17.5 Review Exercises

1. When you are investigating the relationship between two continuous random variables, why is it important to create a scatter plot of the data?
2. What are the strengths and limitations of Pearson's correlation coefficient?
3. How does Spearman's rank correlation differ from the Pearson correlation?
4. If a test of hypothesis indicates that the correlation between two random variables is not significantly different from 0, does this necessarily imply that the variables are independent? Explain.
5. In a study conducted in Italy, 10 patients with hypertriglyceridemia were placed on a low-fat, high-carbohydrate diet. Before the start of the diet, cholesterol and triglyceride measurements were recorded for each subject [3].

Patient	Cholesterol Level (mmol/l)	Triglyceride Level (mmol/l)
1	5.12	2.30
2	6.18	2.54
3	6.77	2.95
4	6.65	3.77
5	6.36	4.18
6	5.90	5.31
7	5.48	5.53
8	6.02	8.83
9	10.34	9.48
10	8.51	14.20

   (a) Construct a two-way scatter plot for these data.
   (b) Does there appear to be any evidence of a linear relationship between cholesterol and triglyceride levels prior to the diet?
   (c) Compute r, the Pearson correlation coefficient.
   (d) At the 0.05 level of significance, test the null hypothesis that the population correlation ρ is equal to 0. What do you conclude?
   (e) Calculate r_s, the Spearman rank correlation coefficient.
   (f) How does the value of r_s compare to that of r?
   (g) Using r_s, again test the null hypothesis that the population correlation is equal to 0. What do you conclude?

6. Thirty-five patients with ischemic heart disease, a suppression of blood flow to the heart, took part in a series of tests designed to evaluate the perception of pain. In one part of the study, the patients exercised until they experienced angina, or chest pain; time until the onset of angina and the duration of the attack were recorded. The data are saved on your disk in the file `ischemic` [4] (Appendix B, Table B.6). Time to angina in seconds is saved under the variable name `time`; the duration of angina, also in seconds, is saved under the name `duration`.
   (a) Create a two-way scatter plot for these data.
   (b) In the population of patients with ischemic heart disease, does there appear to be any evidence of a linear relationship between time to angina and the duration of the attack?
   (c) Calculate Pearson's correlation coefficient.
   (d) Does the duration of angina tend to increase or decrease as time to angina increases?
   (e) Test the null hypothesis

   $$H_0: \rho = 0.$$

   What do you conclude?
   (f) Compute Spearman's rank correlation.
   (g) Using r_s, again test the null hypothesis that the population correlation is equal to 0. What do you conclude?

7. The data set `lowbwt` contains information collected for a sample of 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts [5] (Appendix B, Table B.7). Measurements of systolic blood pressure are saved under the variable name `sbp`, and values of the Apgar score recorded five minutes after birth—an index of neonatal asphyxia or oxygen deprivation—are saved under the name `apgar5`. The Apgar score is an ordinal random variable that takes values between 0 and 10.
   (a) Estimate the correlation of the random variables systolic blood pressure and five-minute Apgar score for this population of low birth weight infants.
   (b) Does Apgar score tend to increase or decrease as systolic blood pressure increases?
   (c) Test the null hypothesis

   $$H_0: \rho = 0.$$

   What do you conclude?

8. Suppose that you are interested in determining whether a relationship exists between the fluoride content in a public water supply and the dental caries experience of children using this water. Data from a study examining 7257 children in 21 cities are saved on your disk in the file `water` [6] (Appendix B, Table B.21). The fluoride content of the public water supply in each city, measured in parts per million, is saved under the variable name `fluoride`; the number of dental caries per 100 children examined is saved under the name `caries`. The total dental caries experience is obtained by summing the numbers of filled teeth, teeth with untreated dental caries, teeth requiring extraction, and missing teeth.
   (a) Construct a two-way scatter plot for these data.
   (b) What is the correlation between the number of dental caries per 100 children and the fluoride content of the water?
   (c) Is this correlation significantly different from 0?

(d) For the 21 cities in the study, the highest fluoride content in a given water supply is 2.6 ppm. If you were to increase the fluoride content of the water to more than 4 ppm, do you believe that the number of dental caries per 100 children would decrease?

9. One of the functions of the Federation of State Medical Boards is to collect data summarizing disciplinary actions taken against nonfederal physicians by medical licensing boards. Serious actions include license revocations, suspensions, and probations. For each of the years 1991 through 1995, the number of serious actions per 1000 doctors was ranked by state from highest to lowest. The ranks are contained in a data set called `actions` [7] (Appendix B, Table B.22); the ranks for 1991 are saved under the variable name `rank91`, those for 1992 under `rank92`, and so on.
(a) Which states have the highest rates of serious actions in each of the five years 1991 through 1995? Which states have the lowest rates?
(b) Construct a two-way scatter plot for the ranks of disciplinary actions in 1992 versus the ranks in 1991.
(c) Does there appear to be a relationship between these two quantities?
(d) Calculate the correlation of the two sets of ranks.
(e) Is this correlation significantly different from 0? What do you conclude?
(f) Calculate the correlations of the ranks in 1991 and those in 1993; those in 1991 and 1994; and those in 1991 and 1995. What happens to the magnitude of the correlation as the years being compared get further apart?
(g) Is each of these three correlations significantly different from 0?
(h) Do you believe that all states are equally strict in taking disciplinary action against physicians?

Bibliography

[1] United Nations Children's Fund, *The State of the World's Children 1994,* New York: Oxford University Press.

[2] Snedecor, G. W., and Cochran, W. G., *Statistical Methods,* Ames, Iowa: The Iowa State University Press, 1980.

[3] Cominacini, L., Zocca, I., Garbin, U., Davoli, A., Compri, R., Brunetti, L., and Bosello, O., "Long-Term Effect of a Low-Fat, High-Carbohydrate Diet on Plasma Lipids of Patients Affected by Familial Endogenous Hypertriglyceridemia," *American Journal of Clinical Nutrition,* Volume 48, July 1988, 57–65.

[4] Miller, P. F., Sheps, D. S., Bragdon, E. E., Herbst, M. C., Dalton, J. L., Hinderliter, A. L., Koch, G. G., Maixner, W., and Ekelund, L. G., "Aging and Pain Perception in Ischemic Heart Disease," *American Heart Journal,* Volume 120, July 1990, 22–30.

[5] Leviton, A., Fenton, T., Kuban, K. C. K., and Pagano, M., "Labor and Delivery Characteristics and the Risk of Germinal Matrix Hemorrhage in Low Birth Weight Infants," *Journal of Child Neurology,* Volume 6, October 1991, 35–40.

[6] Dean, H. T., Arnold, F. A., and Elvove, E., "Domestic Water and Dental Caries," *Public Health Reports,* Volume 57, August 7, 1942, 1155–1179.

[7] Public Citizen Health Research Group, "Ranking of Doctor Disciplinary Actions by State Medical Licensing Boards—1992," *Health Letter,* Volume 9, August 1993, 4–5.

18

Simple Linear Regression

Like correlation analysis, *simple linear regression* is a technique that is used to explore the nature of the relationship between two continuous random variables. The primary difference between these two analytical methods is that regression enables us to investigate the change in one variable, called the *response,* which corresponds to a given change in the other, known as the *explanatory variable.* Correlation analysis makes no such distinction; the two variables involved are treated symmetrically. The ultimate objective of regression analysis is to predict or estimate the value of the response that is associated with a fixed value of the explanatory variable.

An example of a situation in which regression analysis might be preferred to correlation is illustrated by the pediatric growth charts in Figures 18.1 and 18.2. Among children of both sexes, head circumference appears to increase linearly between the ages of 2 and 18 years. Rather than quantifying the strength of this association, we might be interested in predicting the change in head circumference that corresponds to a given change in age. In this case, head circumference is the response, and age is the explanatory variable. An understanding of their relationship helps parents and pediatricians to monitor growth and detect possible cases of macrocephaly and microcephaly.

18.1 Regression Concepts

Suppose that we are interested in the probability distribution of a continuous random variable Y. The outcomes of Y, denoted y, are the head circumference measurements in centimeters for the population of low birth weight infants—defined as those weighing less than 1500 grams—born in two teaching hospitals in Boston, Massachusetts [1]. We are told that the mean head circumference for the infants in this population is

$$\mu_y = 27 \text{ cm}$$

or

(31.66, 42.20).

Because of the extra source of variability, this interval is quite a bit wider than the 95% confidence interval for the predicted mean value of y.

After generating the least-squares regression line, we might wish to have some idea about how well this model fits the observed data. One way to evaluate the fit is to examine the coefficient of determination. In Table 18.1, the value of R^2 is displayed in the top portion of the output on the right-hand side. For the simple linear regression of length on gestational age, the coefficient of determination is

$$R^2 = 0.4559;$$

this means that approximately 45.59% of the variability among the observed values of length is explained by the linear relationship between length and gestational age. The remaining 54.41% is not explained by this relationship. The line of output directly below R^2, labeled Adj R-square (adjusted R^2), will be discussed in the following chapter.

A second technique for evaluating the fit of the least-squares regression line to the sample data involves looking at a two-way scatter plot of the residuals versus the predicted values of length. The residuals are obtained by subtracting the fitted values $\hat{y}_i$ from the actual observations y_i; the calculations may be performed using a computer package. Table 18.3 shows the observed and predicted values of length, along with the differences between them, for the first 10 infants in the sample. Figure 18.16 is a scatter plot of the points $(\hat{y}_i, e_i)$ for all 100 low birth weight infants.

Looking at the residual plot, we see that there is one point with a particularly low residual that appears to be an outlier. We might try removing this point, fitting a new line, and then comparing the two models to see how much of an effect the point has on the estimated regression coefficients. However, there is no evidence that the assumption of homoscedasticity has been violated, or that a transformation of either the response or the explanatory variable is necessary.

TABLE 18.3
Residuals for the first 10 infants in the sample

Length	Predicted Length $\hat{y}$	Residual
41	36.92467	4.075324
40	38.82788	1.172117
38	40.73109	−2.731091
38	38.82788	−0.827883
38	37.87628	0.123720
32	33.11826	−1.118262
33	35.02147	−2.021469
38	36.92467	1.075324
30	35.97307	−5.973073
34	36.92467	−2.924676

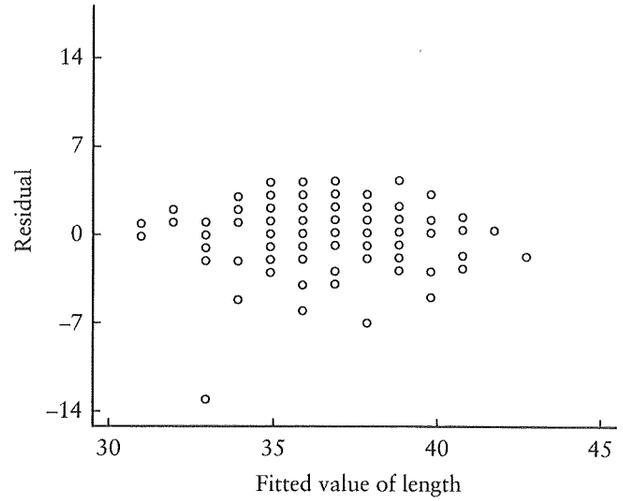

FIGURE 18.16
Residuals versus fitted values of length for a sample of 100 low birth weight infants

18.5 Review Exercises

1. What is the main distinction between correlation analysis and simple linear regression?
2. What assumptions do you make when using the method of least squares to estimate a population regression line?
3. Explain the least-squares criterion for obtaining estimates of the regression coefficients.
4. Why is it dangerous to extrapolate an estimated linear regression line outside the range of the observed data values?
5. Given a specified value of the explanatory variable, how does a confidence interval constructed for the mean of the response differ from a prediction interval constructed for a new, individual value? Explain.
6. Why might you need to consider transforming either the response or the explanatory variable when fitting a simple linear regression model? How is the circle of powers used in this situation?
7. For a given sample of data, how can a two-way scatter plot of the residuals versus the fitted values of the response be used to evaluate the fit of a least-squares regression line?

8. Figure 18.17 displays a two-way scatter plot of CBVR—the response of cerebral blood volume in the brain to changes in carbon dioxide tension in the arteries—versus gestational age for a sample of 17 newborn infants [4]. The graph also shows the fitted least-squares regression line for these data. The investigators who constructed the model determined that the slope of the line β is significantly larger than 0.

(a) Suppose that you are interested in only those infants who are born prematurely. If you were to eliminate the four data points corresponding to newborns whose gestational age is 38 weeks or greater, would you still believe that there is a significant increase in CBVR as gestational age increases?

(b) In an earlier study, the same investigators found no obvious relationship between CBVR and gestational age in newborn infants; gestational age was not useful in predicting CBVR. Would this information cause you to modify your answer above?

9. The data set `lowbwt` contains information for the sample of 100 low birth weight infants born in Boston, Massachusetts [1] (Appendix B, Table B.7). Measurements of systolic blood pressure are saved under the variable name `sbp`, and values of gestational age under the name `gestage`.

(a) Construct a two-way scatter plot of systolic blood pressure versus gestational age. Does the graph suggest anything about the nature of the relationship between these variables?

(b) Using systolic blood pressure as the response and gestational age as the explanatory variable, compute the least-squares regression line. Interpret the estimated slope and y-intercept of the line; what do they mean in words?

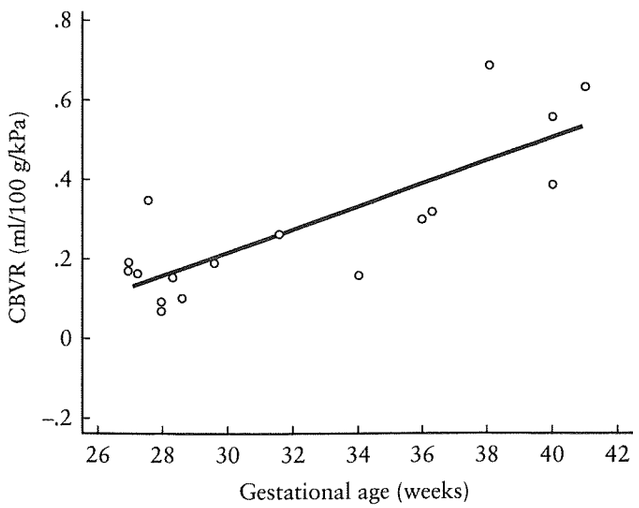

FIGURE 18.17
Response of cerebral blood volume versus gestational age for a sample of 17 newborns

(c) At the 0.05 level of significance, test the null hypothesis that the true population slope β is equal to 0. What do you conclude?

(d) What is the estimated mean systolic blood pressure for the population of low birth weight infants whose gestational age is 31 weeks?

(e) Construct a 95% confidence interval for the true mean value of systolic blood pressure when $x = 31$ weeks.

(f) Suppose that you randomly select a new child from the population of low birth weight infants with gestational age 31 weeks. What is the predicted systolic blood pressure for this child?

(g) Construct a 95% prediction interval for this new value of systolic blood pressure.

(h) Does the least-squares regression model seem to fit the observed data? Comment on the coefficient of determination and a plot of the residuals versus the fitted values of systolic blood pressure.

10. Measurements of length and weight for a sample of 20 low birth weight infants are contained in the data set `twenty` [1] (Appendix B, Table B.23). The length measurements are saved under the variable name `length`, and the corresponding birth weights under `weight`.

(a) Construct a two-way scatter plot of birth weight versus length for the 20 infants in the sample. Without doing any calculations, sketch your best guess for the least-squares regression line directly on the scatter plot.

(b) Now compute the true least-squares regression line. Draw this line on the scatter plot. Does the actual least-squares line concur with your guess?

Based on the two-way scatter plot, it is clear that one point lies outside the range of the remainder of the data. This point corresponds to the ninth infant in the sample of size 20. To illustrate the effect that the outlier has on the model, remove this point from the data set.

(c) Compute the new least-squares regression line based on the sample of size 19, and sketch this line on the original scatter plot. How does the least-squares line change? In particular, comment on the values of the slope and the intercept.

(d) Compare the coefficients of determination (R^2) and the standard deviations from regression ($s_{y|x}$) for the two least-squares regression lines. Explain how these values changed when you removed the outlier from the original data set. Why did they change?

11. The relationship between total fertility rate and the prevalence of contraceptive practice was investigated for a large number of countries around the world [5]. Measuring fertility rate in births per woman 15 to 49 years of age and the proportion of currently married women using any form of contraception as a percentage, the least-squares regression line relating these two quantities is

$$\hat{y} = 6.83 - 0.062x.$$

For the subset of 17 countries in the sub-Saharan region of Africa, fertility rate and the prevalence of contraceptive practice are saved in a data set called `africa` (Appendix B, Table B.24). Measures of fertility rate are saved under the variable name `fertrate`, and values of percentage contraception under `contra`.

(a) Interpret the slope and the intercept of the least-squares regression line generated for the countries around the world. What do they imply in words?
(b) Using the data for the 17 sub-Saharan African countries, construct a two-way scatter plot of total fertility rate versus prevalence of contraceptive practice. Does there appear to be a relationship between these two quantities?
(c) On the scatter plot, sketch the least-squares line estimated based on countries throughout the world.
(e) How do the actual fertility rates for the African nations compare with those that would be predicted based on the fitted regression line?

12. In the 11 years before the passage of the Federal Coal Mine Health and Safety Act of 1969, the fatality rates for underground miners varied little. After the implementation of that act, however, fatality rates decreased steadily until 1979. The fatality rates for the years 1970 through 1981 are provided below [6]; for computational purposes, calendar years have been converted to a scale beginning at 1. This information is contained in the data set `miner` (Appendix B, Table B.25). Values of the response, fatality rate, are saved under the name `rate`, and values of the explanatory variable, calendar year, under the name `year`.

Calendar Year	Year	Fatality Rate per 1000 Employees
1970	1	2.419
1971	2	1.732
1972	3	1.361
1973	4	1.108
1974	5	0.996
1975	6	0.952
1976	7	0.904
1977	8	0.792
1978	9	0.701
1979	10	0.890
1980	11	0.799
1981	12	1.084

(a) Construct a two-way scatter plot of fatality rate versus year. What does this plot suggest about the relationship between these two variables?
(b) To model the trend in fatality rates, fit the least-squares regression line

$$\hat{y} = \hat{\alpha} + \hat{\beta}x,$$

where x represents year. Using both the coefficient of determination R^2 and a plot of the residuals versus the fitted values of fatality rate, comment on the fit of the model to the observed data.
(c) Now transform the explanatory variable x to $\ln(x)$. Create a scatter plot of fatality rate versus the natural logarithm of year.
(d) Fit the least-squares model

$$\hat{y} = \hat{\alpha} + \hat{\beta}\ln(x).$$

Use the coefficient of determination and a plot of the residuals versus the fitted values of fatality rate to compare the fit of this model to the model constructed in (b).
(e) Transform x to $1/x$. Construct a two-way scatter plot of fatality rate versus the reciprocal of year.
(f) Fit the least-squares model

$$\hat{y} = \hat{\alpha} + \hat{\beta}\left(\frac{1}{x}\right).$$

Using the coefficient of determination and a plot of the residuals, comment on the fit of this model and compare it to the previous ones.
(g) Which of the three models appears to fit the data best? Defend your selection.

13. Statistics that summarize personal health care expenditures by state for the years 1966 through 1982 have been examined in an attempt to understand issues related to rising health care costs. Suppose that you are interested in focusing on the relationship between expense per admission into a community hospital and average length of stay in the facility. The data set `hospital` contains information for each state in the United States (including the District of Columbia) for the year 1982 [7] (Appendix B, Table B.26). The measures of mean expense per admission are saved under the variable name `expadm`; the corresponding average lengths of stay are saved under `los`.
(a) Generate numerical summary statistics for the variables expense per admission and length of stay in the hospital. What are the means and medians of each variable? What are their minimum and maximum values?
(b) Construct a two-way scatter plot of expense per admission versus length of stay. What does the scatter plot suggest about the nature of the relationship between these variables?
(c) Using expense per admission as the response and length of stay as the explanatory variable, compute the least-squares regression line. Interpret the estimated slope and y-intercept of this line in words.
(d) Construct a 95% confidence interval for β, the true slope of the population regression line. What does this interval tell you about the linear relationship between expense per admission and length of stay in the hospital?
(e) What is the coefficient of determination for the least-squares line? How is R^2 related to the Pearson correlation coefficient r?
(f) Construct a plot of the residuals versus the fitted values of expense per admission. In what three ways does the residual plot help you to evaluate the fit of the model to the observed data?

Bibliography

[1] Leviton, A., Fenton, T., Kuban, K. C. K., and Pagano, M., "Labor and Delivery Characteristics and the Risk of Germinal Matrix Hemorrhage in Low Birth Weight Infants," *Journal of Child Neurology,* Volume 6, October 1991, 35–40.

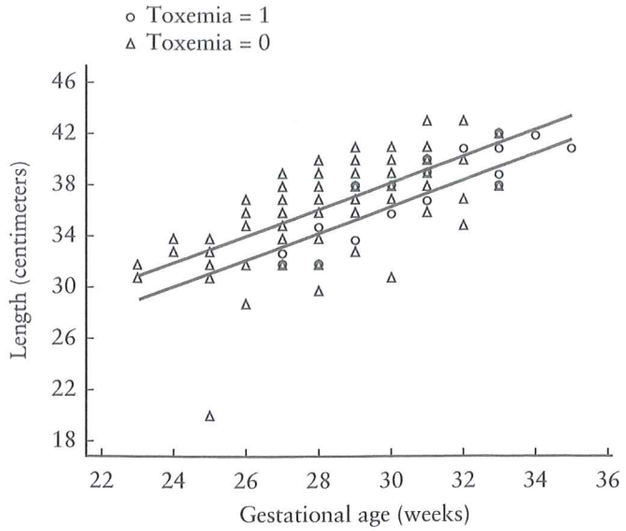

FIGURE 19.6
Fitted least-squares regression lines for different levels of toxemia

TABLE 19.3
Stata output displaying the linear regression of length on gestational age, toxemia, and their interaction

Source	SS	df	MS
Model	619.522097	3	206.507366
Residual	643.237903	96	6.70039483
Total	1262.76	99	12.7551515

Number of obs	= 100
F(3,96)	= 30.82
Prob > F	= 0.0000
R-square	= 0.4906
Adj R-square	= 0.4747
Root MSE	= 2.5885

length	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gestage	1.058458	.1262952	8.381	0.000	.8077647	1.309152
tox	-3.477085	8.5198381	-0.408	0.684	-20.38883	13.43466
gesttox	.0559409	.2794651	0.200	0.842	-4.4987929	.6106747
_cons	6.608269	3.592847	1.893	0.069	-.5234754	13.74001

We are unable to reject the null hypothesis that β_{13}, the coefficient of the interaction term, is equal to 0 ($p = 0.84$); the adjusted R^2 has decreased from 48.0% to 47.5%. Furthermore, the high correlation between toxemia and the gestational age–toxemia interaction—the Pearson correlation coefficient is equal to 0.997—has introduced collinearity into the model. Note that the standard error of the estimated coefficient of toxemia is approximately 12 times larger than it was in the model that did not contain the interaction term. Therefore, we conclude that there is no evidence that gestational age has a different effect on length depending on whether a mother experienced toxemia during pregnancy or not.

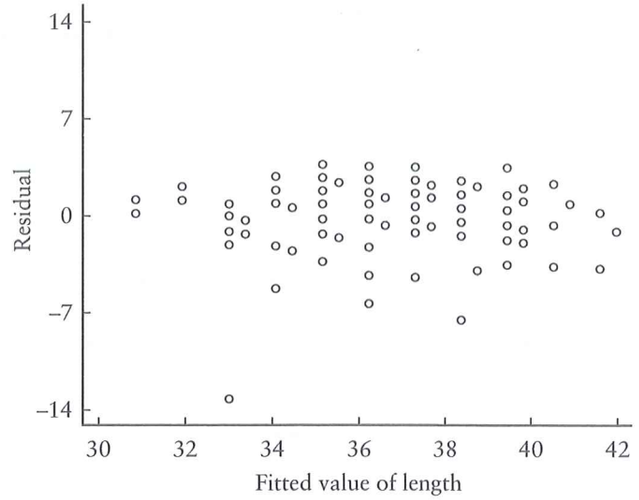

FIGURE 19.7
Residuals versus fitted values of length

Returning to the model that contains gestational age and toxemia status but not their interaction term, a plot of the residuals is displayed in Figure 19.7. There appears to be one outlier in the data set. We might consider dropping this observation, refitting the least-squares equation, and comparing the two models to determine how much of an effect the point has on the estimated coefficients. However, the assumption of homoscedasticity has not been violated, and a transformation of variables does not appear to be necessary.

19.4 Review Exercises

1. What assumptions do you make when using the method of least squares to estimate a population regression equation containing two or more explanatory variables?
2. Given a multiple regression model with a total of q distinct explanatory variables, how would you make inference about a single coefficient β_j?
3. Explain how the coefficient of determination and the adjusted R^2 can be used to help evaluate the fit of a multiple regression model to the observed data.
4. What is the function of an interaction term in a regression model? How is an interaction term created?

5. If you are performing an analysis with a single response and several potential explanatory variables, how would you decide which variables to include in a multiple regression model and which to leave out?

6. How can collinearity between two explanatory variables affect the estimated coefficients in a regression model?

7. In a study designed to examine the effects of adding oats to the typical American diet, individuals were randomly divided into two different groups. Twice a day, the first group substituted oats for other foods containing carbohydrates; the members of the second group did not make any changes to their diet. One outcome of interest is the serum cholesterol level of each individual eight weeks after the start of the study. Explanatory variables that might affect this response include diet group, serum cholesterol level at the start of the study, body mass index, and gender. The estimated coefficients and standard errors from the multiple regression model containing these four explanatory variables are displayed below [2].

Variable	Coefficient	Standard Error
Diet Group	−11.25	4.33
Baseline Cholesterol	0.85	0.07
Body Mass Index	0.23	0.65
Gender	−3.02	4.42

   (a) Conduct tests of the null hypotheses that each of the four coefficients in the population regression equation is equal to 0. At the 0.05 level of significance, which of the explanatory variables have an effect on serum cholesterol level eight weeks after the start of the study?
   (b) If an individual's body mass index were to increase by 1 kg/m^2 while the values of all other explanatory variables remained constant, what would happen to his or her serum cholesterol level?
   (c) If an individual's body mass index were to increase by 10 kg/m^2 while the values of all other explanatory variables remained constant, what would happen to his or her serum cholesterol level?
   (d) The indicator variable gender is coded so that 1 represents a male and 0 a female. Who is more likely to have a higher serum cholesterol level eight weeks after the start of the study, a man or a woman? How much higher would it be, on average?

8. For the population of low birth weight infants, a significant linear relationship was found to exist between systolic blood pressure and gestational age. Recall that the relevant data are in the file `lowbwt` [1] (Appendix B, Table B.7). The measurements of systolic blood pressure are saved under the variable name `sbp`, and the corresponding gestational ages under `gestage`. Also contained in the data set is `apgar5`, the five-minute apgar score for each infant. (The apgar score is an indicator of a child's general state of health five minutes after it is born; although it is actually an ordinal measurement, it is often treated as if it were continuous.)

   (a) Construct a two-way scatter plot of systolic blood pressure versus five-minute apgar score. Does there appear to be a linear relationship between these two variables?
   (b) Using systolic blood pressure as the response and gestational age and apgar score as the explanatory variables, fit the least-squares model

   $$\hat{y} = a + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

   Interpret $\hat{\beta}_1$, the estimated coefficient of gestational age. What does it mean in words? Similarly, interpret $\hat{\beta}_2$, the estimated coefficient of five-minute apgar score.
   (c) What is the estimated mean systolic blood pressure for the population of low birth weight infants whose gestational age is 31 weeks and whose five-minute apgar score is 7?
   (d) Construct a 95% confidence interval for the true mean value of systolic blood pressure when $x_1 = 31$ weeks and $x_2 = 7$.
   (e) Test the null hypothesis

   $$H_0: \beta_2 = 0$$

   at the 0.05 level of significance. What do you conclude?
   (f) Comment on the magnitude of R^2. Does the inclusion of five-minute apgar score in the model already containing gestational age improve your ability to predict systolic blood pressure?
   (g) Construct a plot of the residuals versus the fitted values of systolic blood pressure. What does this plot tell you about the fit of the model to the observed data?

9. The data set `lowbwt` also contains `sex`, a dichotomous random variable designating the gender of each infant.
   (a) Add the indicator variable `sex`—where 1 represents a male and 0 a female—to the model that contains gestational age. Given two infants with identical gestational ages, one male and the other female, which would tend to have the higher systolic blood pressure? By how much, on average?
   (b) Construct a two-way scatter plot of systolic blood pressure versus gestational age. On the graph, draw the two separate least-squares regression lines corresponding to males and to females. Is the gender difference in systolic blood pressure at each value of gestational age significantly different from 0?
   (c) Add to the model a third explanatory variable that is the interaction between gestational age and sex. Does gestational age have a different effect on systolic blood pressure depending on the gender of the infant?
   (d) Would you choose to include sex and the gestational age–sex interaction term in the regression model simultaneously? Why or why not?

10. The Bayley Scales of Infant Development produce two scores—the Psychomotor Development Index (PDI) and the Mental Development Index (MDI)—which can be used to assess a child's level of functioning. As part of a study examining the development and neurologic status of children who underwent reparative heart surgery during the first three months of life, the Bayley Scales were administered

to a sample of one-year-old infants born with congenital heart disease. Prior to heart surgery, the children had been randomized to one of two different treatment groups, called "circulatory arrest" and "low-flow bypass," which differed in the specific way in which the operation was performed. The data for this study are saved in the data set `heart` [3] (Appendix B, Table B.12). PDI scores are saved under the variable name `pdi`, MDI scores under `mdi`, and indicators of treatment group under `trtment`. For the treatment group variable, 0 represents circulatory arrest and 1 is low-flow bypass.

(a) In Chapter 11, the two-sample t-test was used to compare mean PDI and MDI scores for infants assigned to the circulatory arrest and low-flow bypass treatment groups. These analyses could also be performed using linear regression. Fit two simple linear regression models—one with PDI score as the response and the other with MDI score—that both have the indicator of treatment group as the explanatory variable.

(b) Who is more likely to have a higher PDI score, a child assigned to the circulatory arrest treatment group or one assigned to the low-flow bypass group? How much higher would the score be, on average?

(c) Who is more likely to have a higher MDI score? How much higher, on average?

(d) Is the treatment group difference in either PDI or MDI scores statistically significant at the 0.05 level? What do you conclude?

11. In Chapter 18, the relationship between expense per admission into a community hospital and average length of stay in the facility was examined for each state in the United States in 1982. The relevant data are in the file `hospital` [4] (Appendix B, Table B.26); mean expense per admission is saved under the variable name `expadm`, and average length of stay under the name `los`. Also included in the data set is `salary`, the average salary per employee in 1982.

(a) Summarize the average salary per employee both graphically and numerically. What are the mean and median average salaries? What are the minimum and maximum values?

(b) Construct a two-way scatter plot of mean expense per admission versus average salary. What does the graph suggest about the relationship between these two variables?

(c) Fit the least-squares model where mean expense per admission is the response and average length of stay and average salary are the explanatory variables. Interpret the estimated regression coefficients.

(d) What happens to the estimated coefficient of length of stay when average salary is added to the model?

(e) Does the inclusion of salary in addition to average length of stay improve your ability to predict mean expense per admission? Explain.

(f) Examine a plot of the residuals versus the fitted values of expense per admission. What does this plot tell you about the fit of the model to the observed data?

12. A study was conducted to examine the roles of firearms and various other factors in the rate of homicides in the city of Detroit. Information for the years 1961 to 1973 is provided in the data set `detroit` [5] (Appendix B, Table B.27); the number of homicides per 100,000 population is saved under the variable name `homicide`. Other variables in the data set include `police`, the number of full-time police officers per 100,000 population; `unemp`, the percentage of adults who are unemployed; `register`, the number of handgun registrations per 100,000 population; and `weekly`, the average weekly earnings for city residents.

(a) For each of the four explanatory variables listed—`police`, `unemp`, `register`, and `weekly`—construct a two-way scatter plot of homicide rate versus that variable. Are any linear relationships apparent from the graphs?

(b) Fit four simple linear regression models using homicide rate as the response and each of the other variables as the single explanatory variable. Individually, which of the variables have an effect on homicide rate that is significant at the 0.05 level?

(c) List the coefficient of determination for each regression model. Which variable explains the greatest proportion of the observed variation among the values of homicide rate?

(d) Using the method of forward selection, find the "best" multiple regression equation. Each variable contained in the final model should explain a significant amount of the observed variability in homicide rate. What do you conclude?

Bibliography

[1] Leviton, A., Fenton, T., Kuban, K. C. K., and Pagano, M., "Labor and Delivery Characteristics and the Risk of Germinal Matrix Hemorrhage in Low Birth Weight Infants," *Journal of Child Neurology,* Volume 6, October 1991, 35–40.

[2] Van Horn, L., Moag-Stahlberg, A., Liu, K., Ballew, C., Ruth, K., Hughes, R., and Stamler, J., "Effects on Serum Lipids of Adding Instant Oats to Usual American Diets," *American Journal of Public Health,* Volume 81, February 1991, 183–188.

[3] Bellinger, D. C., Jonas, R. A., Rappaport, L. A., Wypij, D., Wernovsky, G., Kuban, K. C. K., Barnes, P. D., Holmes, G. L., Hickey, P. R., Strand, R. D., Walsh, A. Z., Helmers, S. L., Constantinou, J. E., Carrazana, E. J., Mayer, J. E., Hanley, F. L., Castaneda, A. R., Ware, J. H., and Newburger, J. W., "Development and Neurologic Status of Children After Heart Surgery with Hypothermic Circulatory Arrest or Low-Flow Cardiopulmonary Bypass," *The New England Journal of Medicine,* Volume 332, March 2, 1995, 549–555.

[4] Levit, K. R., "Personal Health Care Expenditures, by State: 1966–1982," *Health Care Financing Review*, Volume 6, Summer 1985, 1–25.

[5] Fisher, J. C., "Homicide in Detroit: The Role of Firearms," *Criminology,* Volume 14, 1976, 387–400.