysis by Example, Fourth Edition. By Samprit Chatterjee and Ali S. Hadi

1. Least squares estimates of the regression coefficients are unbiased but are not efficient in the sense that they no longer have minimum variance.

2. The estimate of σ^2 and the standard errors of the regression coefficients may be seriously understated; that is, from the data the estimated standard errors would be much smaller than they actually are, giving a spurious impression of accuracy.

3. The confidence intervals and the various tests of significance commonly employed would no longer be strictly valid.

The presence of autocorrelation can be a problem of serious concern for the preceding reasons and should not be ignored.

We distinguish between two types of autocorrelation and describe methods for dealing with each. The first type is only autocorrelation in appearance. It is due to the omission of a variable that should be in the model. Once this variable is uncovered, the autocorrelation problem is resolved. The second type of autocorrelation may be referred to as pure autocorrelation. The methods of correcting for pure autocorrelation involve a transformation of the data. Formal derivations of the methods can be found in Johnston (1984) and Kmenta (1986).

8.2 CONSUMER EXPENDITURE AND MONEY STOCK

Table 8.1 gives quarterly data from 1952 to 1956 on consumer expenditure (Y) and the stock of money (X), both measured in billions of current dollars for the United States. The data can be found in the book's Web site.[1]

A simplified version of the quantity theory of money suggests a model given by

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \tag{8.1}$$

where β_0 and β_1 are constants, ε_t the error term. Economists are interested in estimating β_1 and its standard error; β_1 is called the *multiplier* and has crucial importance as an instrument in fiscal and monetary policy. Since the observations are ordered in time, it is reasonable to expect that autocorrelation may be present. A summary of the regression results is given in Table 8.2.

The regression coefficients are significant; the standard error of the slope coefficient is 0.115. For a unit change in the money supply the 95% confidence interval for the change in the aggregate consumer expenditure would be $2.30 \pm 2.10 \times 0.115 = (2.06, 2.54)$. The value of R^2 indicates that roughly 96% of the variation in the consumer expenditure can be accounted for by the variation in money stock. The analysis would be complete if the basic regression assumptions were valid. To check on the model assumption, we examine the residuals. If there are indications that autocorrelation is present, the model should be reestimated after eliminating the autocorrelation.

[1] http://www.ilr.cornell.edu/˜hadi/RABE4

Table 8.1 Consumer Expenditure and Money Stock

Year	Quarter	Consumer Expenditure	Money Stock	Year	Quarter	Consumer Expenditure	Money Stock
1952	1	214.6	159.3	1954	3	238.7	173.9
	2	217.7	161.2		4	243.2	176.1
	3	219.6	162.8	1955	1	249.4	178.0
	4	227.2	164.6		2	254.3	179.1
1953	1	230.9	165.9		3	260.9	180.2
	2	233.3	167.9		4	263.3	181.2
	3	234.1	168.3	1956	1	265.6	181.6
	4	232.3	169.7		2	268.2	182.5
1954	1	233.7	170.5		3	270.4	183.3
	2	236.5	171.6		4	275.6	184.3

Source: Friedman and Meiselman (1963), p. 266.

Table 8.2 Results When Consumer Expenditure Is Regressed on Money Stock, X

Variable	Coefficient	s.e.	t-test	p-value
Constant	−154.72	19.850	−7.79	< 0.0001
X	2.30	0.115	20.10	< 0.0001
$n = 20$	$R^2 = 0.957$	$R_a^2 = 0.955$	$\hat{\sigma} = 3.983$	$d.f. = 18$

For time series data a useful plot for analysis is the index plot (plot of the standardized residuals versus time). The graph is given in Figure 8.1. The pattern of residuals is revealing and is characteristic of situations where the errors are correlated. Residuals of the same sign occur in clusters or bunches. The characteristic pattern would be that several successive residuals are positive, the next several are negative, and so on. From Figure 8.1 we see that the first seven residuals are positive, the next seven negative, and the last six positive. This pattern suggests that the error terms in the model are correlated and some additional analysis is required.

This visual impression can be formally confirmed by counting the number of runs in a plot of the signs of the residuals, the residuals taken in the order of the observations. These types of plots are called *sequence plots*. In our present example the sequence plot of the signs of the residuals is

$$+ + + + + + + - - - - - - - + + + + + +$$

and it indicates three runs. With n_1 residuals positive and n_2 residuals negative, under the hypothesis of randomness the expected number of runs μ and its variance

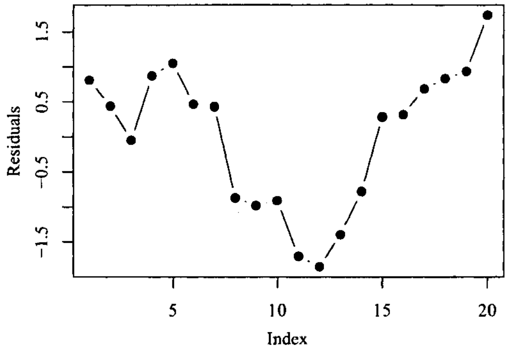

Figure 8.1 Index plot of the standardized residuals.

σ^2 would be

$$\mu = \frac{2n_1n_2}{n_1+n_2}+1,$$

$$\sigma^2 = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}.$$

In our case $n_1 = 13, n_2 = 7$, giving the expected number of runs to be 10.1 and a standard deviation of 1.97. The observed number of runs is three. The deviation of 5.1 from the expected number of runs is more than twice the standard deviation, indicating a significant departure from randomness. This formal *runs test* procedure merely confirms the conclusion arrived at visually that there is a pattern in the residuals.

Many computer packages now have the runs test as an available option. This approximate runs test for confirmation can therefore be easily executed. The runs test as we have described it should not, however, be used for small values of n_1 and n_2 (less than 10). For small values of n_1 and n_2 one needs exact tables of probability to judge significance. For more details on the runs test, the reader should refer to a book on nonparametric statistics such as Lehmann (1975), Gibbons (1993), and Hollander and Wollfe (1999). Besides the graphical analysis, which can be confirmed by the runs test, autocorrelated errors can also be detected by the Durbin-Watson statistic.

8.3 DURBIN-WATSON STATISTIC

The Durbin-Watson statistic is the basis of a popular test of autocorrelation in regression analysis. The test is based on the assumption that successive errors are correlated, namely,

$$\varepsilon_t = \rho\varepsilon_{t-1} + \omega_t, \qquad |\rho| < 1, \tag{8.2}$$

where ρ is the correlation coefficient between ε_t and ε_{t-1}, and ω_t is normally independently distributed with zero mean and constant variance. In this case, the errors are said to have *first-order autoregressive structure* or *first-order autocorrelation*. In most situations the error ε_t may have a much more complex correlation structure. The first-order dependency structure, given in (8.2), is taken as a simple approximation to the actual error structure.

The Durbin-Watson statistic is defined as

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

where e_i is the ith ordinary least squares (OLS) residual. The statistics d is used for testing the null hypothesis $H_0 : \rho = 0$ against an alternative $H_1 : \rho > 0$. Note that when $\rho = 0$ in Equation (8.2), the ε's are uncorrelated.

Since ρ is unknown, we estimate the parameter ρ by $\hat{\rho}$, where

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}. \tag{8.3}$$

An approximate relationship between d and $\hat{\rho}$ is

$$d \doteq 2(1 - \hat{\rho}),$$

($\doteq$ means approximately equal to) showing that d has a range of 0 to 4. Since $\hat{\rho}$ is an estimate of ρ, it is clear that d is close to 2 when $\rho = 0$ and near to zero when $\rho = 1$. The closer the sample value of d to 2, the firmer the evidence that there is no autocorrelation present in the error. Evidence of autocorrelation is indicated by the deviation of d from 2. The formal test for positive autocorrelation operates as follows: Calculate the sample statistic d. Then, if

1. $d < d_L$, reject H_0.
2. $d > d_U$, do not reject H_0.
3. $d_L < d < d_U$, the test is inconclusive.

The values of (d_L, d_U) for different percentage points have been tabulated by Durbin and Watson (1951). A table is provided in the Appendix at the end of the book (Tables A.6 and A.7).

Tests for negative autocorrelation are seldom performed. If, however, a test is desired, then instead of working with d, one works with $(4 - d)$ and follows the same procedure as for the testing of positive autocorrelation.

In our Money Stock and Consumer Expenditure data, the value of d is 0.328. From Table A.6, with $n = 20$, $p = 1$ (the number of predictors), and a significance level of 0.05, we have $d_L = 1.20$ and $d_U = 1.41$. Since $d < d_L$, we conclude that the value of d is significant at the 5% level and H_0 is rejected, showing that autocorrelation is present. This essentially reconfirms our earlier conclusion, which was arrived at by looking at the index plot of the residuals.

If d had been larger than $d_U = 1.41$, autocorrelation would not be a problem and no further analysis is needed. When $d_L < d < d_U$, additional analysis of the equation is optional. We suggest that in cases where the Durbin-Watson statistic lies in the inconclusive region, reestimate the equation using the methods described below to see if any major changes occur.

As pointed out earlier, the presence of correlated errors distorts estimates of standard errors, confidence intervals, and statistical tests, and therefore we should reestimate the equation. When autocorrelated errors are indicated, two approaches may be followed. These are (1) work with transformed variables, or (2) introduce additional variables that have time-ordered effects. We illustrate the first approach with the Money Stock data. The second approach is illustrated in Section 8.6.

8.4 REMOVAL OF AUTOCORRELATION BY TRANSFORMATION

When the residual plots and Durbin-Watson statistic indicate the presence of correlated errors, the estimated regression equation should be refitted taking the autocorrelation into account. One method for adjusting the model is the use of a transformation that involves the unknown autocorrelation parameter, ρ. The introduction of ρ causes the model to be nonlinear. The direct application of least squares is not possible. However, there are a number of procedures that may be used to circumvent the nonlinearity (Johnston, 1984). We use the method due to Cochrane and Orcutt (1949).

From model (8.1), ε_t and ε_{t-1} can be expressed as

$$\begin{aligned} \varepsilon_t &= y_t - \beta_0 - \beta_1 x_t, \\ \varepsilon_{t-1} &= y_{t-1} - \beta_0 - \beta_1 x_{t-1}. \end{aligned}$$

Substituting these in (8.2), we obtain

$$y_t - \beta_0 - \beta_1 x_t = \rho(y_{t-1} - \beta_0 - \beta_1 x_{t-1}) + \omega_t.$$

Rearranging terms in the above equation, we get

$$\begin{array}{ccccccc} y_t - \rho y_{t-1} & = & \beta_0(1-\rho) & + & \beta_1(x_t - \rho x_{t-1}) & + & \omega_t, \\ y_t^* & = & \beta_0^* & + & \beta_1^* \quad x_t^* & + & \omega_t, \end{array} \tag{8.4}$$

where

$$\begin{aligned} y_t^* &= y_t - \rho y_{t-1}, \\ x_t^* &= x_t - \rho x_{t-1}, \\ \beta_0^* &= \beta_0(1-\rho), \\ \beta_1^* &= \beta_1. \end{aligned}$$

Since the ω's are uncorrelated, Equation (8.4) represents a linear model with uncorrelated errors. This suggests that we run an ordinary least squares regression using

y_t^* as a response variable and x_t^* as a predictor. The estimates of the parameters in the original equations are

$$\hat{\beta}_0 = \frac{\hat{\beta}_0^*}{1 - \hat{\rho}} \quad \text{and} \quad \hat{\beta}_1 = \hat{\beta}_1^*. \tag{8.5}$$

Therefore, when the errors in model (8.1) have an autoregressive structure as given in (8.2), we can transform both sides of the equation and obtain transformed variables which satisfy the assumption of uncorrelated errors.

The value of ρ is unknown and has to be estimated from the data. Cochrane and Orcutt (1949) have proposed an iterative procedure. The procedure operates as follows:

1. Compute the OLS estimates of β_0 and β_1 by fitting model (8.1) to the data.

2. Calculate the residuals and, from the residuals, estimate ρ using (8.3).

3. Fit the equation given in (8.4) using the variables $y_t - \hat{\rho} y_{t-1}$ and $x_t - \hat{\rho} x_{t-1}$ as response and predictor variables, respectively, and obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ using (8.5).

4. Examine the residuals of the newly fitted equation. If the new residuals continue to show autocorrelation, repeat the entire procedure using the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ as estimates of β_0 and β_1 instead of the original least squares estimates. On the other hand, if the new residuals show no autocorrelation, the procedure is terminated and the fitted equation for the original data is:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t.$$

As a practical rule we suggest that if the first application of Cochrane-Orcutt procedure does not yield non-autocorrelated residuals, one should look for alternative methods of removing autocorrelation. We apply the Cochrane-Orcutt procedure to the data given in Table 8.1.

The d value for the original data is 0.328, which is highly significant. The value of $\hat{\rho}$ is 0.751. On fitting the regression equation to the variables $(y_t - 0.751 y_{t-1})$ and $(x_t - 0.751 x_{t-1})$, we have a d value of 1.43. The value of d_U for $n = 19$ and $p = 1$ is 1.40 at the 5% level. Consequently, $H_0 : \rho = 0$ is not rejected.[2] The fitted equation is

$$\hat{y}_t^* = -53.70 + 2.64 x_t^*,$$

which, using (8.5), the fitted equation in terms of the original variables is

$$\hat{y}_t = -215.31 + 2.64 x_t.$$

[2] The significance level of the test is not exact because $\hat{\rho}$ was used in the estimation process. The d value of 1.43 may be viewed as an index of autocorrelation that indicates an improvement from the previous value of 0.328.

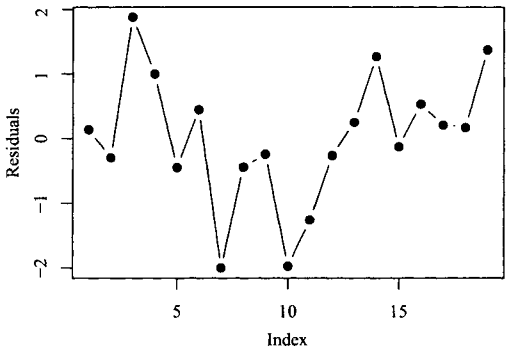

Figure 8.2 Index plot of standardized residuals after one iteration of the Cochrane-Orcutt method.

The estimated standard error for the slope is 0.307, as opposed to the least squares estimate of the original equation, which was $y_t = -154.7 + 2.3x_t$ with a standard error for the slope of 0.115. The newly estimated standard error is larger by a factor of almost 3. The residual plots for the fitted equation of the transformed variables are shown in Figure 8.2. The residual plots show less clustering of the adjacent residuals by sign, and the Cochrane-Orcutt procedure has worked to our advantage.

8.5 ITERATIVE ESTIMATION WITH AUTOCORRELATED ERRORS

One advantage of the Cochrane-Orcutt procedure is that estimates of the parameters are obtained using standard least squares computations. Although two stages are required, the procedure is relatively simple. A more direct approach is to try to estimate values of ρ, β_0, and β_1 simultaneously. The model is formulated as before requiring the construction of transformed variables $y_t - \rho y_{t-1}$ and $x_t - \rho x_{t-1}$. Parameter estimates are obtained by minimizing the sum of squared errors, which is given as

$$S(\beta_0, \beta_1, \rho) = \sum_{t=2}^{n} [y_t - \rho y_{t-1} - \beta_0(1-\rho) - \beta_1(x_t - \rho x_{t-1})]^2.$$

If the value of ρ were known, β_0 and β_1 would be easily obtained by regressing $y_t - \rho y_{t-1}$ on $x_t - \rho x_{t-1}$. Final estimates are obtained by searching through many values of ρ until a combination of ρ, β_0 and β_1 is found that minimizes $S(\rho, \beta_0, \beta_1)$. The search could be accomplished using a standard regression computer program, but the process can be much more efficient with an automated search procedure. This method is due to Hildreth and Lu (1960). For a discussion of the estimation procedure and properties of the estimates obtained, see Kmenta (1986).

Once the minimizing values, say $\tilde{\rho}$, $\tilde{\beta}_0$, and $\tilde{\beta}_1$, have been obtained, the standard error for the estimate of β_1 can be approximated using a version of Equation (2.25)

Table 8.3 Comparison of Regression Estimates

Method	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	s.e.$(\hat{\beta}_1)$
OLS	–	−154.700	2.300	0.115
Cochrane-Orcutt	0.874	−324.440	2.758	0.444
Iterative	0.824	−235.509	2.753	0.436

of Chapter 2. The formula is used as though $y_t - \rho y_{t-1}$ were regressed on $x_t - \rho x_{t-1}$ with ρ known; that is, the estimated standard error of $\tilde{\beta}_1$ is

$$\text{s.e.}(\tilde{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum[x_t - \tilde{\rho}x_{t-1} - \bar{x}(1-\tilde{\rho})]^2}},$$

where $\hat{\sigma}$ is the square root of $S(\tilde{\rho}, \tilde{\beta}_0, \tilde{\beta}_1)/(n-2)$. When adequate computing facilities are available such that the iterative computations are easy to accomplish, then the latter method is recommended. However, it is not expected that the estimates and standard errors for the iterative method and the two-stage Cochrane-Orcutt method would be appreciably different. The estimates from the three methods, OLS, Cochrane-Orcutt, and iterative for the data of Table 8.1, are given in Table 8.3 for comparison.

8.6 AUTOCORRELATION AND MISSING VARIABLES

The characteristics of the regression residuals that suggest autocorrelation may also be indicative of other aspects of faulty model specification. In the preceding example, the index plot of residuals and the statistical test based on the Durbin-Watson statistic were used to conclude that the residuals are autocorrelated. Autocorrelation is only one of a number of possible explanations for the clustered type of residual plot or low Durbin-Watson value.

In general, a plot of residuals versus any one of the list of potential predictor variables may uncover additional information that can be used to further explain variation in the response variable. When an index plot of residuals shows a pattern of the type described in the preceding example, it is reasonable to suspect that it may be due to the omission of variables that change over time. Certainly, when the residuals appear in clusters alternating above and below the mean value line of zero, when the estimated autocorrelation coefficient is large and the Durbin-Watson statistic is significant, it would appear that the presence of autocorrelation is overwhelmingly supported. We shall see that this conclusion may be incorrect. The observed symptoms would be better interpreted initially as a general indication of some form of model misspecification.

All possible correction procedures should be considered. In fact, it is always better to explore fully the possibility of some additional predictor variables before yielding to an autoregressive model for the error structure. It is more satisfying and probably more useful to be able to understand the source of apparent autocorrelation in terms of an additional variable. The marginal effect of that variable can then be estimated and used in an informative way. The transformations that correct for pure autocorrelation may be viewed as an action of last resort.

8.7 ANALYSIS OF HOUSING STARTS

As an example of a situation where autocorrelation appears artificially because of the omission of another predictor variable, consider the following project undertaken by a midwestern construction industry association. The association wants to have a better understanding of the relationship between housing starts and population growth. They are interested in being able to forecast construction activity. Their approach is to develop annual data on regional housing starts and try to relate these data to potential home buyers in the region. Realizing that it is almost impossible to measure the number of potential house buyers accurately, the researchers settled for the size of the 22- to 44-year-old population group in the region as a variable that reflects the size of potential home buyers. With some diligent work they were able to bring together 25 years of historical data for the region (see Table 8.4). The data in Table 8.4 can be obtained from the book's Web site. Their goal was to get a simple regression relationship between housing starts and population,

$$H_t = \beta_0 + \beta_1 P_t + \varepsilon_t. \tag{8.6}$$

Then using methods that they developed for projecting population changes, they would be able to estimate corresponding changes in the requirements for new houses. The construction association was aware that the relationship between population and housing starts could be very complex. It is even reasonable to suggest that housing affects population growth (by migration) instead of the other way around. Although the proposed model is undoubtedly naive, it serves a useful purpose as a starting point for their analysis.

Analysis

The regression results from fitting model (8.6) to the 25 years of data are given in Table 8.5. The proportion of variation in H accounted for by the variability in P is $R^2 = 0.925$. We also see that an increase in population of 1 million leads to an increase in housing starts of about 71,000. The Durbin-Watson statistic and the index plot of the residuals (Figure 8.3) suggest strong autocorrelation. However, it is fairly simple to conjecture about other variables that may further explain housing starts and could be responsible for the appearance of autocorrelation. These variables include the unemployment rate, social trends in marriage and family formation, government programs in housing, and the availability of construction and mortgage funds. The first choice was an index that measures the availability of mortgage money for the region. Adding that variable to the equation the model

Table 8.4 Data for Housing Starts (H), Population Size (P) in millions, and Availability for Mortgage Money Index (D)

Row	H	P	D
1	0.09090	2.200	0.03635
2	0.08942	2.222	0.03345
3	0.09755	2.244	0.03870
4	0.09550	2.267	0.03745
5	0.09678	2.280	0.04063
6	0.10327	2.289	0.04237
7	0.10513	2.289	0.04715
8	0.10840	2.290	0.04883
9	0.10822	2.299	0.04836
10	0.10741	2.300	0.05160
11	0.10751	2.300	0.04879
12	0.11429	2.340	0.05523
13	0.11048	2.386	0.04770
14	0.11604	2.433	0.05282
15	0.11688	2.482	0.05473
16	0.12044	2.532	0.05531
17	0.12125	2.580	0.05898
18	0.12080	2.605	0.06267
19	0.12368	2.631	0.05462
20	0.12679	2.658	0.05672
21	0.12996	2.684	0.06674
22	0.13445	2.711	0.06451
23	0.13325	2.738	0.06313
24	0.13863	2.766	0.06573
25	0.13964	2.793	0.07229

becomes

$$H_t = \beta_0 + \beta_1 P_t + \beta_2 D_t + \varepsilon_t.$$

The introduction of the additional variable has the effect of removing autocorrelation. From Table 8.6 we see that the Durbin-Watson statistic has the new value 1.852, well into the acceptable region. The index plot of the residuals (Figure 8.4) is also improved. The regression coefficients and their corresponding t-values show that there is a significant population effect but that it was overstated by a factor of more than 2 in the first equation. In a certain sense, the effect of changes in the availability of mortgage money for a fixed level of population is more important than a similar change in population.

If each variable in the regression equation is replaced by the standardized version of the variable (the variables transformed so as to have mean 0, and unit variance),

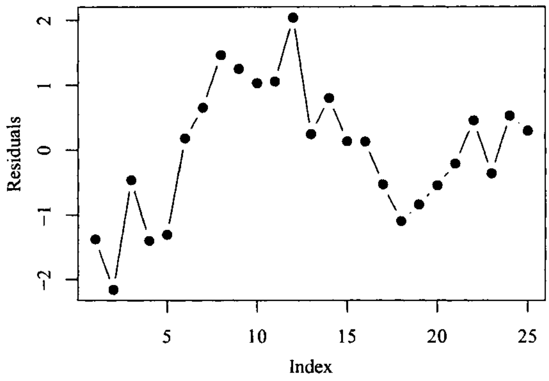

Figure 8.3 Index plot of standardized residuals from the regression of H_t on P_t for the Housing Starts data.

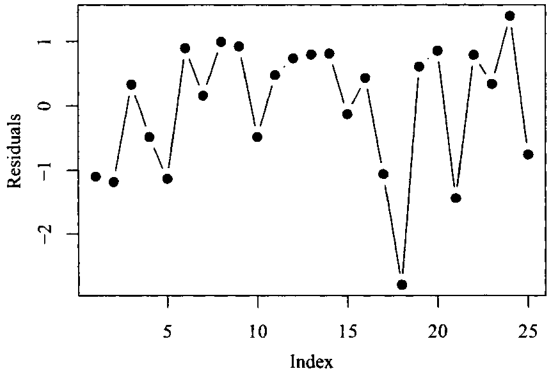

Figure 8.4 Index plot of the standardized residuals from the regression of H_t on P_t and D_t for the Housing Starts data.

Table 8.5 Regression on Housing Starts (H) Versus Population (P)

Variable	Coefficient	s.e.	t-test	p-value
Constant	−0.0609	0.0104	−5.85	< 0.0001
P	0.0714	0.0042	16.90	< 0.0001
$n = 25$	$R^2 = 0.925$	$d = 0.621$	$\hat{\sigma} = 0.0041$	$d.f. = 23$

Table 8.6 Results of the Regression of Housing Starts (H) on Population (P) and Index (D)

Variable	Coefficient	s.e.	t-test	p-value
Constant	−0.0104	0.0103	−1.01	0.3220
P	0.0347	0.0064	5.39	< 0.0001
D	0.7605	0.1216	6.25	< 0.0001
$n = 25$	$R^2 = 0.973$	$d = 1.85$	$\hat{\sigma} = 0.0025$	$d.f. = 22$

the resulting regression equation is

$$\tilde{H}_t = 0.4668\tilde{P}_t + 0.5413\tilde{D}_t \ ,$$

where $\tilde{H}$ denotes the standardized value of H, $\tilde{H} = (H - \bar{H})/s_H$. A unit increase in the standardized value of $\tilde{P}_t$ is worth an additional 0.4668 to the standardized value of H_t; that is, if the population increases by standard deviation then H_t increases by 0.4668 standard deviation. Similarly, if D_t increases by 1 standard deviation H_t increases by 0.5413 standard deviation. Therefore, in terms of the standardized variables, the mortgage index is more important (has a larger effect) than population size.

The example on housing starts illustrates two important points. First, a large value of R^2 does not imply that the data have been fitted and explained well. Any pair of variables that show trends over time are usually highly correlated. A large value of R^2 does not necessarily confirm that the relationship between the two variables has been adequately characterized. Second, the Durbin-Watson statistic as well as the residual plots may indicate the presence of autocorrelation among the errors when, in fact, the errors are independent but the omission of a variable or variables has given rise to the observed situation. Even though the Durbin-Watson statistic was designed to detect first-order autocorrelation it can have a significant value when some other model assumptions are violated such as misspecification of the variables to be included in the model. In general, a significant value of the Durbin-Watson statistic should be interpreted as an indication that a problem exists, and both the possibility of a missing variable or the presence of autocorrelation should be considered.

Table 8.7 Ski Sales Versus PDI

Variable	Coefficient	s.e.	t-test	p-value
Constant	12.3921	2.539	4.88	< 0.0001
PDI	0.1979	0.016	12.40	< 0.0001
$n = 40$	$R^2 = 0.801$	$d = 1.968$	$\hat{\sigma} = 3.019$	$d.f. = 38$

8.8 LIMITATIONS OF DURBIN-WATSON STATISTIC

In the previous examples on Expenditure versus Money Stock and Housing Starts versus Population Size the residuals from the initial regression equations indicated model misspecifications associated with time dependence. In both cases the Durbin-Watson statistic was small enough to conclude that positive autocorrelation was present. The index plot of residuals further confirmed the presence of a time-dependent error term. In each of the two problems the presence of autocorrelation was dealt with differently. In one case (Housing Starts) an additional variable was uncovered that had been responsible for the appearance of autocorrelation, and in the other case (Money Stock) the Cochrane-Orcutt method was used to deal with what was perceived as pure autocorrelation. It should be noted that the time dependence observed in the residuals in both cases is a first-order type of dependence. Both the Durbin-Watson statistic and the pattern of residuals indicate dependence between residuals in adjacent time periods. If the pattern of time dependence is other than first order, the plot of residuals will still be informative. However, the Durbin-Watson statistic is not designed to measure higher-order time dependence and may not yield much valuable information.

As an example we consider the efforts of a company that produces and markets ski equipment in the United States to obtain a simple aggregate relationship of quarterly sales to some leading economic indicator. The indicator chosen is personal disposable income, PDI, in billions of current dollars. The initial model is

$$S_t = \beta_0 + \beta_1 \text{PDI}_t + \varepsilon_t,$$

where S_t is ski sales in period t in millions of dollars and PDI_t is the personal disposable income for the same period. Data for 10 years (40 quarters) are available (Table 5.11). The data can be obtained from the book's Web site. The regression output is in Table 8.7 and the index plot of residuals is given in Figure 8.6.

At first glance the results in Table 8.7 are encouraging. The proportion of variation in sales accounted for by PDI is 0.80. The marginal contribution of an additional dollar unit of PDI to sales is between \$165,420 and \$230,380 ($\hat{\beta}_1 = 0.1979$) with a confidence coefficient of 95%. In addition, the Durbin-Watson statistic is 1.968, indicating no first-order autocorrelation.

It should be expected that PDI would explain a large proportion of the variation in sales since both variables are increasing over time. Therefore, although the R^2

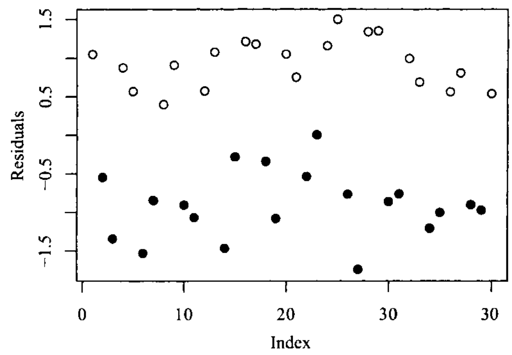

Figure 8.5 Index plot of the standardized residuals. (Quarters 1 and 4 are indicated by an open circle and Quarters 2 and 3 are indicated by a solid circle.)

value of 0.80 is good, it should not be taken as a final evaluation of the model. Also, the Durbin-Watson value is in the acceptable range, but it is clear from Figure 8.5 that there is some sort of time dependence of the residuals. We notice that residuals from the first and fourth quarters are positive, while residuals from the second and third quarters are negative for all the years. Since skiing activities are affected by weather conditions, we suspect that a seasonal effect has been overlooked. The pattern of residuals suggests that there are two seasons that have some bearing on ski sales: the second and third quarters, which correspond to the warm weather season, and the fourth and first quarters, which correspond to the winter season, when skiing is in full progress. This seasonal effect can be simply characterized by defining an indicator (dummy) variable that takes the value 1 for each winter quarter and is set equal to zero for each summer quarter (see Chapter 5). The expanded data set is listed in Table 8.8 and can be obtained from the book's Web site.

8.9 INDICATOR VARIABLES TO REMOVE SEASONALITY

Using the additional seasonal variable, the model is expanded to be

$$S_t = \beta_0 + \beta_1 \mathrm{PDI}_t + \beta_2 Z_t + \varepsilon_t, \qquad (8.7)$$

where Z_t is the zero-one variable described above and β_2 is a parameter that measures the seasonal effect. Note that the model in (8.7) can be represented by the two models (one for the cold weather quarters where $Z_t = 1$) and the other for the warm quarters where $Z_t = 0$):

$$\begin{array}{llcccccc} \text{Winter season :} & S_t & = & (\beta_0 + \beta_2) & + & \beta_1 \,\mathrm{PDI}_t & + & \varepsilon_t, \\ \text{Summer season :} & S_t & = & \beta_0 & + & \beta_1 \,\mathrm{PDI}_t & + & \varepsilon_t. \end{array}$$

Thus, the model represents the assumption that sales can be approximated by a linear function of PDI, in one line for the winter season and one for the summer

Table 8.8 Disposable Income and Ski Sales, and Seasonal Variables for Years 1964–1973

Quarter	Sales	PDI	Season
Q1/64	37.0	109	1
Q2/64	33.5	115	0
Q3/64	30.8	113	0
Q4/64	37.9	116	1
Q1/65	37.4	118	1
Q2/65	31.6	120	0
Q3/65	34.0	122	0
Q4/65	38.1	124	1
Q1/66	40.0	126	1
Q2/66	35.0	128	0
Q3/66	34.9	130	0
Q4/66	40.2	132	1
Q1/67	41.9	133	1
Q2/67	34.7	135	0
Q3/67	38.8	138	0
Q4/67	43.7	140	1
Q1/68	44.2	143	1
Q2/68	40.4	147	0
Q3/68	38.4	148	0
Q4/68	45.4	151	1
Q1/69	44.9	153	1
Q2/69	41.6	156	0
Q3/69	44.0	160	0
Q4/69	48.1	163	1
Q1/70	49.7	166	1
Q2/70	43.9	171	0
Q3/70	41.6	174	0
Q4/70	51.0	175	1
Q1/71	52.0	180	1
Q2/71	46.2	184	0
Q3/71	47.1	187	0
Q4/71	52.7	189	1
Q1/72	52.2	191	1
Q2/72	47.0	193	0
Q3/72	47.8	194	0
Q4/72	52.8	196	1
Q1/73	54.1	199	1
Q2/73	49.5	201	0
Q3/73	49.5	202	0
Q4/73	54.3	204	1

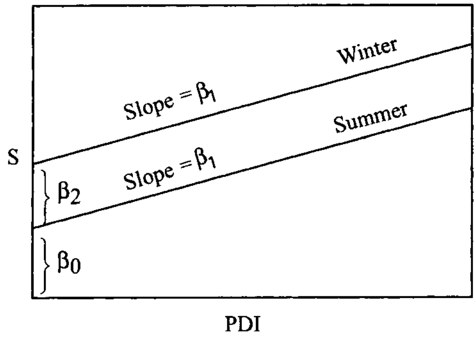

Figure 8.6 Model for Ski Sales and PDI adjusted for season.

Table 8.9 Ski Sales Versus PDI and Seasonal Variables

Variable	Coefficient	s.e.	t-test	p-value
Constant	9.5402	0.9748	9.79	0.3220
PDI	0.1987	0.0060	32.90	< 0.0001
Z	5.4643	0.3597	15.20	< 0.0001
$n = 40$	$R^2 = 0.972$	$d = 1.772$	$\hat{\sigma} = 1.137$	$d.f. = 37$

season. The lines are parallel; that is, the marginal effect of changes in PDI is the same in both seasons. The level of sales, as reflected by the intercept, is different in each season (Figure 8.6).

The regression results are summarized in Table 8.9 and the index plot of the standardized residuals is shown in Figure 8.7. We see that all indications of the seasonal pattern have been removed. Furthermore, the precision of the estimated marginal effect of PDI increased. The confidence interval is now $186,520 to $210,880. Also, the seasonal effect has been quantified and we can say that for a fixed level of PDI the winter season brings between $4,734,109 and $6,194,491 over the summer season (with 95% confidence).

The ski data illustrate two important points concerning autocorrelation. First, the Durbin-Watson statistic is only sensitive to correlated errors when the correlation occurs between adjacent observations (first-order autocorrelation). In the ski data the first-order correlation is −0.001. The second-, fourth-, sixth-, and eighth-order correlations are −0.81, 0.76, −0.71, and 0.73, respectively. The Durbin-Watson test does not show significance in this case. There are other tests that may be used for the detection of higher-order autocorrelations (see Box and Pierce (1970)). But in all cases, the graph of residuals will show the presence of time dependence in the error term when it exists.

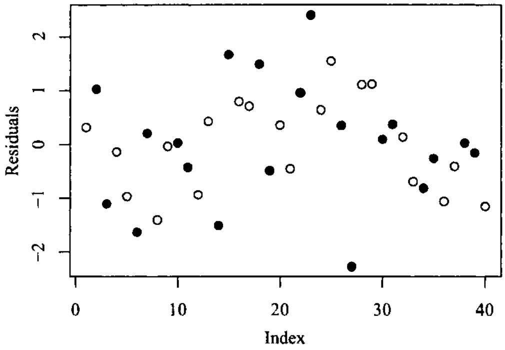

Figure 8.7 Index plot of the standardized residuals with seasonal variables (quarters indicated). (Quarters 1 and 4 are indicated by an open circle and Quarters 2 and 3 are indicated by a solid circle.)

Second, when autocorrelation is indicated the model should be refitted. Often the autocorrelation appears because a time-dependent variable is missing from the model. The inclusion of the omitted variable often removes the observed autocorrelation. Sometimes, however, no such variable is present. Then one has to make a differencing type of transformation on the original variables to remove the autocorrelation.

If the observations are not ordered in time, the Durbin-Watson statistic is not strictly relevant. The statistic may still, however, be a useful diagnostic tool. If the data are ordered by an extraneous criterion, for example, an alphabetic listing, the value of the Durbin-Watson statistic should be near 2.0. Small values are suspicious, and the data should be scrutinized very carefully.

Many data sets are ordered on a criterion that may be relevant to the study. A list of cities or companies may be ordered by size. A low value of the Durbin-Watson statistic would indicate the presence of a significant size effect. A measure of size should therefore be included as a predictor variable. Differencing or Cochrane-Orcutt type of differencing would not be appropriate under these conditions.

8.10 REGRESSING TWO TIME SERIES

The data sets analyzed in this chapter all have the common characteristic that they are time series data (i.e., the observations arise in successive periods of time). This is quite unlike the data sets studied in previous chapters (a notable exception being the bacteria data in Chapter 6), where all the observations are generated at the same point in time. The observations in these examples were contemporaneous and gave rise to *cross-sectional data*. When the observations are generated simultaneously (and relate to a single time period), we have cross-sectional data. The contrast between time series and cross-sectional data can be seen by comparing the ski sales data discussed in this chapter (data arising sequentially in time), and the supervisor performance data in Section 3.3, where all the data were collected in an attitude survey and relate to one historical point in time.

Regression analysis of one time series on another is performed extensively in economics, business, public health, and other social sciences. There are some special features in time series data that are not present in cross-sectional data. We draw attention to these features and suggest possible techniques for handling them.

The concept of autocorrelation is not relevant in cross-sectional data. The ordering of the observations is often arbitrary. Consequently, the correlation of adjacent residuals is an artifact of the organization of the data. For time series data, however, autocorrelation is often a significant factor. The presence of autocorrelation shows that there are hidden structures in the data (often time related) which have not been detected. In addition, most time series data exhibit seasonality, and an investigator should look for seasonal patterns. A regular time pattern in the residuals (as in the ski data) will often indicate the presence of seasonality. For quarterly or monthly data, introduction of indicator variables, as has been pointed out, is a satisfactory solution. For quarterly data, four indicator variables would be needed but only three used in the analysis (see the discussion in Chapter 4). For monthly data, we will need 12 indicator variables but use only 11, to avoid problems of collinearity (this is discussed in Chapter 5). Not all of the indicator variables will be significant and some of them may well be deleted in the final stages of the analysis.

In attempting to find a relationship between y_t and $x_{1t}, x_{2t}, \ldots, x_{pt}$ one may expand the set of predictor variables by including lagged values of the predictor variables. A model such as

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{1t-1} + \beta_3 x_{2t} + \varepsilon_t$$

is meaningful in an analysis of time series data but not with cross-sectional data. The model given above implies that the value of Y in a given period is affected not only by the values of X_1 and X_2 of that period but also by the value of X_1 in the preceding period (i.e., there is a lingering effect of X_1 on Y for one period). Variables lagged by more than one period are also possibilities and could be included in the set of predictor variables.

Time series data are also likely to contain trends. Data in which time trends are likely to occur are often analyzed by including variables that are direct functions of time (t). Variables such as t and t^2 are included in the list of predictor variables. They are used to account for possible linear or quadratic trend. Simple first differencing $(y_t - y_{t-1})$, or more complex lagging of the type $(y_t - ay_{t-1})$ as in the Cochrane-Orcutt procedure, are also possibilities. For a fuller discussion, the reader should consult a book on time series analysis such as Shumway (1988), Hamilton (1994).

To summarize, when performing regression analysis with time series data the analyst should be watchful for autocorrelation and seasonal effects, which are often present in the data. The possibility of using lagged predictor variables should also be explored.

EXERCISES

8.1 Fit model (8.6) to the data in Table 8.4.

(a) Compute the Durbin-Watson statistic *d*. What conclusion regarding the presence of autocorrelation would you draw from *d*?

(b) Compare the number of runs to their expected value and standard deviation when fitting model (8.6) to the data in Table 8.4. What conclusion regarding the presence of autocorrelation would you draw from this comparison?

8.2 Oil Production Data: Refer to the oil production data in Table 6.17. The index plot of the residuals obtained after fitting a linear regression of log(OIL) on Year show a clear cyclical pattern.

(a) Compute the Durbin-Watson statistic *d*. What conclusion regarding the presence of autocorrelation would you draw from *d*?

(b) Compare the number of runs to their expected value and standard deviation. What conclusion regarding the presence of autocorrelation would you draw from this comparison?

8.3 Refer to the Presidential Election Data in Table 5.17. Since the data come over time (for 1916–1996 election years), one might suspect the presence of the autocorrelation problem when fitting the model in (5.11) to the data.

(a) Do you agree? Explain.

(b) Would adding a time trend (e.g., year) as an additional predictor variable improve or exacerbate the autocorrelation? Explain.

8.4 Dow Jones Industrial Average (DJIA): Tables 8.10 and 8.11 contain the values of the daily DJIA for all the trading days in 1996. The data can be found in the book's Web site.[3] DJIA is a very popular financial index and is meant to reflect the level of stock prices in the New York Stock Exchange. The Index is composed of 30 stocks. The variable Day denotes the trading day of the year. There were 262 trading days in 1996, and as such the variable Day goes from 1 to 262.

(a) Fit a linear regression model connecting DJIA with Day using all 262 trading days in 1996. Is the linear trend model adequate? Examine the residuals for time dependencies.

(b) Regress $\text{DJIA}_{(t)}$ against $\text{DJIA}_{(t-1)}$, that is, regress DJIA against its own value lagged by one period. Is this an adequate model? Are there any evidences of autocorrelation in the residuals?

(c) The variability (volatility) of the daily DJIA is large, and to accommodate this phenomenon the analysis is carried out on the logarithm of DJIA.

[3] http://www.ilr.cornell.edu/~hadi/RABE4

Table 8.10 DJIA Data for the First Six Months of 1996

Day	Date	DJIA	Day	Date	DJIA	Day	Date	DJIA
1	1/1/96	5117.12	45	3/1/96	5536.56	89	5/2/96	5498.27
2	1/2/96	5177.45	46	3/4/96	5600.15	90	5/3/96	5478.03
3	1/3/96	5194.07	47	3/5/96	5642.42	91	5/6/96	5464.31
4	1/4/96	5173.84	48	3/6/96	5629.77	92	5/7/96	5420.95
5	1/5/96	5181.43	49	3/7/96	5641.69	93	5/8/96	5474.06
6	1/8/96	5197.68	50	3/8/96	5470.45	94	5/9/96	5475.14
7	1/9/96	5130.13	51	3/11/96	5581.00	95	5/10/96	5518.14
8	1/10/96	5032.94	52	3/12/96	5583.89	96	5/13/96	5582.60
9	1/11/96	5065.10	53	3/13/96	5568.72	97	5/14/96	5624.71
10	1/12/96	5061.12	54	3/14/96	5586.06	98	5/15/96	5625.44
11	1/15/96	5043.78	55	3/15/96	5584.97	99	5/16/96	5635.05
12	1/16/96	5088.22	56	3/18/96	5683.60	100	5/17/96	5687.50
13	1/17/96	5066.90	57	3/19/96	5669.51	101	5/20/96	5748.82
14	1/18/96	5124.35	58	3/20/96	5655.42	102	5/21/96	5736.26
15	1/19/96	5184.68	59	3/21/96	5626.88	103	5/22/96	5778.00
16	1/22/96	5219.36	60	3/22/96	5636.64	104	5/23/96	5762.12
17	1/23/96	5192.27	61	3/25/96	5643.86	105	5/24/96	5762.86
18	1/24/96	5242.84	62	3/26/96	5670.60	106	5/27/96	5762.86
19	1/25/96	5216.83	63	3/27/96	5626.88	107	5/28/96	5709.67
20	1/26/96	5271.75	64	3/28/96	5630.85	108	5/29/96	5673.83
21	1/29/96	5304.98	65	3/29/96	5587.14	109	5/30/96	5693.41
22	1/30/96	5381.21	66	4/1/96	5637.72	110	5/31/96	5643.18
23	1/31/96	5395.30	67	4/2/96	5671.68	111	6/3/96	5624.71
24	2/1/96	5405.06	68	4/3/96	5689.74	112	6/4/96	5665.71
25	2/2/96	5373.99	69	4/4/96	5682.88	113	6/5/96	5697.48
26	2/5/96	5407.59	70	4/5/96	5682.88	114	6/6/96	5667.19
27	2/6/96	5459.61	71	4/8/96	5594.37	115	6/7/96	5697.11
28	2/7/96	5492.12	72	4/9/96	5560.41	116	6/10/96	5687.87
29	2/8/96	5539.45	73	4/10/96	5485.98	117	6/11/96	5668.66
30	2/9/96	5541.62	74	4/11/96	5487.07	118	6/12/96	5668.29
31	2/12/96	5600.15	75	4/12/96	5532.59	119	6/13/96	5657.95
32	2/13/96	5601.23	76	4/15/96	5592.92	120	6/14/96	5649.45
33	2/14/96	5579.55	77	4/16/96	5620.02	121	6/17/96	5652.78
34	2/15/96	5551.37	78	4/17/96	5549.93	122	6/18/96	5628.03
35	2/16/96	5503.32	79	4/18/96	5551.74	123	6/19/96	5648.35
36	2/19/96	5503.32	80	4/19/96	5535.48	124	6/20/96	5659.43
37	2/20/96	5458.53	81	4/22/96	5564.74	125	6/21/96	5705.23
38	2/21/96	5515.97	82	4/23/96	5588.59	126	6/24/96	5717.79
39	2/22/96	5608.46	83	4/24/96	5553.90	127	6/25/96	5719.27
40	2/23/96	5630.49	84	4/25/96	5566.91	128	6/26/96	5682.70
41	2/26/96	5565.10	85	4/26/96	5567.99	129	6/27/96	5677.53
42	2/27/96	5549.21	86	4/29/96	5573.41	130	6/28/96	5654.63
43	2/28/96	5506.21	87	4/30/96	5569.08			
44	2/29/96	5485.62	88	5/1/96	5575.22			

Table 8.11 DJIA Data for the Second Six Months of 1996

Day	Date	DJIA	Day	Date	DJIA	Day	Date	DJIA
131	7/1/96	5729.98	175	8/30/96	5616.21	219	10/31/96	6029.38
132	7/2/96	5720.38	176	9/2/96	5616.21	220	11/1/96	6021.93
133	7/3/96	5703.02	177	9/3/96	5648.39	221	11/4/96	6041.68
134	7/4/96	5703.02	178	9/4/96	5656.90	222	11/5/96	6081.18
135	7/5/96	5588.14	179	9/5/96	5606.96	223	11/6/96	6177.71
136	7/8/96	5550.83	180	9/6/96	5659.86	224	11/7/96	6206.04
137	7/9/96	5581.86	181	9/9/96	5733.84	225	11/8/96	6219.82
138	7/10/96	5603.65	182	9/10/96	5727.18	226	11/11/96	6255.60
139	7/11/96	5520.50	183	9/11/96	5754.92	227	11/12/96	6266.04
140	7/12/96	5510.56	184	9/12/96	5771.94	228	11/13/96	6274.24
141	7/15/96	5349.51	185	9/13/96	5838.52	229	11/14/96	6313.00
142	7/16/96	5358.76	186	9/16/96	5889.20	230	11/15/96	6348.03
143	7/17/96	5376.88	187	9/17/96	5888.83	231	11/18/96	6346.91
144	7/18/96	5464.18	188	9/18/96	5877.36	232	11/19/96	6397.60
145	7/19/96	5426.82	189	9/19/96	5867.74	233	11/20/96	6430.02
146	7/22/96	5390.94	190	9/20/96	5888.46	234	11/21/96	6418.47
147	7/23/96	5346.55	191	9/23/96	5894.74	235	11/22/96	6471.76
148	7/24/96	5354.69	192	9/24/96	5874.03	236	11/25/96	6547.79
149	7/25/96	5422.01	193	9/25/96	5877.36	237	11/26/96	6528.41
150	7/26/96	5473.06	194	9/26/96	5868.85	238	11/27/96	6499.34
151	7/29/96	5434.59	195	9/27/96	5872.92	239	11/28/96	6499.34
152	7/30/96	5481.93	196	9/30/96	5882.17	240	11/29/96	6521.70
153	7/31/96	5528.91	197	10/1/96	5904.90	241	12/2/96	6521.70
154	8/1/96	5594.75	198	10/2/96	5933.97	242	12/3/96	6442.69
155	8/2/96	5679.83	199	10/3/96	5932.85	243	12/4/96	6422.94
156	8/5/96	5674.28	200	10/4/96	5992.86	244	12/5/96	6437.10
157	8/6/96	5696.11	201	10/7/96	5979.81	245	12/6/96	6381.94
158	8/7/96	5718.67	202	10/8/96	5966.77	246	12/9/96	6463.94
159	8/8/96	5713.49	203	10/9/96	5930.62	247	12/10/96	6473.25
160	8/9/96	5681.31	204	10/10/96	5921.67	248	12/11/96	6402.52
161	8/12/96	5704.98	205	10/11/96	5969.38	249	12/12/96	6303.71
162	8/13/96	5647.28	206	10/14/96	6010.00	250	12/13/96	6304.87
163	8/14/96	5666.88	207	10/15/96	6004.78	251	12/16/96	6268.35
164	8/15/96	5665.78	208	10/16/96	6020.81	252	12/17/96	6308.33
165	8/16/96	5689.45	209	10/17/96	6059.20	253	12/18/96	6346.77
166	8/19/96	5699.44	210	10/18/96	6094.23	254	12/19/96	6473.64
167	8/20/96	5721.26	211	10/21/96	6090.87	255	12/20/96	6484.40
168	8/21/96	5689.82	212	10/22/96	6061.80	256	12/23/96	6489.02
169	8/22/96	5733.47	213	10/23/96	6036.46	257	12/24/96	6522.85
170	8/23/96	5722.74	214	10/24/96	5992.48	258	12/25/96	6522.85
171	8/26/96	5693.89	215	10/25/96	6007.02	259	12/26/96	6546.68
172	8/27/96	5711.27	216	10/28/96	5972.73	260	12/27/96	6560.91
173	8/28/96	5712.38	217	10/29/96	6007.02	261	12/30/96	6549.37
174	8/29/96	5647.65	218	10/30/96	5993.23	262	12/31/96	6448.27

Repeat the above exercises using log(DJIA) instead of DJIA. Are your conclusions similar? Do you notice any differences?

8.5 Refer again to the DJIA data in Exercise 8.4.

(a) Use the form of the model you found adequate in Exercise 8.4 and refit the model but using only the trading days in the first six months of 1996 (the 130 days in Table 8.10). Compute the residual mean square.

(b) Use the above model to predict the daily DJIA for the first fifteen trading days in July 1996 (Table 8.11). Compare your pedictions with the actual values of the DJIA in Table 8.11 by computing the *prediction errors*, which is the difference between the actual values of the DJIA for the first 15 days of July, 1996 and their corresponding values predicted by the model.

(c) Compute the average of the squared prediction errors and compare with the residual mean square.

(d) Repeat the above exercise but using the model to predict the daily DJIA for the second half of the year (132 days).

(e) Explain the results you obtained above in the light of the scatter plot the DJIA versus Day.

8.6 Continuing with modeling the DJIA data in Exercises 8.4 and 8.5. A simplified version of the so-called *random walk model* of stock prices states that the best prediction of the stock index at Day t is the value of the index at Day $t - 1$. In regression model terms it would mean that for the models fitted in Exercises 8.4 and 8.5 the constant term is 0, and the regression coefficient is 1.

(a) Carry out the appropriate statistical tests of significance. (Test the values of the coefficients individually and then simultaneously.) Which test is the appropriate one: the individual or the simultaneous?

(b) The random walk theory implies that the first differences of the index (the difference between successive values) should be independently normally distributed with zero mean and constant variance. Examine the first differences of DJIA and log(DJIA) to see if the this hypothesis holds.

(c) DJIA is widely available. Collect the latest values available to see if the findings for 1996 hold for the latest period.