Introduction to Regression Modeling

Bovas Abraham
University of Waterloo
Johannes Ledolter
University of Iowa

THOMSON
BROOKS/COLE™

Australia • Canada • Mexico • Singapore • Spain
United Kingdom • United States

THOMSON

BROOKS/COLE

Acquisitions Editor: *Carolyn Crockett*
Assistant Editor: *Ann Day*
Editorial Assistant: *Rhonda Letts*
Technology Project Manager: *Burke Taft*
Marketing Manager: *Tom Ziolkowski*
Marketing Assistant: *Jessica Bothwell*
Advertising Project Manager: *Nathaniel Bergson-Michelson*
Project Manager, Editorial Production: *Belinda Krohmer*
Art Director: *Rob Hugel*
Print/Media Buyer: *Karen Hunt*
Permissions Editor: *Kiely Sexton*

Printed in the United States of America
1 2 3 4 5 6 7 09 08 07 06 05

For more information about our products, contact us at:
Thomson Learning Academic Resource Center
1-800-423-0563

For permission to use material from this text or product, submit a request online at
http://www.thomsonrights.com.

Any additional questions about permissions can be submitted by email to
thomsonrights@thomson.com.

Library of Congress Control Number: 2004109839

ISBN 0-534-42075-3

Production Service: *Matrix Productions*
Copy Editor: *Dan Hays*
Illustrator: *International Typesetting and Composition*
Cover Designer: *Carolyn Deacy*
Cover Image: © *Alberto Incrocci/Getty Images*
Cover Printer: *Quebecor World, Taunton*
Compositor: *International Typesetting and Composition*
Printer: *Quebecor World, Taunton*

Thomson Higher Education
10 Davis Drive
Belmont, CA 94002-3098
USA

Asia (including India)
Thomson Learning
5 Shenton Way #01-01
UIC Building
Singapore 068808

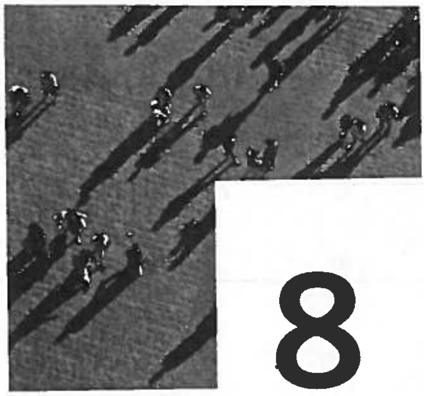
Australia/New Zealand
Thomson Learning Australia
102 Dodds Street
Southbank, Victoria 3006
Australia

Canada
Thomson Nelson
1120 Birchmount Road
Toronto, Ontario M1K 5G4
Canada

UK/Europe/Middle East/Africa
Thomson Learning
High Holborn House
50/51 Bedford Row
London WC1R 4LR
United Kingdom

Latin America
Thomson Learning
Seneca, 53
Colonia Polanco
11560 Mexico D.F.
Mexico

Spain (includes Portugal)
Thomson Paraninfo
Calle Magallanes, 25
28015 Madrid, Spain



8 Case Studies in Linear Regression

In this chapter, we apply what we have learned about regression to the analysis of several data sets. We have selected one project on the educational achievement of Iowa students. One project deals with the price of Bordeaux wine, whereas another tries to predict the auction price of livestock. A fourth project addresses the prediction of U.S. presidential elections. The scope of these projects is broad enough so that we can illustrate various aspects of regression modeling, including model selection, model estimation, and diagnostic checking. Although we do not cover all details, we give the reader suggestions of what other analyses one could try. We conclude the chapter by presenting guidelines and examples of reader-driven projects in which students of this text select the projects and collect their own data for analysis.

8.1 EDUCATIONAL ACHIEVEMENT OF IOWA STUDENTS

Data on average test scores for 325 school districts in the State of Iowa are given in the file **iowastudent**. The data set from the 2000–2001 school year includes average scores on the mathematics, reading, and science portions of the Iowa Test of Basic Skills for grades 4, 8, and 11. These tests are administered each year to monitor the progress of Iowa students. The data set also includes information on the size of the school district (number of students in the district), average teacher salary (in $), and average teacher experience (in years).

Here, we address two types of issues:

1. Achievement issues: How are test scores related across fields (math, reading, and science), how are achievements related across grades (4th, 8th, and 11th grade), and how are test scores related to the size of the district?
2. Salary issues: Do teacher salaries depend on the size of the district and teacher experience?

TABLE 8.1 PAIRWISE CORRELATION COEFFICIENTS: IOWA EDUCATIONAL ACHIEVEMENT

Correlations among math, reading, and science scores at various grades

	4th math	8th math	11th math	4th read	8th read	11th read	8th sci
8th math	0.415						
11th math	0.256	0.290					
4th read	**0.663**	0.378	0.232				
8th read	0.369	**0.691**	0.260	0.386			
11th read	0.238	0.340	**0.616**	0.241	0.327		
8th sci	0.188	0.521	0.186	0.240	0.530	0.274	
11th sci	0.199	0.230	0.462	0.219	0.269	**0.646**	0.312

Correlations among discipline averages (averaged over years)

	Math avg	Read avg
Read avg	**0.773**	
Sci avg	0.494	**0.617**

Correlations among grade averages (averaged over disciplines)

	4th avg	8th avg
8th avg	0.385	
11th avg	0.318	0.385

8.1.1 ANALYSIS OF ACHIEVEMENT SCORES

Pairwise correlations among math, reading, and science scores for 4th, 8th, and 11th graders are listed in Table 8.1. Scatter plots are not shown here, but you can check that the relationships in most graphs are well described by linear models. Correlations that exceed 0.60 are set in bold type. Reading and math scores at all three grade levels and grade 11 reading and science scores are correlated most strongly. School districts that score high in reading tend to score high on math also, indicating that "good" school districts tend to be good in both areas.

Correlations of test scores across grades are also shown in Table 8.1. Although it is true that test scores are correlated across grades, the correlation across grades is considerably weaker than the correlation across disciplines.

Are test scores related to the size of the district? In the following analysis, we divide school districts into three size groups: large districts (more than 2,000 students), medium-sized districts (between 1,000 and 2,000 students), and small districts (less than 1,000 students).

We use the one-way classification analysis of variance (ANOVA) model in Section 5.3 with average district test score as response and three parameters to represent the mean scores of the three groups. That is,

$$y_i = \mu_1 x_{i1} + \mu_2 x_{i2} + \mu_3 x_{i3} + \varepsilon_i = \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i3} + \varepsilon_i \tag{8.1}$$

where x_{i1}, x_{i2}, x_{i3} are indicator variables that are one if an observation comes from a small, medium, or large district; see the discussion in Chapter 5. The model can be parameterized without an intercept and three parameters μ_1, μ_2, and μ_3, which are the means of the three groups, or it can be written as a regression model with an intercept and two parameters that relate the size effects to group 1; that is, $\beta_0 = \mu_1$, $\beta_1 = \mu_2 - \mu_1$, and $\beta_2 = \mu_3 - \mu_1$.

The results are given in Table 8.2. For 4th and 8th grades there is little support for size differences among the test scores. However, there is strong evidence that test scores for 11th graders increase with the size of the district. Large school

TABLE 8.2 ANOVA FOR EDUCATIONAL ACHIEVEMENT SCORES AND SIZE OF THE SCHOOL DISTRICT. MINITAB OUTPUT

ANOVA: 4th grade averages

```
Analysis of Variance for 4th average
Source       DF          SS          MS           F          P
Size          2       151.2        75.6        0.76      0.467
Error       319     31597.9        99.1
Total       321     31749.2
                                             Individual 95% CIs For Mean
                                             Based on Pooled StDev
Level         N        Mean       StDev      -+---------+---------+---------+-----
small       222      71.092      10.124                (----*-----)
medium       67      69.716       9.893       (---------*--------)
large        33      72.121       8.817             (------------*------------)
                                             -+---------+---------+---------+-----
Pooled StDev =        9.953                 67.5      70.0      72.5      75.0
```

ANOVA: 8th grade averages

```
Analysis of Variance for 8th average
Source       DF          SS          MS           F          P
Size          2        53.7        26.8        0.40      0.671
Error       261     17525.2        67.1
Total       263     17578.9
                                             Individual 95% CIs For Mean
                                             Based on Pooled StDev
Level         N        Mean       StDev      ---+---------+---------+---------+---
small       178      73.623       8.417            (-----*-----)
medium       55      73.576       7.329      (----------*----------)
large        31      75.011       8.339            (-------------*-------------)
                                             ---+---------+---------+---------+---
Pooled StDev =        8.194                  72.0      74.0      76.0      78.0
```

ANOVA: 11th grade averages

```
Analysis of Variance for 11th average
Source       DF          SS          MS           F          P
Size          2       838.7       419.3        4.28      0.015
Error       250     24489.1        98.0
Total       252     25327.8
```

(Continued)

TABLE 8.2 (Continued)

```
                                            Individual 95% CIs For Mean
                                            Based on Pooled StDev
Level         N       Mean      StDev    ----------+---------+---------+------
small       167     73.608     10.729    (----*----)
medium       54     75.642      8.792       (--------*--------)
large        32     79.000      6.432                (----------*-----------)
                                         ----------+---------+---------+------
Pooled StDev =       9.897                       75.0      78.0      81.0
```

ANOVA: 11thgrade math averages

```
Analysis of Variance for 11th math average
Source       DF         SS        MS       F        P
Size          2      362.5     181.3    1.84    0.160
Error       314    30871.2      98.3
Total       316    31233.8
                                            Individual 95% CIs For Mean
                                            Based on Pooled StDev
Level         N       Mean      StDev    -----+---------+---------+---------+-
small       215     78.316     10.551       (----*-----)
medium       68     78.632      9.431    (---------*--------)
large        34     81.824      5.744             (------------*-------------)
                                         -----+---------+---------+---------+-
Pooled StDev =       9.915                  77.5      80.0      82.5      85.0
```

ANOVA: 11thgrade reading averages

```
Analysis of Variance for 11th reading average
Source       DF         SS        MS         F        P
Size          2       1042       521      3.63    0.028
Error       310      44516       144
Total       312      45558
                                            Individual 95% CIs For Mean
                                            Based on Pooled StDev
Level        'N       Mean      StDev    --+---------+---------+---------+----
small       212      71.12      13.05     (---*----)
medium       68      73.71      10.13       (--------*-------)
large        33      76.61       7.33            (-----------*-----------)
                                         --+---------+---------+---------+----
Pooled StDev =       11.98              70.0      73.5      77.0      80.5
```

ANOVA: 11thgrade science averages

```
Analysis of Variance for 11th science average
Source       DF         SS        MS      F        P
Size          2        963       482   2.64    0.073
Error       253      46183       183
Total       255      47146
                                            Individual 95% CIs For Mean
                                            Based on Pooled StDev
Level         N       Mean      StDev    --------+---------+---------+--------
small       169      72.85      14.54    (-----*-----)
medium       54      76.30      11.18          (---------*---------)
large        33      77.76      11.25           (------------*------------)
                                         --------+---------+---------+--------
Pooled StDev =       13.51                      73.5      77.0      80.5
```

districts score significantly higher than small and medium-sized districts. The test of H_0: $\mu_1 = \mu_2 = \mu_3$, or $\beta_1 = \beta_2 = 0$, leads to the F statistic = 4.28 with probability value 0.015. Examining 11th grade math, reading, and science scores separately, we find that the size of the district matters most for achievement on the reading portion of the test (F ratio = 3.63, probability value = 0.028). The reason why 11th graders score higher in large districts may have to do with the additional educational opportunities large districts can provide.

Table 8.2 lists the output from the one-way ANOVA command of the Minitab statistical software package. Minitab refers to the square root of the mean square error of the regression model in Eq. (8.1) as the pooled standard deviation. That is,

$$\text{Pooled StDev} = \sqrt{\text{SSE}/\text{df}_{\text{Error}}}$$

where SSE is the error sum of squares, and df_{Error} are its degrees of freedom. It is an estimate of $\sigma = \sqrt{V(\varepsilon_i)}$. The 95% confidence interval for the group mean μ_i in this output is calculated from

$$\bar{y}_i \pm t(0.975;\ \text{df}_{\text{Error}}) \frac{\text{Pooled StDev}}{\sqrt{n_i}}$$

where $\bar{y}_i$ is the sample mean of the n_i observations in group i. These confidence intervals are shown in Table 8.2 as dashed lines.

8.1.2 ANALYSIS OF TEACHER SALARIES

Let us consider the average teacher salary in the school district as the response variable y. Does size of the school district matter? The box plots in Figure 8.1 show that average salaries increase with size of the district, but that the variability

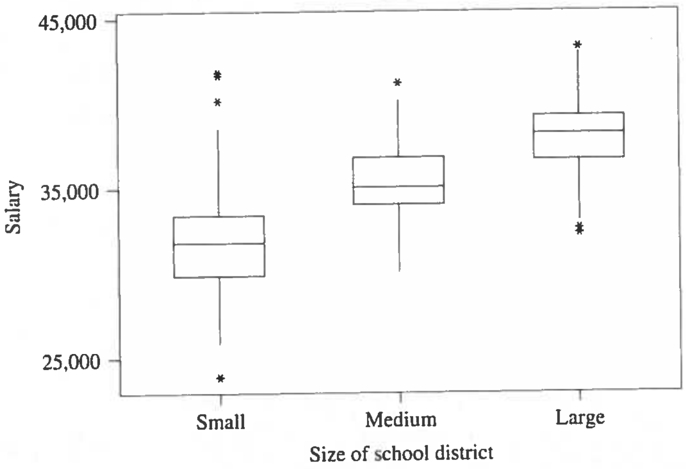


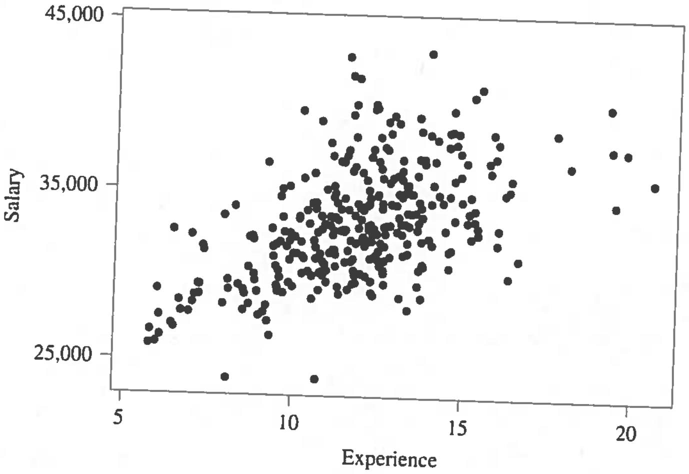
FIGURE 8.1 Box plots of teacher salaries and size of the district. MINITAB uses Q1 − 1.5 (Q3 − Q1) and Q3 + 1.5 (Q3 − Q1) for the endpoints of the lower and upper whiskers; Q1 and Q3 are the first and third quartiles. Observations beyond the endpoints of the whiskers are considered outliers and are denoted by asterisks. A test for the equality of the variances of the three groups was also considered but was found insignificant at the 0.05 significance level

TABLE 8.3 ANOVA FOR AVERAGE TEACHER SALARY AND SIZE OF THE SCHOOL DISTRICT. MINITAB OUTPUT

ANOVA: Salary versus size

```
Analysis of Variance for salary
Source      DF          SS          MS        F       P
Size         2   1.475E+09   737470964   103.30   0.000
Error      322   2.299E+09     7138790
Total      324   3.774E+09

                                          Individual 95% CIs For Mean
                                          Based on Pooled StDev
Level        N        Mean       StDev    ---+---------+---------+---------+---
small      223       31799        2820    (-*-)
medium      68       35326        2249                (---*--)
large       34       37834        2425                           (---*----)
                                          ---+---------+---------+---------+---
Pooled StDev =        2672                32000     34000     36000     38000
```

FIGURE 8.2 Scatter Plot of Teacher Salaries against Teacher Experience



(i.e., the width of the boxes) is approximately constant across the three size groups. This is important because the regression (ANOVA) model in Eq. (8.1) with average teacher salary as the response assumes constant error variance.

The results of the one-way ANOVA model in Eq. (8.1) for teacher salaries and the three district size groups are given in Table 8.3. The results show quite convincingly that teacher compensation increases with the size of the district. This effect may have to do with the cost of living. Large school districts are mostly located in urban settings, and it is more expensive to live in urban areas than in small rural towns.

Does experience have an effect on salaries? One would expect that teachers with more experience get paid more, and districts with larger average experience have higher average salaries. The scatter plot in Figure 8.2 confirms

TABLE 8.4 REGRESSION OF AVERAGE TEACHER SALARY ON SIZE OF THE SCHOOL DISTRICT AND TEACHER EXPERIENCE

```
The regression equation is
salary = 24450 + 631 experience + 3053 med size + 5522 large size

Predictor                 Coef       SE Coef          T        P
Constant               24450.1         581.4      42.06    0.000
Experience              630.98         48.34      13.05    0.000
Medium                  3052.8         301.8      10.11    0.000
Large                   5522.4         400.1      13.80    0.000

S = 2163            R-Sq = 60.2%          R-Sq(adj) = 59.8%

Analysis of Variance

Source                      DF            SS          MS        F         P
Regression                   3    2271932689   757310896   161.88     0.000
Residual Error             321    1501699566     4678192
Total                      324    3773632255
```

Diagnostics: Cases with large standardized residuals

```
Case School District         Exper      Size        Salary      Fit    Residual
 79  (Davies County)         16.34      medium       30051    37813      -7762
113  (Sioux Center)          11.82      small        41656    31908       9748
115  (Hudson)                11.63      small        41805    31788      10017
230  (Kingsley-Pierson)      12.31      small        40151    32217       7934
325  (Lineville-Clio)        10.72      small        23912    31214      -7302
```

this hypothesis. However, we already learned that size of the district matters. Adding the size of the district to the model leads to the regression

$$y_i = \beta_0 + \beta_1 \text{Experience} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \tag{8.2}$$

The indicators for size were explained previously. The small district group becomes the reference in our comparison. The parameter β_2 measures, for fixed experience, the difference in the average salaries of medium-size and small school districts. The parameter β_3 measures, for fixed experience, the difference in average salaries of large and small school districts. The difference $\beta_3 - \beta_2$ measures, for fixed experience, the difference in average salaries of large and medium-size school districts. The parameter β_1 measures, for given size of the school district, the benefit of an additional year of experience on average pay.

The regression summary in Table 8.4 shows that there is approximately a \$3,000 difference in the average salaries of small and medium-size districts and a \$2,500 difference in the average salaries of medium-size and large districts. Each additional year of experience "costs" the district (and earns the teachers) approximately \$630.

Leverages of six cases are larger than three times the average leverage, 0.037; the largest leverage is 0.047. However, the largest Cook's statistic (0.07569) is rather unremarkable. The histogram of the standardized residuals in Figure 8.3 shows that there are several large residuals. The five cases with standardized

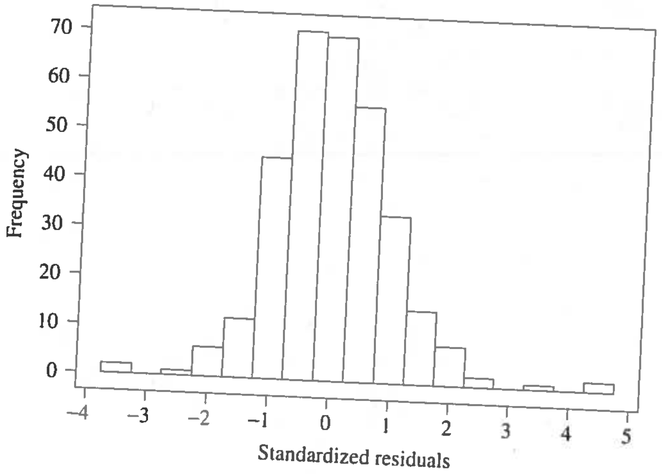


FIGURE 8.3 Histogram of Standardized Residuals from the Regression Model (8.2)

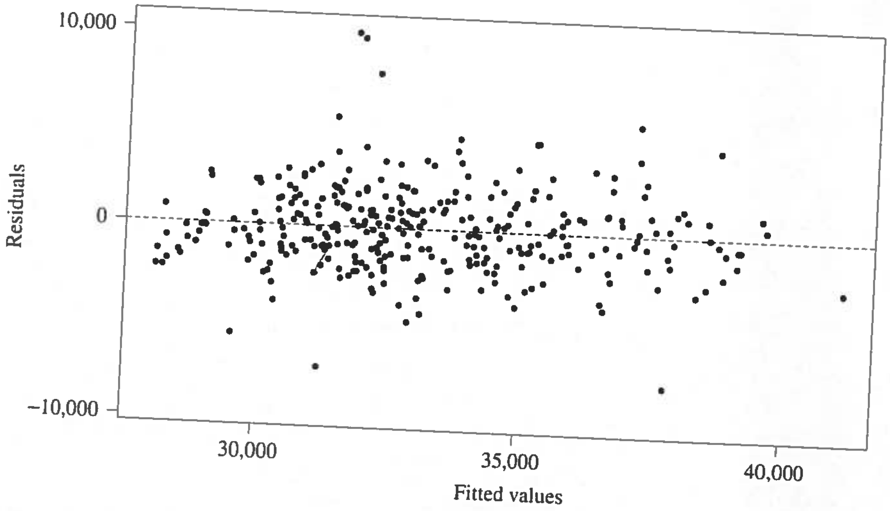


FIGURE 8.4 Scatter Plot of the Residuals against the Fitted Values, Regression Model (8.2)

residuals outside ±3 are listed in Table 8.4. The scatter plot of residuals against fitted values in Figure 8.4 fails to reveal major problems with model (8.2).

8.1.3 CONCLUDING COMMENTS

The average test scores of a school district depend on many factors, such as

- the intellectual ability of the incoming students;
- the support students get from their parents; and
- the instruction that is provided by the school and the teachers.

Of course, all three factors are difficult to measure. One could try to get a measure of student "ability" such as an average intelligence score on the student cohort that enters each district. However, because of the sensitive nature of such data, it may be almost impossible to obtain this information. One could try to assess—through surveys—the role families play in student education. The average number of hours children are tutored or the number of extracurricular activities students are exposed to could be used as proxies. Quality of instruction is very difficult to measure. One could try to obtain proxies such as the average amount of money a school district spends on each child.

Economic factors, such as the average income or the poverty rate of the school district's county, are often used to "explain" educational achievement. Regression results will find strong relationships among test scores and poverty; we already saw this in the example in Section 2.8. However, the mechanisms of the relationship are unclear because wealth (or poverty) of a school district affects all three factors listed previously. Economic conditions will attract smarter students to certain school districts (by families moving into desirable areas), affect the amount of outside help parents provide to their children, and impact the resources the district can provide. A claim that good test scores are due to excellence in instruction may be premature. The ability and composition of the student body and the family support children receive also play a major role.

8.2 PREDICTING THE PRICE OF BORDEAUX WINE

The following data are discussed in Ashenfelter, Ashmore, and Lalonde (1995). For additional discussion, see *Barron's* (December 30, 1996, pp. 17–19) and Chapter 6 of Fair (2002).

Traditionally, the quality of a Bordeaux vintage is first evaluated by experts in March of the following year. These first ratings, however, are rather unreliable because a 4-month old wine is a rather foul mixture of fermenting grape juice and little like the magnificent stuff it can become years later. Wouldn't it be wonderful to be able to rate the quality (and hence predict the price) of the most recent Bordeaux vintage immediately, compared to having to wait several months before a first, and usually inaccurate, assessment of its quality can be made?

Price data are obtained from the London market, a main market for fine wines. Price information from 1990–1991 auctions representing six chateaus (Latour, Lafite, Cheval Blanc, Pichon-Lalande, Cos d'Estournel, and Montrose) are averaged for each vintage, and the average price is expressed as a fraction of the price of the 1961 vintage (which was truly outstanding).

The theory is that the quality of the wine, and hence vintage price, depends on weather variables—such as the average temperature during the growing season (April–September, in degrees centigrade), the amount of rain during the harvest season (August and September, in total millimeters), and the amount of rain in the preceding October–March period—and the age of the vintage. It is thought that conditions for the vintage are best when the growing season is warm, August and September are dry, and the previous winter was wet. Furthermore, because of storage expenses older wines should cost more than younger ones.

Data for the years 1952–1980 are listed in Table 8.5. The 1962 price of 0.331 in column 2, for example, implies that the price of the 1962 vintage amounted to

TABLE 8.5 PRICE AND GROWING CONDITIONS OF BORDEAUX WINE, 1952–1980[a]

Vintage	Price (1961 = 1)	Average Temperature, April–September (°C)	Rainfall, August–September (ml)	Rainfall Previous, October–March (ml)	Age (1983 = 0)
1952	0.368	17.12	160	600	31
1953	0.635	16.73	80	690	30
1954	*	*	*	*	29
1955	0.446	17.15	130	502	28
1956	*	*	*	*	27
1957	0.221	16.13	110	420	26
1958	0.180	16.42	187	582	25
1959	0.658	17.48	187	485	24
1960	0.139	16.42	290	763	23
1961	1.000	17.33	38	830	22
1962	0.331	16.30	52	697	21
1963	0.168	15.72	155	608	20
1964	0.306	17.27	96	402	19
1965	0.106	15.37	267	602	18
1966	0.473	16.53	86	819	17
1967	0.191	16.23	118	714	16
1968	0.105	16.20	292	610	15
1969	0.117	16.55	244	575	14
1970	0.404	16.67	89	622	13
1971	0.272	16.77	112	551	12
1972	0.101	14.98	158	536	11
1973	0.156	17.07	123	376	10
1974	0.111	16.30	184	574	9
1975	0.301	16.95	171	572	8
1976	0.253	17.65	247	418	7
1977	0.107	15.58	87	821	6
1978	0.270	15.82	51	763	5
1979	0.214	16.17	122	717	4
1980	0.136	16.00	74	578	3

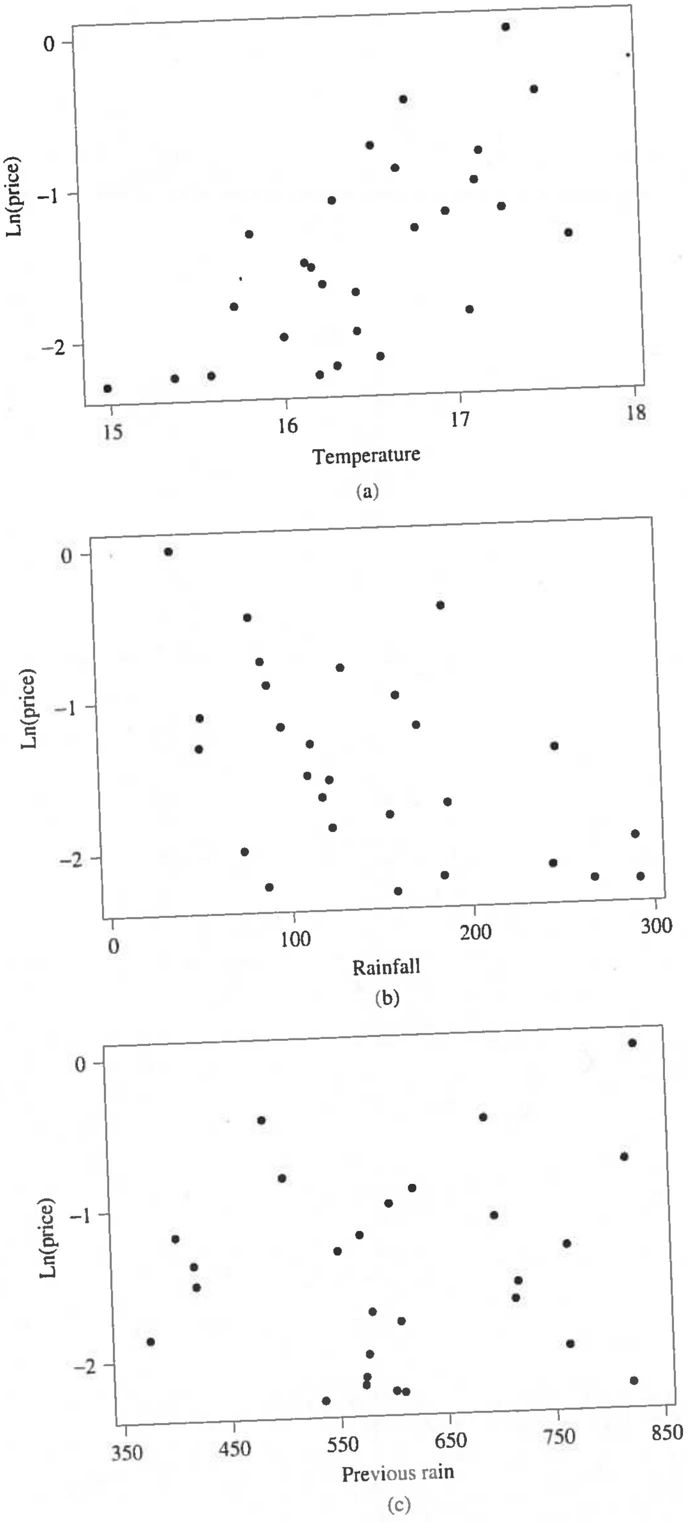
[a] The data are stored in the file **wine**.

33.1% of the price of the 1961 vintage. The prices for the 1954 and 1956 vintages are missing. Prices for these two vintages could not be established because the 1954 and 1956 vintages were poor and very little wine was sold.

Scatter diagrams of price against each of the four predictor variables suggest that the relationship between price and these predictor variables is not linear and that a logarithmic transformation of the response may be beneficial (these graphs are not shown here; we encourage you to check this conclusion). Scatter diagrams of the logarithm of price against the four predictor variables are shown in Figures 8.5a–8.5d. The results of fitting the linear model

$$\text{Ln(Price)} = \beta_0 + \beta_1 \text{ Temp} + \beta_2 \text{ Rain} + \beta_3 \text{ PRain} + \beta_4 \text{ Age} + \varepsilon \qquad (8.3)$$

FIGURE 8.5 Logarithm of Price against (a) Temperature, (b) Rainfall, (c) Previous Rain, and (d) Age and (e) Plot of Residuals against Fitted Values, Model (8.3)

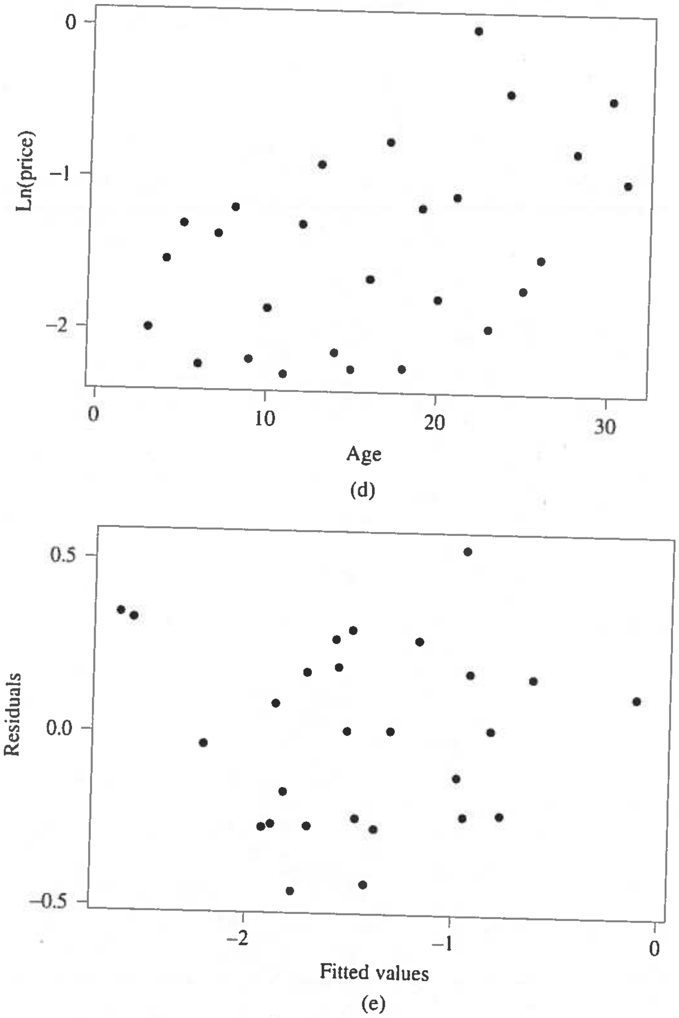


(a)

(b)

(c)

FIGURE 8.5
(Continued)



are shown in Table 8.6. The coefficients have the anticipated signs and are statistically significant, indicating that it is not possible to simplify the model. Backtransformation of Eq. (8.3) through exponentiation shows that the regression coefficients are related to percentage changes in price when changing covariates by one unit. The positive coefficient on age (0.0239) implies that the average price increases by $100(e^{0.0239} - 1) = 2.4\%$ annually (assuming, of course, that all other covariates are unchanged). The model in Eq. (8.3) explains about 83% of the variation in the response. The plot of the residuals against the fitted values in Figure 8.5e is rather unremarkable; it shows no gross violations of the fitted model. Also, none of the leverages and none of the Cook's distances are unusually large; the largest residual (case 6 for 1959) barely exceeds 2 standard deviations.

The fitting results and the model diagnostics show that Eq. (8.3) represents a fairly respectable model. The model lends support to the theory that the price of a

TABLE 8.6 MINITAB REGRESSION OUTPUT OF MODEL (8.3)

```
Regression Analysis: Ln(Price) versus Temp, Rain, PRain, Age

The regression equation is
Ln(Price) = - 12.2 + 0.617 Temp - 0.00387 Rain + 0.00117 PRain + 0.0239 Age

Predictor           Coef      SE Coef        T        P
Constant         -12.159        1.686    -7.21    0.000
Temp             0.61699      0.09502     6.49    0.000
Rain          -0.0038659    0.0008062    -4.80    0.000
PRain          0.0011710    0.0004814     2.43    0.024
Age             0.023901     0.007155     3.34    0.003

S = 0.2861               R-Sq = 82.8%     R-Sq(adj) = 79.7%

Analysis of Variance

Source              DF          SS        MS        F          P
Regression           4      8.6795    2.1699    26.51      0.000
Residual Error      22      1.8004    0.0818
Total               26     10.4799

Unusual Observations
Obs               Temp   Ln(Price)       Fit   SE Fit   Residual   St Resid
  6               17.5     -0.4186   -0.9550   0.1124     0.5364      2.04R

R denotes an observation with a large standardized residual
```

Bordeaux wine (i.e., its quality) depends on the growing conditions. Temperature and last year's rainfall are beneficial, whereas excess rain during the growing season is detrimental. The costs of storing the wine are reflected in the positive coefficient of age.

However, the real test is the model performance when using the model to predict the prices for vintages that are not part of the sample data that were used to construct the model. How successful is the regression approach in obtaining out-of-sample predictions? Can we use the model to predict the price of a new vintage?

Ray C. Fair, in his 2002 book *Predicting Presidential Elections and Other Things* (Chapter 6), lists the growing conditions for the years 1987–1991. They are given in Table 8.7. The age variable is a counting variable that is set at 0 in 1983, and hence its values for 1987 and beyond are negative. Note that it does not matter which year is taken as 0 (here, 1983); in a linear model such as Eq. (8.3), all that matters is that the variable changes by the same constant amount.

We use model (8.3) to get out-of-sample price predictions for 1987–1991. Using the new growing conditions, we calculate

$$\begin{aligned}\text{Ln(Price)} &= -12.159 + 0.617\text{Temp} - 0.00387\text{Rain}\\ &\quad + 0.00117\text{PRain} + 0.0239\text{Age}\end{aligned}$$

TABLE 8.7 GROWING CONDITIONS FOR THE PREDICTION SET, 1987–1991[a]

Vintage	Average Temperature	Rainfall	Previous Rainfall	Age (1983 = 0)	Prediction for Ln(Price)	Prediction for Price	Actual Price
1987	16.98	115	452	−4	−1.69301	0.184	0.135
1988	17.10	59	808	−5	−1.00952	0.364	0.271
1989	18.60	82	443	−6	−0.62425	0.536	0.432
1990	18.70	80	468	−7	−0.54945	0.578	0.568
1991	17.70	183	570	−8	−1.46909	0.230	0.142

[a] Predictions and actual prices are shown in columns 7 and 8.

and

$$\begin{aligned}\text{Price} = \exp(&-12.159 + 0.617\text{Temp} - 0.00387\text{Rain}\\ &+ 0.00117\text{PRain} + 0.0239\text{Age})\end{aligned}$$

The predictions of price are given in column 7 of Table 8.7.

Fair also lists price information that he obtained from an East Coast wine distributor. His data imply the following average prices for the 1961 and the 1987–1991 vintages: \$258.33 (1961), \$35.00 (1987), \$70.00 (1988), \$111.67 (1989), \$146.67 (1990), and \$36.67 (1991). From these numbers, he calculates the price of the 1987–1991 vintages relative to 1961 as $35.00/258.33 = 0.135$, $70.00/258.33 = 0.271$, $111.67/258.33 = 0.432$, $146.67/258.33 = 0.568$, and $36.67/258.33 = 0.142$, respectively. These are the entries in the last column of Table 8.7. A comparison of the last two columns in this table addresses the accuracy of the out-of-sample predictions. The model is correct in identifying the worst vintage (1987) and the best vintage (1990). In fact, it ranks the prices on all five vintages correctly. One can calculate percentage absolute errors, such as $100(0.184 - 0.135)/0.135 = 36.3\%$ for 1987. The mean absolute percentage error for the five periods is approximately 32%.

8.3 FACTORS INFLUENCING THE AUCTION PRICE OF IOWA COWS

Table 8.8 lists the results of livestock sales by Wapello Livestock Sales in Wapello, Iowa. This data set is part of a 2001 student project by Jay Heindel at the University of Iowa. It contains the selling price of a cow as well as various characteristics of the animal that is sold. We have also listed an explanation of how these factors can be expected to influence the price. A random sample of 115 sales over a period of 19 weeks (mid-September to the end of January 2000) is analyzed. Explanatory factors include

- Age of the animal: Cows in the mid-range may be more valuable because they have shown the ability to produce calves but are still young enough for subsequent breeding.

TABLE 8.8 AUCTION PRICE (y) AND FACTORS (x) THAT MAY HELP EXPLAIN THE PRICE[a]

Price y ($)	Age (years)	Indicator for Bred	Indicator for Angus	Frame (Large)	Weight (100 lb)	Indicator for Conditioned	Indicator for Registered
1,000	3	1	1	1	10.15	1	0
1,250	3	1	1	1	11.00	1	0
980	5	1	0	0	11.15	1	0
1,015	4	0	1	0	11.00	1	0
995	5	1	0	1	10.00	0	0
825	7	1	0	0	9.80	0	0
850	6	1	0	0	10.25	0	1
1,150	2	1	1	0	10.50	1	1
1,150	2	1	1	1	10.75	1	1
1,200	3	1	1	1	11.75	1	0
1,200	2	1	1	1	11.60	1	0
1,000	6	1	1	0	11.00	1	0
1,000	7	0	1	0	10.00	0	0
1,050	6	1	0	0	10.00	0	1
1,075	4	1	1	1	9.90	0	1
1,165	5	1	1	0	11.35	1	0
780	2	0	1	0	8.85	0	0
800	2	0	1	0	9.50	0	1
1,180	3	0	0	1	11.45	1	1
1,000	4	0	0	1	11.45	0	1
1,200	2	1	0	1	11.50	1	1
1,000	3	1	1	0	10.00	1	0
1,025	3	1	1	0	9.75	1	0
1,175	2	0	1	1	11.35	1	1
800	5	1	0	0	12.00	1	0
915	4	0	1	0	11.85	1	1
1,185	6	1	1	1	12.00	0	0
1,020	5	1	1	0	11.25	0	0
775	7	1	1	0	12.00	1	0
850	7	1	0	1	12.00	1	1
1,200	2	1	1	0	10.00	1	1
1,200	3	1	1	0	10.15	1	1
775	8	1	0	0	12.50	1	0
775	7	1	0	0	11.85	1	0
1,200	3	1	0	1	11.85	0	1
1,135	4	1	0	0	12.00	1	0
1,000	7	1	0	0	11.50	1	0
1,185	3	0	0	1	12.00	0	1
1,155	3	1	0	0	11.85	1	0
1,155	2	0	0	1	12.00	1	1
1,175	2	1	1	1	11.75	1	0
1,200	2	1	0	1	11.50	1	0
1,165	3	1	1	0	11.65	1	0
1,000	6	1	1	0	12.25	1	0
1,200	2	0	1	1	10.25	0	0
1,175	2	1	0	0	10.25	0	1

(Continued)

TABLE 8.8 (Continued)

Price *y* ($)	Age (years)	Indicator for Bred	Indicator for Angus	Frame (Large)	Weight (100 lb)	Indicator for Conditioned	Indicator for Registered
1,000	5	0	0	0	10.00	0	1
1,125	5	0	0	0	10.00	1	0
1,150	4	1	1	1	12.25	1	0
1,125	2	0	0	1	11.25	1	1
875	6	0	0	0	10.00	0	0
1,000	6	1	0	1	11.25	0	0
1,185	3	1	1	0	10.00	0	0
1,185	4	1	1	0	10.00	1	1
1,000	4	1	1	0	11.00	1	0
950	4	1	1	1	10.25	0	0
875	6	1	0	0	11.25	1	0
775	6	0	0	0	10.15	0	0
1,100	2	0	1	0	9.85	0	1
1,125	2	0	1	0	10.35	1	1
1,200	2	0	0	0	10.65	1	1
1,175	4	1	1	1	12.30	1	0
1,000	4	1	0	1	11.75	0	0
1,000	3	1	0	1	11.85	0	0
1,000	6	1	1	0	11.25	0	0
1,000	6	1	0	1	11.25	1	0
985	5	0	1	0	10.50	0	0
1,150	3	1	1	0	11.75	1	1
1,150	4	1	1	0	11.50	1	0
1,075	7	1	0	0	11.45	1	0
1,050	6	0	0	0	11.50	0	0
885	6	0	0	0	10.00	0	0
1,200	4	1	1	1	11.50	1	1
1,150	3	1	1	0	11.50	1	1
1,000	4	1	0	0	10.85	0	0
1,075	2	1	1	0	10.85	0	1
980	4	0	0	0	12.50	1	0
980	3	0	0	0	12.45	1	0
1,000	6	1	1	1	12.00	1	0
1,085	3	1	0	0	11.25	1	0
1,175	5	1	0	0	11.55	1	0
1,150	5	1	0	1	11.15	0	1
1,175	3	1	0	0	11.15	1	0
1,000	2	0	1	0	10.00	1	0
875	7	1	1	1	11.65	1	0
995	5	1	1	1	11.55	1	1
1,000	3	1	1	1	10.15	1	1
1,250	3	1	1	1	11.00	1	1
1,150	2	1	0	1	11.00	1	0
1,150	5	1	0	1	10.85	1	0
1,200	5	1	0	1	10.95	1	1
980	8	1	1	0	12.75	1	0
995	3	0	0	0	10.00	0	0

TABLE 8.8 (Continued)

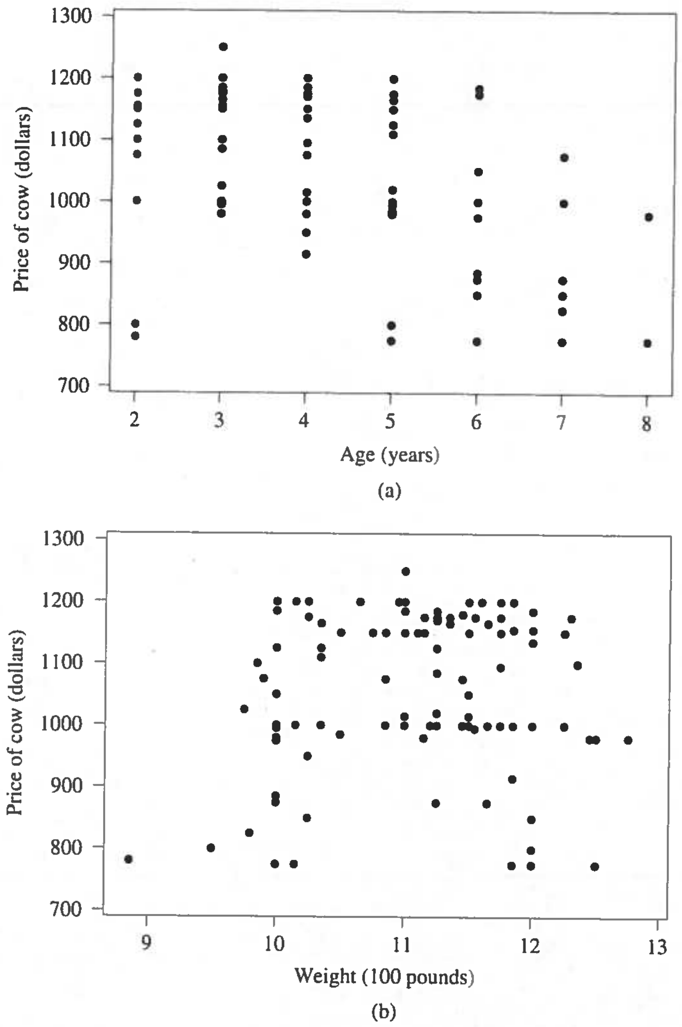
Price y ($)	Age (years)	Indicator for Bred	Indicator for Angus	Frame (Large)	Weight (100 lb)	Indicator for Conditioned	Indicator for Registered
1,000	3	1	1	1	10.35	1	0
1,165	3	1	0	1	10.35	1	0
1,175	5	0	1	0	11.25	1	1
1,000	6	1	1	1	11.20	1	0
775	5	0	1	1	10.00	0	0
1,000	3	1	1	1	11.65	1	0
1,015	4	1	0	1	11.50	1	0
1,200	4	1	1	1	11.50	1	0
1,175	6	1	1	1	11.75	1	1
1,095	4	0	1	1	11.75	1	0
1,100	3	1	1	0	12.35	1	1
980	3	0	0	0	10.00	0	0
1,110	5	0	1	1	10.35	0	0
1,200	2	1	1	1	11.00	1	0
1,185	4	0	1	1	11.25	1	0
1,185	3	0	1	0	11.00	1	0
1,170	4	0	1	0	11.25	1	0
1,150	3	1	0	0	11.10	1	0
1,000	4	1	1	1	11.00	1	1
975	6	1	0	0	10.00	0	0
1,200	4	0	1	0	11.00	1	0
1,185	3	1	1	0	11.00	1	0

[a] The data are stored in the file cows.

- Weight of the animal: Cows are intended for the production of calves, and they are not being sold for meat. Hence, the weight of the cow may not be the most deciding factor because it does not tell about the cow's ability to produce healthy marketable calves.
- Whether the cow has been bred (i.e., currently carrying a calf), indicating that a "free" calf comes with the sale.
- Frame size of the animal: A large size may avoid birthing problems in the future.
- Whether the cow is registered (recorded through a breed organization as a legitimate bloodline of a particular breed).
- Whether the cow is in good condition (i.e., well fed and "filled out").
- Whether an Angus cow is involved: Angus cattle may be more valuable because consumers are willing to pay more for their leaner meat.

Scatter plots of price against the explanatory variables age and weight are shown in Figure 8.6. Box plots of weight against the categorical variables (whether the cow has been bred, is conditioned, its frame size, whether it is an Angus cow,

FIGURE 8.6 Scatter Plots of Auction Price against Age and Weight of the Cow



(a)

(b)

and registered) were also drawn but are not shown here. The graphs in Figure 8.6 suggest that one should allow for quadratic effects of age and weight. It appears that very young and old cows fetch less money than cows in the mid-range; the same applies for weight.

We start by fitting all possible regressions using the seven original explanatory variables plus the two constructed ones (squares of age and weight). The results are shown in Table 8.9. The results indicate that a model with three explanatory variables appears to give an acceptable representation. The Mallows C_p criteria of the two identified models (in boldface type) are quite acceptable; for an acceptable model with three regressor variables we expect a C_p of approximately 4. There appears to be no bias with the three-regressor models. Weight and its square, and the age of the cow (or, alternatively, its square), seem to

TABLE 8.9 BEST SUBSETS REGRESSION OF PRICE ON THE NINE EXPLANATORY VARIABLES

Vars	R-Sq	R-Sq(adj)	C-p	S	Age	Age**2	Weight	Weight**2	Breed	Angus	Frame	Condition	Register	
1	29.7	29.1	34.6	106.40		X								
1	27.7	27.1	38.7	107.89	X									
2	36.2	35.1	23.1	101.80		X	X							
2	35.6	34.4	24.5	102.32		X		X						
3	**46.4**	**44.9**	**4.1**	**93.753**		**X**	**X**	**X**						
3	**46.2**	**44.8**	**4.5**	**93.913**	**X**		**X**	**X**						(age, weight, weight**2)
4	47.5	45.6	3.8	93.220		X	X	X		X				
4	47.3	45.4	4.1	93.337	X		X	X		X				
5	48.4	46.0	3.9	92.823		X	X	X		X			X	
5	48.1	45.7	4.6	93.119	X		X	X		X			X	
6	48.9	46.1	4.8	92.782		X	X	X	X	X			X	
6	48.8	45.9	5.1	92.904		X	X	X		X		X	X	
7	49.2	45.9	6.3	92.973		X	X	X	X	X		X	X	
7	49.0	45.7	6.7	93.142		X	X	X	X	X	X		X	
8	49.3	45.5	8.0	93.289		X	X	X	X	X	X	X	X	
8	49.2	45.4	8.3	93.401	X	X	X	X	X	X		X	X	
9	49.3	45.0	10.0	93.722	X	X	X	X	X	X	X	X	X	

[a] Only the two best models for each group are shown. The placement of the "X" symbol indicates which variables are included.

matter most. Since a linear term of age is easier to interpret than the square of age without the linear component, we consider the model with weight, weight square, and age in more detail. With this model, we explain approximately 46.2% of the variability in price. The standard deviation of the errors is 93.9 (dollars). This needs to be compared with the standard deviation of the selling price by itself, without taking advantage of the cow's characteristics; this turns out to be 126.4 (dollars). Although this is not a striking reduction, the data do not provide us with a better model. The buyers bidding on cows must use other factors that are not recorded.

Detailed estimation results for this model are shown at the top of Table 8.10. Would the additional information that the cow in question is an Angus cow make a difference? We add the indicator for Angus beef and run the extended

TABLE 8.10 ESTIMATION RESULTS FOR TWO REGRESSION MODELS

Regression Analysis: Price versus Age, Weight, Weight2**

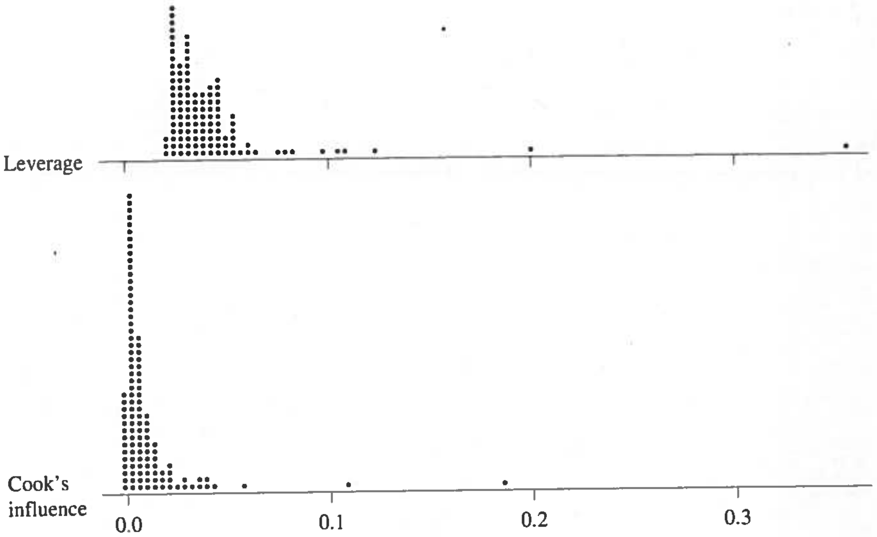
```
The regression equation is
Price = - 6745 - 42.0 Age + 1423 Weight - 63.1 Weight**2

Predictor          Coef     SE Coef        T        P
Constant          -6745        1509    -4.47    0.000
Age             -41.997       5.553    -7.56    0.000
Weight           1422.6       275.8     5.16    0.000
Weight**2        -63.07       12.58    -5.02    0.000

S = 93.91         R-Sq = 46.2%       R-Sq(adj) = 44.8%

Analysis of Variance

Source               DF        SS        MS        F        P
Regression            3    841123    280374    31.79    0.000
Residual Error      111    978981      8820
Total               114   1820104
```

Regression Analysis: Price versus Age, Weight, Weight2, Angus**

```
The regression equation is
Price = - 6887 - 40.4 Age + 1444 Weight - 64.0 Weight**2 + 27.6 Angus

Predictor          Coef     SE Coef        T        P
Constant          -6887        1503    -4.58    0.000
Age             -40.431       5.612    -7.20    0.000
Weight           1443.9       274.5     5.26    0.000
Weight**2        -64.02       12.51    -5.12    0.000
Angus             27.59       17.91     1.54    0.126

S = 93.34        R-Sq = 47.3%        R-Sq(adj) = 45.4%

Analysis of Variance

Source               DF        SS        MS        F        P
Regression            4    861812    215453    24.73    0.000
Residual Error      110    958292      8712
Total               114   1820104
```

regression model. The results in the bottom of Table 8.10 show that, as expected, the effect of Angus beef is positive, indicating that an Angus cow of equal weight and age would bring \$27.6 more on average. However, the partial t ratio, $27.59/17.91 = 1.54$, is not quite statistically significant. The probability value of a one-sided test of the hypothesis that Angus cattle do bring in more money is $0.126/2 = 0.063$; note that the two-sided probability value in the table needs to be cut in half. It is close to the commonly used 5% significance level. Strictly speaking, with a significance level of 5% one cannot reject the null hypothesis that Angus cattle are not priced higher. However, the p value is close to 0.05, and the results do suggest that Angus cattle may fetch a somewhat higher price.

FIGURE 8.7 Dot Plots of Leverages and Cook's Influence Measures from the Model with Age, Weight, Square of Weight, and Indicator for Angus Cattle



What is the optimal weight for a cow? One can take the derivative of the model equation with respect to weight and determine the optimum. The optimum is given by $1,444/[(2)(64)] = 11.28$, and since the second derivative [the coefficient of $(\text{weight})^2$ is -64] is negative, it is indeed a maximum. This means that the optimal weight is 1,128 pounds.

What about the diagnostics of the model that includes age, weight, the square of weight, and the indicator for Angus beef? Do we see serious problems? It should be mentioned that we do not know about the time arrangement of the 115 cases in Table 8.8. All we know is that the data are a sample of 115 auction sales within a 6-month period. Since we cannot assume that the arrangement is sequential in time, the Durbin–Watson statistic is not meaningful in this context. Are there high-leverage cases, and do some cases exert high influence on the regression estimates? Leverages and Cook's influence measures are calculated for the 115 cases and they are graphed in Figure 8.7. The highest leverage is 0.353; it originates from case 17, a cow with the smallest weight (885 pounds). The leverage is quite unusual and certainly much larger than three times the average leverage (the average leverage is $5/115 = 0.0435$). However, this case does not exert much influence on the regression coefficients. The largest Cook's influence measure is 0.184; it comes from case 92, a cow with the largest weight (1,275 pounds). However, it is not unusually large (our usual warning limit is 0.50 or higher) to raise suspicion. None of the cases exert unusual influence on the fitted regression. The histogram of the (standardized) residuals is shown in Figure 8.8. None of the residuals are unusually large, and the shape of the histogram suggests that the normal distribution assumption is adequate. The approximate linear appearance of the normal probability plot (not shown) also confirms that the usual error assumptions are met.

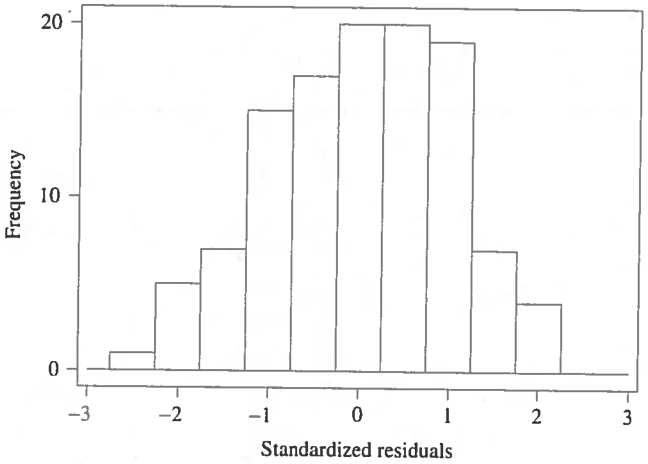


FIGURE 8.8 Histogram of the Standardized Residuals from the Model with Age, Weight, Square of Weight, and Indicator for Angus Cattle

8.4 PREDICTING U.S. PRESIDENTIAL ELECTIONS

8.4.1 A PURELY ECONOMICS-BASED MODEL PROPOSED BY RAY FAIR

Ray Fair, an economics professor at Yale University, uses the state of the economy prior to the election to predict the incumbent vote share in presidential elections. By incumbent vote share we mean the vote share for the candidate of the party occupying the White House during the election. There is a certain attraction to such forecasting models because the explanatory economic variables measure the state of the economy several months prior to the actual election. They are readily available for making real-time predictions of the election outcome.

The data in Table 8.11 are taken from Fair's book, *Predicting Presidential Elections and Other Things* (2002). Chapters 1, 3, and 4 and references in the chapter notes are relevant to our discussion. For detailed discussion and definition of the variables, see Fair's Chapter 3.

- Vote share, the response variable, represents the incumbent share of the two-party vote. By taking the two-party vote share and not the incumbent share of the total vote, one assumes that third-party candidates take about the same amount from each party. This is a reasonable assumption for most elections, except for the 1924 election. There is evidence that La Follette (of the Progressive Party) took more voters from Davis (the Democrat) than from Coolidge (the Republican). It is estimated that 76.5% of the votes for La Follette would have gone to Davis, whereas only 23.5% of the La Follette votes would have gone to Coolidge. This information is incorporated into the listed 1924 incumbent share.
- Growth rate represents the per capita growth rate of real gross domestic product (GDP) in the first three quarters (9 months) of the election year.
- Inflation rate is the average (absolute) inflation rate during the 15 quarters prior to the election (i.e., all the quarters of the administration except the

TABLE 8.11 INCUMBENT VOTE SHARE AND ITS DETERMINANTS: FAIR'S MODEL[a]

Year	Party in Power	Election Outcome	Incumbent Vote Share (%)	Growth Rate (%)	Inflation Rate (%)	Good News Quarters	Duration Value	War	President Running	Party Variable
1916	D	President Wilson beats Hughes	51.7	2.2	4.3	3	0.00	0	1	1
1920	D	Cox loses to Harding	36.1	−11.5	16.5	5	1.00	1	0	1
1924	R	Pres. Coolidge beats Davis and LaFollette	58.2	−3.9	5.2	10	0.00	0	1	0
1928	R	Hoover beats Smith	58.8	4.6	0.2	7	1.00	0	0	0
1932	R	Pres. Hoover loses to Roosevelt	40.8	−14.9	7.1	4	1.25	0	1	0
1936	D	Pres. Roosevelt beats Landon	62.5	11.9	2.4	9	0.00	0	1	1
1940	D	Pres. Roosevelt beats Willkie	55.0	3.7	0.0	8	1.00	0	1	1
1944	D	Pres. Roosevelt beats Dewey	53.8	4.1	5.7	14	1.25	1	1	1
1948	D	Pres. Truman beats Dewey	52.4	1.8	8.7	5	1.50	1	1	1
1952	D	Stevenson loses to Eisenhower	44.6	0.6	2.3	6	1.75	0	0	1
1956	R	Pres. Eisenhower beats Stevenson	57.8	−1.5	1.9	5	0.00	0	1	0
1960	R	Nixon loses to Kennedy	49.9	0.1	1.9	5	1.00	0	0	0
1964	D	Pres. Johnson beats Goldwater	61.3	5.1	1.2	10	0.00	0	1	1
1968	D	Humphrey loses to Nixon	49.6	4.8	3.2	7	1.00	0	0	1
1972	R	Pres. Nixon beats McGovern	61.8	6.3	4.8	4	0.00	0	1	0
1976	R	Ford loses to Carter	48.9	3.7	7.7	4	1.00	0	0	0
1980	D	Pres. Carter loses to Reagan	44.7	−3.8	8.1	5	0.00	0	1	1
1984	R	Pres. Reagan beats Mondale	59.2	5.4	5.4	7	0.00	0	1	0
1988	R	G. Bush beats Dukakis	53.9	2.1	3.3	6	1.00	0	0	0
1992	R	Pres. G. Bush loses to Clinton	46.5	2.3	3.7	1	1.25	0	1	0
1996	D	Pres. Clinton beats Dole	54.7	2.9	2.3	3	0.00	0	1	1

[a] The data are stored in the file **election(Fair)**.

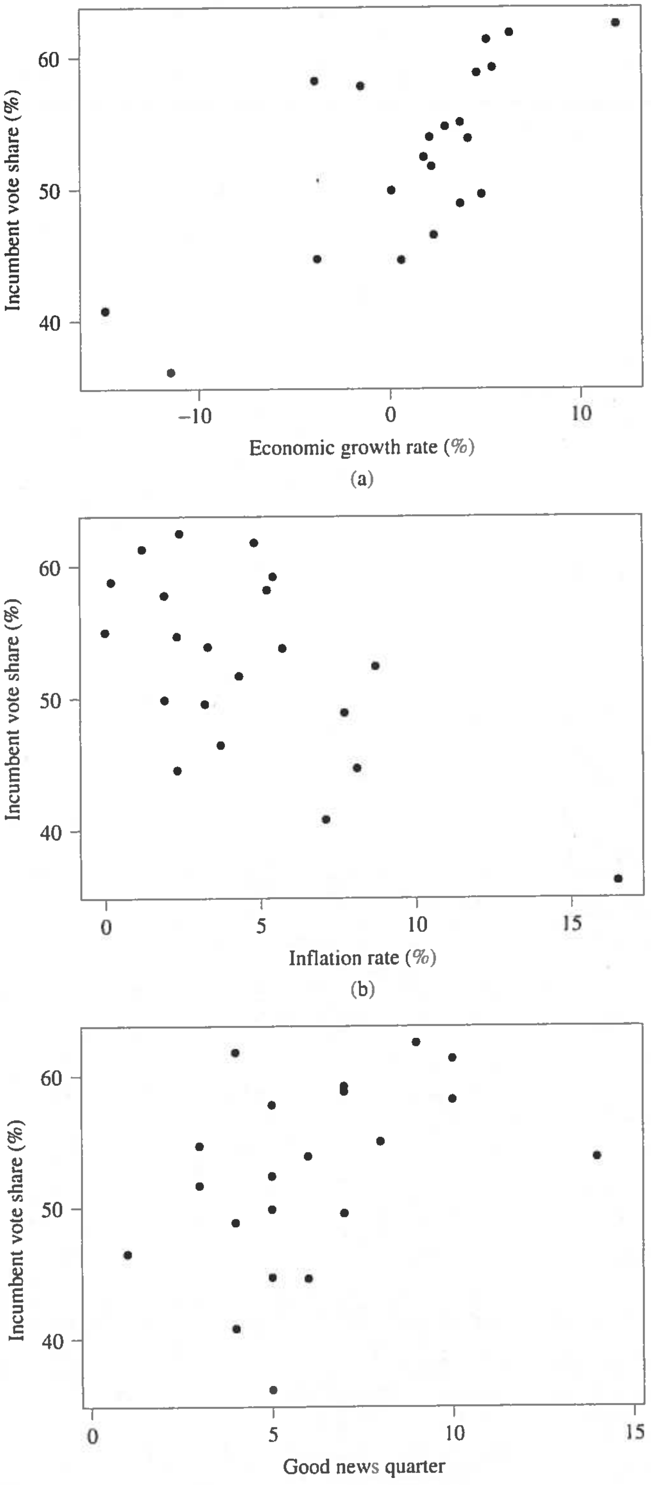
last one). Inflation and deflation (i.e., negative inflation) are treated symmetrically. Deflation is assumed to be just as bad as inflation.

- Good news is the number of quarters out of the 15 quarters prior to the election in which the per capita growth rate exceeds 3.2%.
- Duration takes the value 0 if the incumbent party has been in office for only one consecutive term. It takes the value 1.00 for two consecutive terms, 1.25 for three consecutive terms, 1.50 for four consecutive terms, and 1.75 for five consecutive terms.
- President running variable takes the value 1 if the president is running for reelection; otherwise, the value is 0. Vice presidents who become president during the administration are also given the value 1 if they run for president. The exception is Ford, who is given a 0 because he was not part of the 1972 ticket.
- Party takes the value 1 if the incumbent party is Democratic and 0 if the incumbent party is Republican. It measures the "pure party" effect.
- War: Because of World Wars I and II, the 1920, 1944, and 1948 elections are treated differently. Fair includes a war variable that takes the value 1 for years 1920, 1944, and 1948, and 0 otherwise.

Scatter plots of the incumbent vote share against growth rate, inflation rate, good news variable, and duration are shown in Figures 8.9a–8.9d. The incumbent vote share increases with growth rate and the number of good news quarters, and it decreases with inflation rate and duration. The patterns in these figures are expected because candidates of the incumbent party tend to be elected if the economy is strong. Figure 8.9e shows an interaction diagram of incumbent vote share against incumbent party and whether the incumbent president is running. The graph shows that for Democrats the incumbent effect is weak unless the president is running for reelection.

The summary results from fitting all possible regressions with the listed predictors and the interaction (product) between the party and the president running variables are given in Table 8.12. The model with the five predictors—growth rate, inflation rate, duration, party, and the war variable—leads to an acceptable C_p statistic. The model explains 90.3% of the variation. Although this may seem like a very good fit, the standard deviation of the unexplained component is 2.6%, indicating that the 95% prediction error margins are approximately ±5.2%. Also note that we are fitting a regression model with six parameters to just $n = 21$ cases. The detailed fitting results for this model are shown in Table 8.13. Growth rate (with positive coefficient) and inflation rate (with negative coefficient) have the expected signs. The negative effect of duration indicates that incumbency wears out if the incumbent party has occupied the White House for a long time. The negative coefficient for party indicates that Democrats in power tend to do worse. The war variable is significant and positive, indicating that in times of war voters appear to rally around the incumbent party.

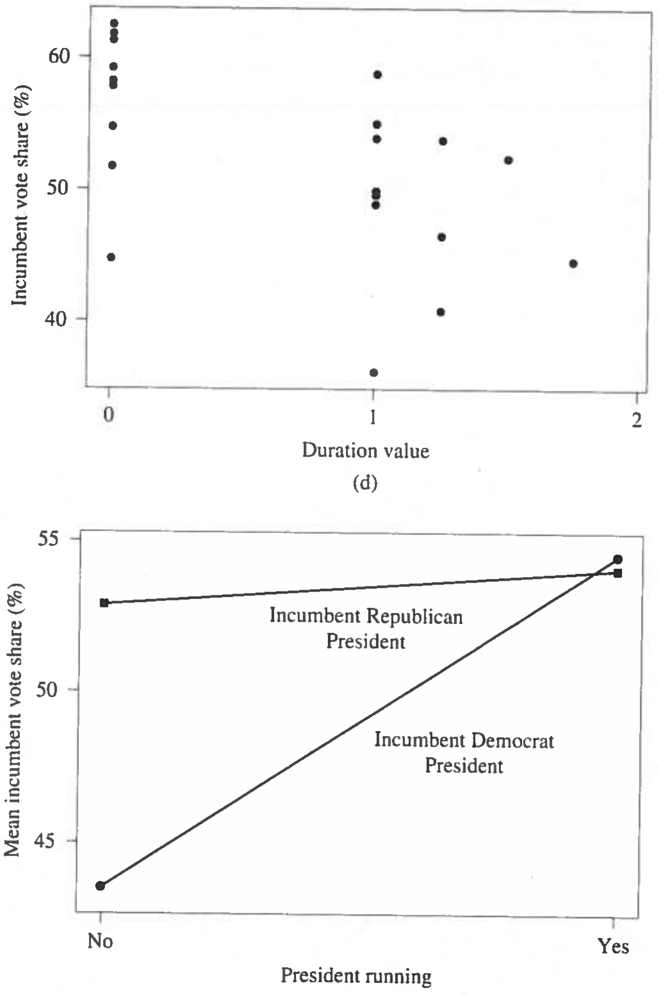
FIGURE 8.9 Scatter Plots of Incumbent Vote Share against (a) Economic Growth Rate, (b) Inflation Rate, (c) Good News Variable, and (d) Duration Value and (e) Interaction Diagram—Mean Incumbent Vote Share against Incumbent Party and President Running



(a)

(b)

FIGURE 8.9 (Continued)



(d)

(e)

An interaction term for incumbent party and incumbent president running was quite visible in the interaction plot in Figure 8.9e. However, the final model does not include such a term. This is not necessarily a contradiction. The interaction graph gives a marginal view of the data and "collapses" the data over all other variables except the two considered in the graph. It so happens that the five predictors in the model explain this interaction and that conditional on these five variables, the need for an interaction has disappeared.

The Durbin–Watson test is appropriate here because we estimate the regression on time series data. Its value, DW = 1.82, is quite close to 2 and does not indicate problems with serial correlation at lag 1. The variance inflation factors are unremarkable and smaller than commonly-used cutoff values; there is not an undue amount of multicollinearity. A scatter plot of the standardized residuals

TABLE 8.12 BEST SUBSETS REGRESSION OF INCUMBENT VOTE SHARE ON ALL LISTED PREDICTOR VARIABLES

Vars	R-Sq	R-Sq(adj)	C-p	S	Growth R	Inflation	Good News	Duration	War	President	Party	Pres*party
1	54.6	52.3	56.9	4.9290	X							
1	38.5	35.3	83.1	5.7375		X						
2	71.0	67.7	32.3	4.0526	X			X				
2	64.7	60.8	42.5	4.4665	X					X		
3	76.5	72.3	25.3	3.7521	X		X	X				
3	76.2	72.0	25.8	3.7777	X			X			X	
4	85.1	81.4	13.3	3.0804	X		X	X			X	
4	81.7	77.1	18.9	3.4161	X	X		X	X			
5	**90.3**	**87.0**	**6.8**	**2.5687**	**X**	**X**		**X**	**X**		**X**	
5	87.5	83.3	11.4	2.9126	X		X	X		X	X	
6	92.3	89.0	5.6	2.3703	X	X	X	X	X		X	
6	90.4	86.3	8.6	2.6403	X	X		X	X	X	X	
7	92.6	88.7	7.0	2.4025	X	X	X	X	X	X	X	
7	92.4	88.2	7.4	2.4457	X	X	X	X	X		X	X
8	92.6	87.7	9.0	2.5001	X	X	X	X	X	X	X	X

against the fitted values is shown in Figure 8.10a; the scatter plot shows no appreciable patterns. The normal probability plot of the standardized residuals is shown in Figure 8.10b; the linear pattern confirms the normality of the residuals. Dot diagrams of leverages and Cook's distances are given in Figure 8.10c; no case exerts an undue influence on the regression results.

Now that we have obtained an acceptable regression model, let us use it to predict the vote share of the incumbent party (represented by candidate Gore) in the 2000 presidential election. In 2000, the growth rate was 2.2%, inflation was at 1.7%, and the growth rates of 7 of the 15 quarters during the 1996–2000 term exceeded 3.2%. The duration variable has value 1 because the Democrats (the incumbent party) have been in power for two consecutive terms. The war indicator is zero and the party indicator is 1 for Democrat. The prediction from the regression equation is

$$\begin{aligned}\text{Incumbent vote share} &= 62.6 + 0.496\,\text{growth rate} - 1.18\,\text{inflation}\\ &\quad - 6.78\,\text{duration value} - 4.64\,\text{party} + 11.0\,\text{war}\end{aligned}$$

TABLE 8.13 ESTIMATION RESULTS FOR THE REGRESSION MODEL

```
The regression equation is
Incumbent Vote Share (%) = 62.6 + 0.496 Growth Rate − 1.18 Inflation
       − 6.78 Duration Value − 4.64 Party + 11.0 War

Predictor          Coef     SE Coef        T        P     VIF
Constant         62.636     1.728      36.25    0.000
Growth R          0.4963    0.1308      3.80    0.002     1.8
Inflatio         −1.1788    0.2673     −4.41    0.001     3.0
Duration         −6.780     1.089      −6.22    0.000     1.4
Party            −4.638     1.272      −3.65    0.002     1.3
War              10.979     2.672       4.11    0.001     2.8

S = 2.569        R-Sq = 90.3%   R-Sq(adj) = 87.0%

Analysis of Variance

Source              DF         SS        MS        F        P
Regression           5     918.73    183.75    27.85    0.000
Residual Error      15      98.97      6.60
Total               20    1017.71

Durbin-Watson statistic = 1.82
```

or

$$\begin{aligned}\text{Incumb vote share} &= 62.6+0.496(2.2)-1.18(1.7)-6.78(1)-4.64(1)+11.0(0)\\ &= 50.306\end{aligned}$$

The actual vote share for Gore in the election was 50.3%. The prediction is right on target. However, note that quite a bit of "luck" was involved. The standard error of the prediction error, calculated from Eq. (4.29) in Section 4.3, amounts to 2.79, and a 95% prediction interval for the incumbent vote share extends from 44.3 to 56.3. Here, we have used the 97.5 percentile of the *t* distribution with 15 degrees of freedom. A Gore result of, for example, 46 or 54% would not have been out of the ordinary either. Of course, proponents of prediction models tend to "tout their horns" if they get so close to the true value, and they credit this to the quality of their model. On the other hand, if they are not very close to the actual value, they point to the large prediction intervals to excuse their miss.

A Comment about the War Variable

Fair includes a war variable that takes the value 1 for years 1920, 1944, and 1948, and 0 otherwise. Fair argues that inflation and "good news" are irrelevant in these elections. He does not use the actual values of inflation and the good news variable for these three elections but instead sets them equal to zero. He claims that this implies that voters do not take into account past inflation and good news when deciding to vote during the three war-dominated periods.

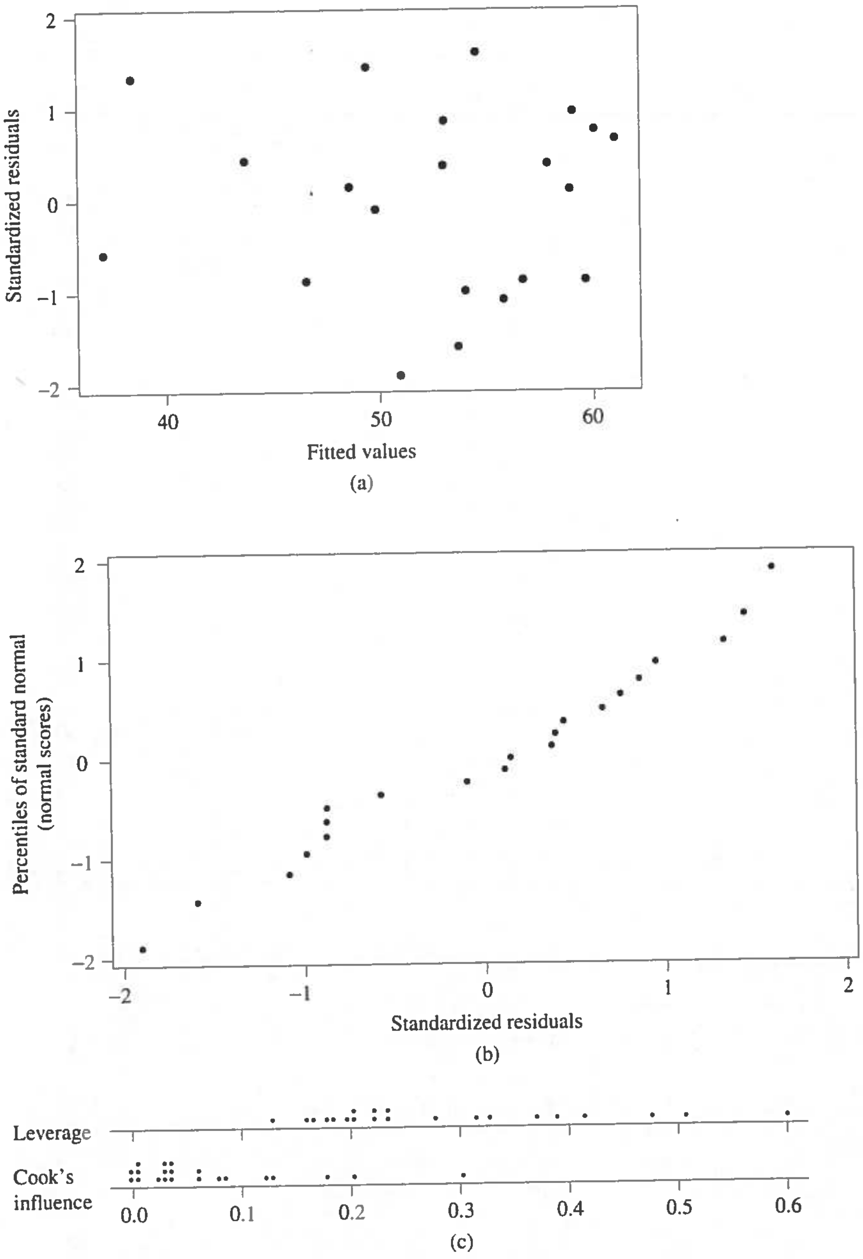


FIGURE 8.10
(a) Scatter plot of standardized residuals against fitted values, (b) normal probability plot of standardized residuals, and (c) dot plots of leverages and Cook's influence measures. Regarding Fig. 8.10(b). Note that the normal probability plot created by Minitab has switched the axes. Minitab plots the percentiles of the standard normal distribution (the normal scores) against the residuals, compared to the graphs in Chapter 6, which plot the residuals against the normal scores

However, this argument is not entirely correct. The decision to plug in zeros for inflation and good news is arbitrary. A better approach uses the original values of the regressor variables but allows for three separate indicator variables—one for each war year (1920, 1944, and 1948). This specification implies that the war years need separate adjustments, but it does not tie the adjustments to specific variables. We have estimated the regression model with the four explanatory variables in

Table 8.13 (growth rate, inflation, duration value, and party) and three separate war indicators. The results show that the war effects for 1920, 1944 and 1948 are about the same. We compare the fit of the full model with the three war indicators to the fit of the restricted model that uses the same indicator for all three war years. The F statistic for testing this restriction is $F = \dfrac{(98.97 - 84.91)/2}{84.91/(21-8)} = 1.08$. Its probability value $P(F \geq 1.08) = 0.362$ indicates that the three war years can be treated the same.

8.4.2 PREDICTION MODELS PROPOSED BY POLITICAL SCIENTISTS

Political scientists refer to Fair's model as an "economy incumbency model" because it uses economic variables to predict the incumbent vote share. They criticize his model for failing to incorporate measures of public opinion that are available prior to the election. There is a large literature in political science on the prediction of presidential elections. The book by Campbell and Garand (2000) gives a good summary of several competing models. This book also lists the data for all elections from 1948 through 1996 that were used in the estimation of these models as well as the model predictions for 2000. Data for five popular models (the models by Campbell, Abramowitz, Lewis-Beck and Tien, Holbrook, and Lockerbie) are included in Exercise 8.1. Here, we focus on the model by Michael S. Lewis-Beck and Charles Tien. They use data on the economy and survey responses to several questions on the Gallup poll of the last July prior to the election. Their data are listed in Table 8.14. Note that their model does not include the inflation rate, a significant factor in Fair's model.

- *Incumbent vote* is the percentage of the two-party vote received by the candidate of the president's party.

TABLE 8.14 INCUMBENT VOTE SHARE AND ITS DETERMINANTS[a]

Year	Incumbent Vote	July Popularity	Peace and Prosperity	Future Problems	Leading Indicators	GNP Change	Second Term
1948	52.37	39	*	46.67	0.00	2.42	0
1952	44.60	32	82.41	*	3.88	0.07	0
1956	57.76	69	122.84	56.90	0.00	0.26	1
1960	49.91	49	101.80	49.30	−3.08	1.42	0
1964	61.34	74	140.37	60.32	3.96	3.11	1
1968	49.60	40	85.94	46.55	0.00	2.88	0
1972	61.79	56	106.68	57.35	5.06	4.18	1
1976	48.95	45	80.40	34.85	6.07	2.33	0
1980	44.70	21	113.43	47.37	−5.67	−1.38	1
1984	59.17	52	104.51	51.32	0.00	3.95	1
1988	53.90	51	106.12	53.52	3.23	1.91	0
1992	46.55	32	94.81	45.95	2.48	1.46	0
1996	54.66	57	109.30	52.56	1.69	1.85	1

[a] The data are stored in the file **election(Beck&Tien)**.

- *July popularity* refers to the presidential popularity as measured by the Gallup poll in July before the election.
- *Peace and prosperity* is an index created by adding the percentage of two-party respondents who favored the incumbent party on keeping the United States out of war and on keeping the country prosperous (Gallup question).
- *Future problems* is the percentage of two-party respondents who favored the incumbent party on handling the country's most important problems (Gallup question).
- *Leading indicators* is the percentage change in the government's index of leading indicators during the first two quarters of the election year. It is set at zero if the change in one direction was not sustained for at least 3 months.
- *GNP change* is the nonannualized percentage change in GNP (constant dollars) from the fourth quarter of the year before the election to the second quarter of the election year.
- *Second term* is an indicator variable for a party's second consecutive term in the White House. It is coded 1 if the party is heading into its second term and 0 otherwise.

Scatter plots of incumbent vote against July popularity, response on questions regarding peace and prosperity, response on questions regarding future problems, leading indicators, and GNP change are shown in Figure 8.11. They show some relationships, although the associations are relatively weak.

The summary of all possible regressions is shown in Table 8.15. The model with July popularity, GNP change, and an indicator for a second term seems to give an acceptable fit with little bias; Mallow's C_p statistic (value of 4.9) is close to what one would expect for a good model (it should be approximately 4 for a model with three regressors). Detailed fitting results for this model are shown in Table 8.16. Since the four parameters in this model are estimated on only 13 cases, any conclusion from the estimated model must be made with great caution. The signs of the regression coefficients make sense; one can expect that popularity in July and an increase in GNP help increase the vote share. Running for a second term also increases the incumbent vote share. The Durbin–Watson statistic is unremarkable, indicating that there is no problem with serial correlation. The scatter plot of the standardized residuals in Figure 8.12, the normal probability plot of the standardized residuals (not shown), and the leverages and influence measures (not shown) do not reveal any serious problems with this model. The highest leverage originates from the 1980 election, with a leverage of 0.80 (which is approximately 2.5 times higher than the average leverage of $4/13 = 0.31$). However, Cook's influence measure shows that the influence of this case on the parameter estimates is not out of line. The standard deviation of the residuals amounts to approximately 1.7 percentage points, resulting in an approximate 95% prediction interval with half width ±3.4%.

FIGURE 8.11 Scatter Plots of Incumbent Vote against (a) July Popularity, (b) Peace and Prosperity Response, (c) Future Problems Response, (d) Index of Leading Indicators, and (e) GNP Change

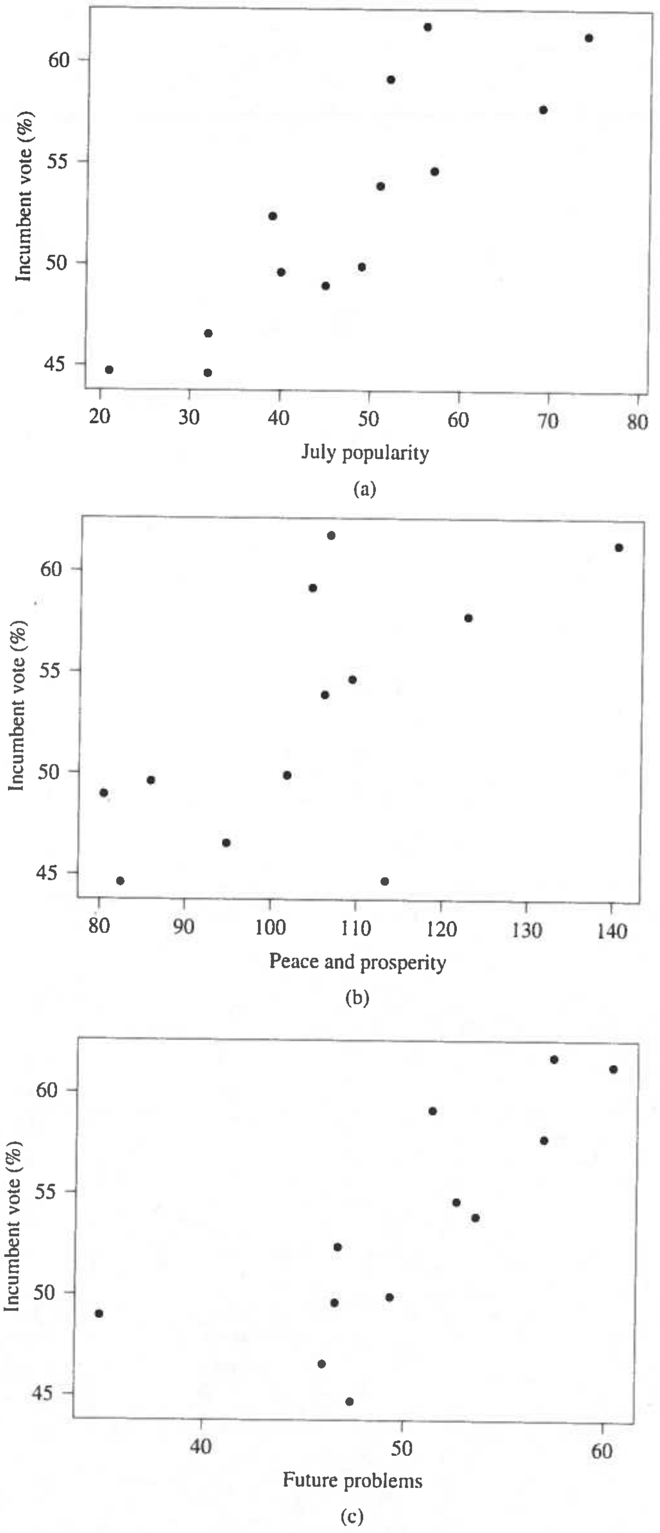
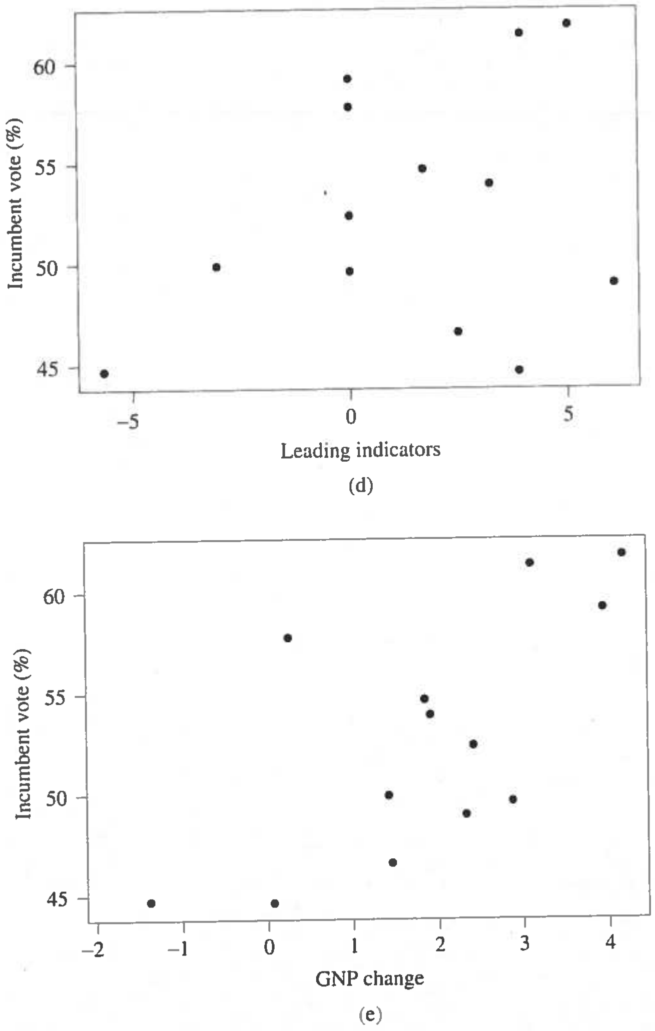


FIGURE 8.11 (Continued)



How does this model perform in predicting Gore's 2000 vote share? July's popularity for the incumbent party's candidate was 59, the GNP change from the fourth quarter of the year before the election to the second quarter of the election year was 2.52, and the indicator for a straight second term is 0. The prediction is given by

$$\text{Incumbent vote} = 38.3 + 0.202\,\text{July popularity} + 1.58\,\text{GNP change} \\ + 4.07\,\text{second term}$$

or

$$\text{Incumbent vote} = 38.3 + 0.202(59) + 1.58(2.52) + 4.07(0) = 54.2$$

with 95% prediction intervals extending from 49.8 to 58.6. Here, we have used the standard deviation of the prediction error in Eq. (4.29) and the 97.5 percentile of

TABLE 8.15 BEST SUBSETS REGRESSION OF INCUMBENT VOTE SHARE ON ALL LISTED PREDICTOR VARIABLES

Best Subsets Regression: Incumbent versus July Popular, Peace and Pr, . . .

```
Response is Incumbent Vote

11 cases used 2 cases contain missing values.

                                      J  P  F     G  S
                                      u  e  u  L  N  e
                                      l  a  t  e  P  c
                                      y  c  u  a     o
                                         e  r  d  C  n
                                      P     e  i  h  d
                                      o  a     n  a
Vars  R-Sq  R-Sq(adj)   C-p        S  p  n  P  g  n  T

   1  74.5       71.6  29.4   3.1800  X
   1  58.3       53.7  52.5   4.0647        X

   2  84.3       80.4  17.4   2.6457     X        X
   2  83.5       79.4  18.5   2.7109        X     X

   3  94.5       92.1   4.9   1.6756  X           X  X
   3  91.8       88.2   8.8   2.0475  X     X     X

   4  96.7       94.6   3.6   1.3904  X     X     X  X
   4  95.0       91.7   6.1   1.7248  X  X        X  X

   5  97.0       93.9   5.3   1.4721  X  X  X     X  X
   5  96.9       93.9   5.4   1.4779  X     X  X  X  X

   6  97.2       93.0   7.0   1.5803  X  X  X  X  X  X
```

the *t* distribution with 9 degrees of freedom. The prediction intervals are somewhat narrower than the ones obtained by Fair (which extended from 44.3 to 56.3). However, the point prediction misses the true value (50.3) by a larger amount than the point prediction in Fair's model.

8.4.3 CONCLUDING COMMENTS

How successful are these models in predicting the incumbent vote share in presidential elections? We notice that the standard deviation of the forecast errors amounts to approximately 2 to 3 percentage points. This implies that a 52 or 53% (point) prediction is not yet sufficient to justify concluding victory for the incumbent party.

The models are based on very few years of data, and they provide only a very rough glimpse of what can be expected at the next election. Although these models provide a useful yardstick of what can be expected, one should not bet one's fortune on their predictions.

TABLE 8.16 ESTIMATION RESULTS FOR THE MODEL WITH JULY POPULARITY, GNP CHANGE, AND AN INDICATOR FOR A SECOND TERM[a]

```
The regression equation is
Incumbent vote = 38.3 + 0.202 July Popularity + 1.58 GNP
          Change + 4.07 Second Term

Predictor              Coef    SE Coef          T        P    VIF
Constant             38.296      1.700      22.53    0.000
July Popularity     0.20156    0.04363       4.62    0.001    1.7
GNP Change           1.5821     0.3667       4.31    0.002    1.3
Second Term           4.066      1.108       3.67    0.005    1.3

S = 1.717           R-Sq = 93.8%        R-Sq(adj) = 91.8%

Analysis of Variance

Source               DF        SS           MS         F          P
Regression            3    402.49       134.16     45.52      0.000
Residual Error        9     26.53         2.95
Total                12    429.02

Durbin-Watson statistic = 2.04
```

[a] Observe that the summary statistics [s, R-Sq, R-Sq(adj)] are not identical to the summary statistics of the best subset regressions in Table 8.15. This has to do with the 2 years (cases) in which some of the regressor variables had missing values (peace and prosperity in 1948 and future problems in 1952). The best subset regression omits the entire case even if only one regressor variable has a missing value for that case, and hence the analysis in Table 8.15 is based on only 11(= 13 − 2) cases. The model in Table 8.16, on the other hand, uses all 13 cases because its regressor variables do not include the variables peace/prosperity and future problems, which had missing values on 2 years.

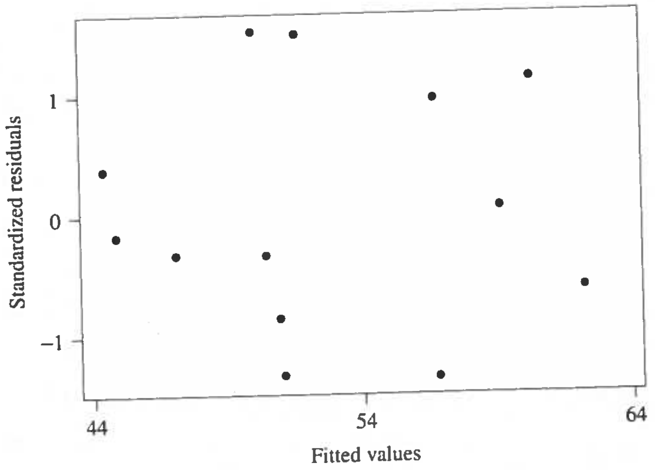


FIGURE 8.12 Scatter Plot of Standardized Residuals against Fitted Values

8.5 STUDENT PROJECTS LEADING TO ADDITIONAL CASE STUDIES

Courses on applied statistics such as regression should incorporate projects in which the student or the reader of this book selects the problems to study, gathers the data, analyzes the information using suitable computer software, and communicates the results in a report. Most textbooks on applied statistics, such as this book, include data. However, the included data sets are typically small to moderately sized, and the data are usually given to the reader as "numbers waiting to be analyzed." You are asked to analyze the data, run specific regression models, and find the best (appropriate) regression models that relate a specified response to certain specified explanatory variables. Such exercises are important because they teach the mechanical aspects of data analysis. However, they are incomplete in that they do not expose you to problem formulation and the difficulties of acquiring relevant data. The message communicated to you is that statistics starts after the data have been collected.

Each course should also contain projects that are relatively unstructured by the instructor. It is your own questions and data that provide the project structure. For such projects, you must generate an interesting problem, figure out what data to gather and where to get them, analyze the information, and put the analysis into words by writing a report on your findings. The data analysis is just one step in this process. The steps of formulating the problem, deciding what data to collect, managing the data acquisition process, and checking its integrity are often the more instructive and challenging parts of the project. This activity teaches you that statistics is more than just analyzing numbers. It engages you in research as you search for solutions to relevant and interesting questions. Active engagement is an important aspect of learning. Much is learned by being involved in writing survey questions and dealing first-hand with such issues as random and nonrandom sampling, nonresponse, and poorly designed questions. Also, when performing your own experiments you think about how to set up the experiment and you struggle with such issues as randomization, blocking, replication, and how your plans are impacted by practical issues. You learn to appreciate the difficulty of obtaining relevant data.

Several acronyms have been used to describe the various steps of the problem-solving cycle. For example, the five-step cycle "DMAIC" consists of defining the problem (D), measure (M), analyze (A), improve (I), and control (C). Deming talks about the Deming–Shewhart wheel PDSA, where P stands for plan, D for do, S for study, and A for act.

It is also very worthwhile to carry out such projects in groups, and we recommend that you form study groups. Our experience shows that group projects work well with committed students who are eager to learn. In such situations, the sharing of information and expertise is very beneficial to learning. Of course, group projects are likely to fail if motivation and group spirit are lacking.

The project output should be a report. The report should start with a short, concise executive summary that describes the problem and the main findings of your study. The write-up should discuss the motivation behind the project, describe the data and the way they were obtained, and discuss the statistical analysis. Furthermore, the report should discuss the appropriateness of the analysis and should reflect on any possible shortcomings. The findings must be interpreted, and the conclusions and implications of the study should be spelled out clearly. Relevant statistical tables and computer output should be put in an appendix. A listing of the raw data (only a subset of the data for large data sets) and a summary of the data definitions and the data sources should be included.

Writing a good research paper is a challenging task. The topic may be interesting and the statistical analysis may be competent, but the project may still fall short due to poor writing. The writing must be well organized and grammatically correct, and it must keep the reader's interest. Fortunately, help on writing is available online because many universities have developed excellent online writing resources. Here are a few particularly good links:

- Purdue University Online Writing Lab: http://owl.english.purdue.edu
- University of Florida—The Reading and Writing Center: http://web.cwoc.ufl.edu/owl
- University of Wisconsin Writing Center: http://www.wisc.edu/writing
- University of Victoria (Canada) Writer's Guide: http://web.uvic.ca/wguide
- Guide to Writing and Grammar: http://webster.commnet.edu/grammar/index.htm

These resources discuss the general structure of research papers and they contain many useful suggestions. They discuss how to cite the work of others and how to avoid plagiarism. They give strategies for clarifying logic and avoiding "deadly" sins, and they help with grammar and style.

An oral presentation, with a subsequent discussion that involves all students in the class, is also very useful. It helps you practice your oral presentation skills, and it teaches you how to respond to questions and criticism. Furthermore, a session for oral presentations exposes you to a wide variety of other topics that have been considered by the various groups.

Your instructor should give you examples of previous projects—good projects as well as bad ones—and should provide suggestions for appropriate topics. Many of the data sets in this book in fact originated from such projects. Successful past projects in our courses have examined the following issues (see Ledolter, 1995):

- "Success" at the university: What are the relationships between college GPA and high school GPA, number of hours studied, etc.? Does gender play a role? Does it help (hurt) your GPA if one lives in a sorority/fraternity? What about the effect of drinking or smoking?

In projects such as these, you have to construct questionnaires and survey your fellow students. You need to think about sampling issues and how to draw representative samples.

- Relationships between preferences (such as listening or buying preferences) and demographic characteristics (such as gender and occupation).
- Sports statistics (football, basketball, and baseball): Find the key variables that explain individual or team performance. Projects that attempt to explain the salaries of baseball players are always fun. Compare the salaries of pitchers, infielders, etc.
- Relationships between CEO compensation and performance; relationship between professor salaries and performance: Much of the needed information can be obtained from Web sites. Salaries at most public universities are readily available. Who wouldn't be interested in finding out the salary of one's professor and learning whether he or she "deserves" it?
- Sales forecasts; marketing applications; applications of statistics to finance and portfolio selection; tracking the performance of investment strategies; predicting economic indicators.
- Effects of legislation on society: For example, investigate the impact of changes in the maximum speed limit on the number and severity of traffic accidents, or study the impact of the motorcycle helmet law on motor cycle accidents that resulted in severe head injuries.
- Experiments with paper helicopters, catapults, rubber balls, and sticky pads. For example, vary certain design characteristics on a simple paper helicopter and study how the settings affect the flying time of the helicopter. The paper by Hunter (1977) is a good reference if you want to conduct such experiments.

Data for projects may be obtained from Internet sources, company data, surveys, statistical reference books, or experiments that you carry out yourself. The *Statistical Data Abstract of the United States* (U.S. Bureau of the Census, U.S. Government Printing Office, Washington, DC, 1879–) is an annual compendium of summary statistics on the political, social, industrial, and economic life of the United States. Four major monthly governmental periodicals provide the majority of the current statistics available on the economy and its operation. The *Survey of Current Business* contains approximately 2500 statistical series on income, expenditures, production, and prices of commodities. Historical figures for the statistical data published in the *Survey of Current Business* are available in a supplement titled *Business Statistics,* published in odd-numbered years. A second source is the *Federal Reserve Bulletin,* which publishes data with emphasis on financial statistics. The third major governmental publication is the *Monthly Labor Review,* which publishes data on work and labor conditions, wage rates, consumer price indices, and the like. The fourth is the *Business Conditions Digest,* which contains several hundred economic time series in a form convenient for forecasters and business analysts.

EXERCISES

8.1. Incumbent Vote Share in Presidential Elections

The model by Campbell includes two predictors: the percentage of support for the inparty candidate in the preference poll conducted by Gallup in early September of the election year and the second-quarter (of the election year) rate of growth in GDP.

Abramowitz models the incumbent vote share as a function of the president's approval rating in the Gallup poll of early July and the annual growth rate of real GDP during the first two quarters of the election year (which are released in August of the election year).

The model developed by Holbrook uses the following predictors: the president's approval rating, a measure of retrospective personal finances indicating whether one is better or worse off financially now than a year ago, and an indicator measuring tenure in office.

Lockerbie uses two measures of change in disposable income, a measure of anticipated financial well-being a year from now and the number of years the incumbent party has controlled the White House.

The data sets and a description of the variables are given below. The data sets are stored in the files **campbell, abramowitz, holbrook**, and **lockerbie**.

a. Use the data to estimate the various regression models. Assess the significance of the estimated regression coefficients. Investigate whether all listed regressor variables are needed. Simplify the regression models if possible.

b. Check the models for violations of the regression assumptions. Identify high-leverage values and influential observations. Also, check for autocorrelation among the residuals.

c. Discuss how these models can be used for prediction. Discuss how these models can be used for scenario forecasts. That is, develop scenarios for the predictor variables under which the incumbent party candidate has a good chance of holding on to the White House. Your analysis should also incorporate the uncertainty.

d. Discuss in more detail the difference between "fitting models" and "forecasting." One could drop a particular year (case) from the regression, estimate the model without the data from that year, and predict the incumbent share for the year that one has omitted. Why would this be different (better or worse) than using all observations and looking at the residuals? Discuss.

e. Summarize the performance of these models, and compare them to Fair's model discussed in Section 8.4.1. Are these worthwhile models?

Data Used by Campbell

Year	Incumbent Party Vote	September Trial Heat	GDP Growth Rate
1948	52.32	45.61	0.91
1952	44.59	42.11	0.27
1956	57.75	55.91	0.64
1960	49.92	50.54	−0.26
1964	61.34	69.15	0.81
1968	49.60	41.89	1.63
1972	61.79	62.89	1.73
1976	48.95	40.00	1.17
1980	44.70	48.72	−2.43
1984	59.17	60.22	1.79
1988	53.90	54.44	0.79
1992	46.55	41.94	0.35
1996	54.74	60.67	1.04

Incumbent vote: Percentage of the two-party vote received by the candidate of the president's party.

September trial heat: Two-party percentage of support for the in-party candidate in the preference poll conducted by Gallup in early September of the election year.

GDP growth rate: Second quarter rate of growth (nonannualized) in the GDP.

Data Used by Abramowitz

Year	Incumbent Vote	Term	Presidential Popularity	GDP Growth Rate
1948	52.3	1	39	1.8
1952	44.6	1	32	−0.2
1956	57.8	0	69	0.5
1960	49.9	1	49	0.4
1964	61.3	0	74	1.1
1968	49.6	1	40	1.3
1972	61.8	0	56	2.2
1976	48.9	1	45	1.1
1980	44.7	0	21	−2.3
1984	59.2	0	52	1.8
1988	53.9	1	51	0.8
1992	46.6	1	31	0.4
1996	54.6	0	56	1.6

Incumbent vote: Percentage of the two-party vote received by the candidate of the president's party.

Term: Binary variable coded 1 if the president's party has held the White House for 8 years or longer and 0 otherwise.

Presidential popularity: President's approval rating in the Gallup poll in early June.

GDP growth rate: Annual growth rate of the real GDP during the first two quarters of the election year.

Data Used by Holbrook

Year	Incumbent Vote	Presidential Popularity	Personal Finances	Tenure in Office
1948	52.4	37.5	116	1
1952	44.6	30.0	94	1
1956	57.8	69.0	110	0
1960	49.9	62.6	108	1
1964	61.3	74.3	120	0
1968	49.6	42.2	114	1
1972	61.8	59.0	129	0
1976	48.9	46.6	103	1
1980	44.7	36.8	79	0
1984	59.2	53.8	121	0
1988	53.9	49.4	111	1
1992	46.5	39.0	97	1
1996	54.6	54.0	114	0

Incumbent vote: Percentage of the two-party vote received by the candidate of the president's party.

Presidential popularity: Average percentage of the public (responding to Gallup polls in the second quarter) who said that they approved of the way the president is handling his job.

Personal finances: Measure of retrospective personal finances based on responses to the Survey of Consumers' question, "Would you say that you (and your family living here) are better off or worse off financially than you were a year ago?"

Tenure in office: Binary variable coded 1 for presidential candidates whose party has held the White House for two or more terms and 0 for candidates whose party has held the White House for one term.

Data Used by Lockerbie

Year	Incumbent Vote	Disposable Income 1	Disposable Income 2	Next Year Better	Tenure
1956	57.80	1.60	3.70	36.00	4
1960	49.90	1.65	2.26	35.00	8
1964	61.30	4.74	1.78	37.00	4
1968	49.60	2.91	3.31	33.00	8
1972	61.79	1.61	2.39	38.00	4
1976	48.90	2.03	0.68	32.00	8
1980	44.70	−1.26	1.59	26.00	4
1984	59.20	2.67	1.67	37.33	4

				Next	
	Incumbent	Disposable	Disposable	Year	
Year	Vote	Income 1	Income 2	Better	Tenure
1988	53.90	1.73	−0.05	36.33	8
1992	46.50	2.10	−1.03	36.33	12
1996	54.74	2.07	2.33	35.67	4

Incumbent vote: Percentage of the two-party vote received by the candidate of the president's party.

Disposable Income 1: Change in per capita real disposable income from the second quarter of the year prior to the election to the second quarter of the election year.

Disposable Income 2: Change in per capita real disposable income from 2 years prior to the election to the year immediately prior to the election.

Next year better: A component of the Index of Consumer Sentiment. This variable is based on the following question: "Now looking ahead—Do you think that a year from now you (and your family living here) will be better off financially, or worse off, or just about the same as now?"

Tenure: Number of years a party has controlled the White House.

8.2. Height and Weight of Boys and of their Mothers and Fathers

This data set is taken from a larger study examining the associations between childhood treatment with methylphenidate (MPH) and adult height and weight [Kramer, J. R., Loney, J., Ponto, L. B., Roberts, M. A., and Grossman, S. Predictors of adult height and weight in boys treated with methylphenidate for childhood behavior problems. *Journal of the American Academy of Child and Adolescent Psychiatry,* 39, 517–524, 2000]. The 93 boys in this study were 6 to 13 years of age, had behavior problems, were referred to a child psychiatry outpatient clinic, treated clinically with MPH for an average of 36 months, and reevaluated between ages 15 and 19. The information for the first five boys and the data description are listed here; the complete data set is contained in the file **height&weight**.

Part 1 In part 1 of this exercise, we ask you to examine the relationship between the height (and weight) of the boys and their age. Pediatricians have access to elaborate nonlinear growth curves, that they obtain from measurements on a very large number of children. With a small data set such as this, it may not be possible to fit elaborate models. However, it may be feasible to obtain useful linear approximations.

a. Consider the measurements taken at referral. Model the relationship between the height and the child's age at referral. Model the relationship between the weight and the child's age. Investigate whether or not the weight at birth has some additional explanatory power.

b. Consider the measurements taken at the follow-up visit. Model the relationship between the height and the child's age at the follow-up visit. Model the relationship between the weight and the child's age.

c. Combine the data set, and use all measurements on weight and height. Ignore the fact that the two measurements on weight (and height) are taken on the same child. Model the relationship between the height and the child's age. Model the relationship between the weight and the child's age.

d. Discuss your findings.

Part 2 The data set also lists the height and weight for mothers and fathers.

a. Investigate, for mothers and fathers separately, relationships among their weight and height.

b. Investigate whether there are relationships between mother's and father's heights and mother's and father's weights. Can you claim that "thin" tends to be attracted by "thin"?

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Col 11
112	56.5	75.0	199	72	165	9.38	64	135	75	230
89	50.0	64.0	184	62	101	8.41	68	127	70	181
100	53.0	85.0	162	63	92	7.40	68	250	69	190
133	56.8	91.8	206	67	147	7.00	65	138	70	181
111	52.5	67.3	187	69	139	8.88	66	179	72	210

Col 1: Age at referral (months)

Col 2: Height at referral (in.)

Col 3: Weight at referral (lb)

Col 4: Age at follow-up (months)

Col 5: Height at follow-up (in.)

Col 6: Weight at follow-up (lb)

Col 7: Birth weight (lb)

Col 8: Height of mother (in.)

Col 9: Weight of mother (lb)

Col 10: Height of father (in.)

Col 11: Weight of father (lb)

8.3. Modeling Softdrink Sales:

The data in file **softdrink** represent weekly sales and prices of two competing products, brand P and brand C. The sales are recorded in ounces, whereas prices are given in dollars per ounce. Logarithms of the sales of 12-packs for both brands (lnSalesP12 and lnSalesC12) are given, as well as the logarithms of prices for 6-, 12-, and 24-packs (lnPriceP6, lnPriceP12, and lnPriceP24 for brand P and lnPriceC6, lnPriceC12, and lnPriceC24 for brand C).

a. Analysis for brand P
Price effects on sales are measured by regressing the logarithm of sales on the logarithm of prices. Consider the regression model

$$\begin{aligned}\textbf{M1:}\quad \text{lnSalesP12}_t = {} & \beta_0 + \beta_1 \ln \text{PriceP6}_t \\ & + \beta_2 \ln \text{PriceP12}_t \\ & + \beta_3 \ln \text{PriceP24}_t + \varepsilon_t\end{aligned}$$

The slope coefficients in the log/log model represent price elasticities (see the discussion in Section 6.5). A 1% change in the product's own price translates into a percentage change in sales of magnitude β_2. We expect this elasticity to be negative because sales decrease when prices are increased. On the other hand, the price elasticities for 6- and 24 packs, β_1 and β_3, should be positive. Because of product substitution within the same brand family, we expect increased sales of 12-packs when prices of the other pack sizes are increased.

Estimate the model M1 using regression software of your choice. Interpret the results. Check whether the elasticities have the expected signs. Discuss whether sales of 12-packs are more responsive to price changes in 24-packs than to price changes in 6-packs.

Check the model assumptions, in particular investigate the autocorrelations of the residuals. You will find that there is some autocorrelation in the residuals, and you will revisit this data set in Exercise 10.8. Look for residual outliers, leverage, and Cook's distance (you will find several large ones).

b. Analysis for brand C
Apply the analysis in (a) to the sales of brand C and consider the model

$$\begin{aligned}\textbf{M2:}\quad \text{lnSalesC12}_t = {} & \alpha_0 + \alpha_1 \ln \text{PriceC6}_t \\ & + \alpha_2 \ln \text{PriceC12}_t \\ & + \alpha_3 \ln \text{PriceC24}_t + \varepsilon_t\end{aligned}$$

Estimate the model. Interpret the estimates. Check the model. Discuss whether or not the results for 12-packs of brand C in model M2 are similar to the results for 12-packs of brand P in model M1.

c. Analysis incorporating prices of both brands

Consider models that do not just include the prices of the considered brand but also the prices of the competing product. Estimate the model

$$\begin{aligned}\textbf{M3:}\quad \text{lnSalesP12}_t = {} & \beta_0 + \beta_1 \ln \text{PriceP6}_t \\ & + \beta_2 \ln \text{PriceP12}_t \\ & + \beta_3 \ln \text{PriceP24}_t \\ & + \alpha_1 \ln \text{PriceC6}_t \\ & + \alpha_2 \ln \text{PriceC12}_t \\ & + \alpha_3 \ln \text{PriceC24}_t + \varepsilon_t\end{aligned}$$

Interpret the coefficients in this log/log model. Discuss whether the results are much of an improvement over the simpler models M1 and M2. You will find that the price elasticities of the competing brand are essentially zero and not overly significant. This means that sales of brand P are mostly driven by the prices of brand P and not by the prices of brand C.

Repeat the analysis by regressing the log sales of 12-packs of brand C on all six prices.

d. Additional models
Consider the ratio SalesP12/SalesC12 as the response. This is equivalent to analyzing $S/(1-S)$, where $S =$ SalesP12/[SalesP12 + SalesC12] is the market share of brand P. Estimate the model

$$\begin{aligned}\textbf{M4:}\quad & \ln(\text{SalesP12}_t/\text{SalesC12}_t) \\ & = \beta_0 + \beta_1 \ln \text{PriceP6}_t \\ & \quad + \beta_2 \ln \text{PriceP12}_t \\ & \quad + \beta_3 \ln \text{PriceP24}_t \\ & \quad + \alpha_1 \ln \text{PriceC6}_t \\ & \quad + \alpha_2 \ln \text{PriceC12}_t \\ & \quad + \alpha_3 \ln \text{PriceC24}_t + \varepsilon_t\end{aligned}$$

Confirm that the elasticities have the expected signs. Confirm that the response decreases with increasing 12-pack price of brand P and decreasing 12-pack price of brand C. The signs of the two price coefficients are different, but their magnitude is approximately the same. Confirm that the same can be said for the price coefficients of the other pack sizes, except that now the signs are reversed. These regression results lead us to a model that contains logarithms of price ratios (i.e., differences of the log prices) as explanatory variables. Consider the model

$$\begin{aligned}\textbf{M5:}\quad & \ln(\text{SalesP12}_t/\text{SalesC12}_t) \\ & = \beta_0 + \beta_1 \ln \text{PriceRatio6}_t \\ & \quad + \beta_2 \ln \text{PriceRatio12}_t \\ & \quad + \beta_3 \ln \text{PriceRatio24}_t + \varepsilon_t\end{aligned}$$

Interpret the model and compare it to model M4. You will find that the R^2 is not much worse, but the model may be easier to interpret.

8.4. Complete a project as outlined in Section 8.5.