11.8 EXERCISES

Conceptual Exercises

1. Alcohol Metabolism. The subjects in the study were given alcohol on two consecutive days. On one of the days it was administered orally and on the other it was administered intravenously. The type of administration given on the first day was decided by a random mechanism. Why was this precaution taken?

2. Alcohol Metabolism. Here are two models for explaining the mean first-pass metabolism:

$$\text{Model 1: } \beta_0 + \beta_1 gast + \beta_2 fem$$
$$\text{Model 2: } \beta_0 + \beta_1 gast + \beta_2 gast \times fem.$$

(a) Why are there no formal tools for comparing these two models? (b) For a given value of gastric activity, what is the mean first-pass metabolism for men minus the mean first-pass metabolism for women (i) from Model 1? (ii) from Model 2?

3. Alcohol Metabolism. What would be the meaning of a third-order interactive effect of gastric activity, sex, and alcoholism on the mean first-pass metabolism?

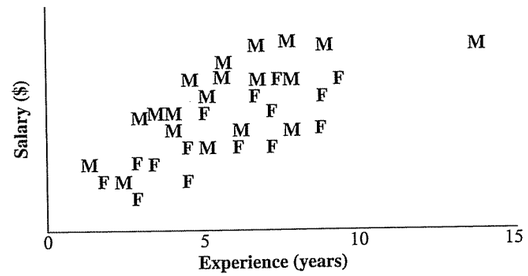
4. Blood–Brain Barrier. (a) How should rats have been randomly assigned to treatment groups? How many treatment groups were there? What is the name of this type of experimental design and this type of treatment structure? (b) How should rats have been randomly assigned to treatments if the researchers suspected that the sex of the rat might be associated with the response? What is the name of this type of experimental design?

5. Blood–Brain Barrier. The residual plot in Display 11.6 contains two distinct groups of points: a cluster on the left and a cluster on the right. (a) Why is this? (b) Does it imply any problems with the model?

6. Robustness and resistance are different properties. (a) What is the difference? (b) Why are they both relevant when the response distribution is "long-tailed"?

7. Display 11.19 shows a hypothetical scatterplot of salary versus experience, with different codes for males and females. The male and female slopes differ significantly if the male with the most experience is included, but not if he is excluded. What course of action should be taken, and why?

DISPLAY 11.19 Hypothetical scatterplot of salary versus experience for males and females



8. (a) Why does a case with large leverage have the *potential* to be influential? Why is it not necessarily influential? (b) Draw a hypothetical scatterplot of Y versus a single X, where one observation has a high leverage but is not influential. (c) Draw a hypothetical scatterplot of Y versus a single X, where one observation has a high leverage and is influential.

9. Suppose it is desired to obtain partial residuals in order to plot Y versus X_2 after getting the effect of X_1 out of the way. The first task is to fit the regression of Y on X_1 and X_2. Let the estimated coefficients be $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, and let the residuals from this fit be represented by res_i. The definition of the ith partial residual is $pres_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i}$. The alternative computational formula is $res_i + \hat{\beta}_2 X_{2i}$. Why are these formulas equivalent?

Computational Exercises

10. Pollen Removal. Reconsider the pollen removal data in Exercise 3.27. (a) Draw a coded scatterplot of the log of the proportion of pollen removed relative to the proportion unremoved (if p is the proportion removed, take $Y = \log[p/(1-p)]$) versus the log of the duration of visit, with a code to distinguish queens from workers. (b) Fit the regression of Y on the two explanatory variables and their interaction, and obtain a residual plot. Does the residual plot indicate any problems? (c) Obtain a set of case influence statistics. Are there any problem observations? What is the most advisable course of action? (d) Does a significant interaction appear to exist, or can the simpler parallel regression lines model be used?

11. Chernobyl Fallout. The data in Display 11.20 are the cesium ($^{134}Cs + {}^{137}Cs$) concentrations (in Bq/kg) in soil and in mushrooms at 17 wooded locations in Umbria, Central Italy, from August 1986 to November 1989. Researchers wished to investigate the cesium transfer from contaminated soil to plants—after the Chernobyl nuclear power plant accident in April 1986—by describing the distribution of the mushroom concentration as a function of soil concentration. (a) Obtain a set of case influence statistics from the simple linear regression fit of mushroom concentration on soil concentration. What do these indicate about case number 17? (b) Repeat part (a) after taking the logarithms of both variables. (Data from R. Borio et al., "Uptake of Radiocesium by the Mushrooms," *Science of the Total Environment* 106 (1991): 183–90.)

DISPLAY 11.20 Cesium ($^{134}Cs + {}^{137}Cs$) concentrations (in Bq/kg) in soil and in mushrooms at 17 locations in Italy, after the Chernobyl nuclear power plant accident; first 5 of 17 rows

Mushroom	Soil
1	33
9	55
14	138
17	319
20	415

12. Brain Weights. Reconsider the brain weight data of Display 9.4. (a) Fit the regression of brain weight on body weight, gestation, and log litter size, using no transformations. Obtain a set of case-influence statistics. Is any mammal influential in this fit? (b) Refit the regression without the influential observation, and obtain the new set of case influence statistics. Are there any influential observations from this fit? (c) What lessons about the connection between the need for a log transformation and influence can be discerned?

13. Brain Weights. Identifying which mammals have larger brain weights than were predicted by the regression model might point the way to further variables that can be examined. Fit the regression of log brain weight on log body weight, log gestation, and log litter size, and compute the studentized

residuals. Which mammals have substantially larger brain weights than were predicted by the model? Do any mammals have substantially smaller brain weights than were predicted by the model?

14. Corn Yield and Rainfall. Reconsider the data in Exercise 9.15. Fit the regression of corn yield on rainfall, rainfall-squared, and year. (a) Obtain the partial residuals of corn yield, adjusted for rainfall, and plot them versus year. (b) Obtain the augmented partial residual of corn yield, adjusted for year, and plot these values versus rainfall. (c) In your opinion, do these plots provide any clarification over the ordinary scatterplots?

15. Election Fraud. Reconsider the disputed election data from Exercise 8.20. (a) Draw a scatterplot of the Democratic percentage of absentee votes versus the Democratic percentage of machine votes for the 22 elections. Fit the simple linear regression of Democratic percentage of absentee votes on the Democratic percentage of machine votes and include the line on the plot. (b) Find the *internally* studentized residual for case number 22 from this fit. (c) Find the *externally* studentized residual for case number 22 from this fit. (d) Are the externally and internally studentized residuals for case 22 very different? If so, what can explain the difference? (e) What do the studentized residuals for case 22 indicate about the unusualness of the absentee ballot percentage for election 22 relative to the pattern of absentee and machine percentages established from the other 21 elections?

16. First-Pass Metabolism. Calculate the leverage, the studentized residual, and Cook's Distance for the 32nd case. Use the model with gastric activity, a sex indicator variable, and the interaction of these two.

17. Blood–Brain Barrier. (a) Using the data in Display 11.4, compute "jittered" versions of treatment, days after inoculation, and an indicator variable for females by adding small random numbers to each (uniform random numbers between -0.15 and 0.15 work well). (b) Obtain a matrix of the correlation coefficients among the same five variables (not jittered!). (c) In pencil, write the relevant correlation (two digits is enough) in a corner of each of the scatterplots in the matrix of scatterplots. (d) On the basis of this, what can be said about the relationship between the covariates (sex and days after inoculation), the response, and the design variables (treatment and sacrifice time)?

18. Blood–Brain Barrier. Using the data in Display 11.4, fit the regression of the log response (brain tumor-to-liver antibody ratio) on all covariates, the treatment indicator, and sacrifice time, treated as a factor with four levels (include three indicator variables, for sacrifice time = 3, 24, and 72 hours). (a) Obtain a set of case influence statistics, including a measure of influence, the leverage, and the studentized residual. (b) Discuss whether any influential observations or outliers occur with respect to this fit.

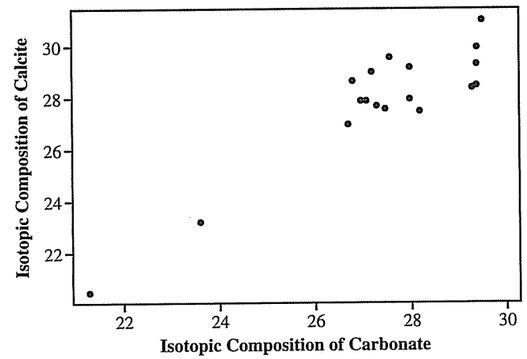
19. Blood–Brain Barrier. (a) Using the data in Display 11.4, fit the regression of the log response (brain tumor-to-liver antibody ratio) on an indicator variable for treatment and on sacrifice time treated as a factor with four levels (include three indicator variables, for sacrifice time = 3, 24, and 72 hours). Use the model to find the estimated mean of the log response at each of the eight treatment combinations (all combinations of the two infusions and the four sacrifice times). (b) Let X represent log of sacrifice time. Fit the regression of the log response on an indicator variable for treatment, X, X^2, and X^3. Use the estimated model to find the estimated mean of the log response at each of the eight treatment combinations. (c) Why are the answers to parts (a) and (b) the same?

20. Warm-Blooded *T. Rex*? The data in Display 11.21 are the isotopic composition of structural bone carbonate (X) and the isotopic composition of the coexisting calcite cements (Y) in 18 bone samples from a specimen of the dinosaur *Tyrannosaurus rex*. Evidence that the mean of Y is positively associated with X was used in an argument that the metabolic rate of this dinosaur resembled warm-blooded more than cold-blooded animals. (Data from R. E. Barrick and W. J. Showers, "Thermophysiology of *Tyrannosaurus rex*: Evidence from Oxygen Isotopes," *Science* 265 (1994): 222–24.) (a) Examine the effects on the p-value for significance of regression and on

R-squared of deleting (i) the case with the smallest value of X, and (ii) the two cases with the smallest values of X. (b) Why does R-squared change so much? (c) Compute the case influence statistics, and discuss interesting cases. (d) Recompute the case statistics when the case with the smallest X is deleted. (e) Comment on the differences in the two sets of case statistics. Why may pairs of influential observations not be found with the usual case influence statistics? (f) What might one conclude about the influence of the two unusual observations in this data set?

DISPLAY 11.21 Isotopic composition of carbonate and of calcite cements in 18 samples of bone from a *Tyrannosaurus rex* specimen

Carbonate	Calcite
21.3	20.5
23.6	23.2
26.7	27.0
26.8	28.7
27.0	27.9
27.1	27.9
27.2	29.0
27.3	27.7
27.5	27.6
27.6	29.6
28.0	28.0
28.0	29.2
28.2	27.5
29.3	28.4
29.4	28.5
29.4	29.3
29.4	30.0
29.5	31.0



21. Calculus Problem. The weighted least squares problem in multiple linear regression is to find the parameter values that minimize the weighted sum of squares,

$$\mathrm{SS}_w(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} w_i (Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_p X_{pi})^2,$$

with all $w_i > 0$. (a) Setting the partial derivatives of SS_w with respect to each of the parameters equal to zero, show that the solutions must satisfy this set of normal equations:

$$\begin{aligned}
\beta_0 \textstyle\sum w_i &+ \beta_1 \sum w_i X_{1i} &+ \beta_2 \sum w_i X_{2i} &+ \cdots + \beta_p \sum w_i X_{pi} &= \sum w_i Y_i \\
\beta_0 \textstyle\sum w_i X_{1i} &+ \beta_1 \sum w_i X_{1i}^2 &+ \beta_2 \sum w_i X_{1i} X_{2i} &+ \cdots + \beta_p \sum w_i X_{1i} X_{pi} &= \sum w_i X_{1i} Y_i \\
\beta_0 \textstyle\sum w_i X_{2i} &+ \beta_1 \sum w_i X_{2i} X_{1i} &+ \beta_2 \sum w_i X_{2i}^2 &+ \cdots + \beta_p \sum w_i X_{2i} X_{pi} &= \sum w_i X_{2i} Y_i \\
\vdots & \quad\vdots & \quad\vdots & \quad\vdots & \quad\vdots \\
\beta_0 \textstyle\sum w_i X_{pi} &+ \beta_1 \sum w_i X_{pi} X_{1i} &+ \beta_2 \sum w_i X_{pi} X_{2i} &+ \cdots + \beta_p \sum w_i X_{pi}^2 &= \sum w_i X_{pi} Y_i.
\end{aligned}$$

(b) Show that solutions to the normal equations minimize SS.

Data Problems

22. Deforestation and Debt. It has been theorized that developing countries cut down their forests to pay off foreign debt. Two researchers examined this belief using data from 11 Latin American nations. (Data from R. T. Gullison and E. C. Losos, "The Role of Foreign Debt in Deforestation in Latin America," *Conservation Biology* 7(1) (1993): 140–7.) The data on debt, deforestation, and population appear in Display 11.22. Does the evidence significantly support the theory that debt causes deforestation? Does debt exert any effect after the effect of population on deforestation is accounted for? Describe the effect of debt, after accounting for population.

DISPLAY 11.22 Foreign debt, annual deforestation area, and population for 11 Latin American countries

Country	Debt (millions of dollars)	Deforestation (thousands of hectares)	Population (thousands of people)
Brazil	86,396	12,150	128,425.0
Mexico	79,613	2,680	74,194.5
Ecuador	6,990	1,557	8,750.5
Colombia	10,101	1,500	27,254.0
Venezuela	24,870	1,430	16,170.5
Peru	10,707	1,250	18,496.5
Nicaragua	3,985	550	3,021.5
Argentina	36,664	400	29,400.5
Bolivia	3,810	300	5,970.5
Paraguay	1,479	250	3,424.5
Costa Rica	3,413	90	2,439.5

23. Air Pollution and Mortality. Does pollution kill people? Data in one early study designed to explore this issue came from five Standard Metropolitan Statistical Areas (SMSA) in the United States, obtained for the years 1959–1961. (Data from G. C. McDonald and J. A. Ayers, "Some Applications of the 'Chernoff Faces': A Technique for Graphically Representing Multivariate Data," in *Graphical Representation of Multivariate Data*, New York: Academic Press, 1978.) Total age-adjusted mortality from all causes, in deaths per 100,000 population, is the response variable. The explanatory variables listed in Display 11.23 include mean annual precipitation (in inches); median number of school years completed, for persons of age 25 years or older; percentage of 1960 population that is nonwhite; relative pollution potential of oxides of nitrogen, NO_X; and relative pollution potential of sulfur dioxide, SO_2. "Relative pollution potential" is the product of the tons emitted per day per square kilometer and a factor correcting for SMSA dimension and exposure. The first three explanatory variables are a subset of climate and socioeconomic variables in the original data set. (*Note*: Two cities—Lancaster and York—are heavily populated by members of the Amish religion,

DISPLAY 11.23 Air pollution and mortality data for 5 U.S. cities, 1959–1961, first 5 of 60 rows

City	Mortality	Precipitation	Education	Nonwhite	NO_x	SO_2
San Jose, CA	790.73	13	12.2	3.0	32	3
Wichita, KS	823.76	28	12.1	7.5	2	1
San Diego, CA	839.71	10	12.1	5.9	66	20
Lancaster, PA	844.05	43	9.5	2.9	7	32
Minneapolis, MN	857.62	25	12.1	2.0	11	26

who prefer to teach their children at home. The lower years of education for these two cities do not indicate a social climate similar to other cities with similar years of education.) Is there evidence that mortality is associated with either of the pollution variables, after the effects of the climate and socioeconomic variables are accounted for? Analyze the data and write a report of the findings, including any important limitations of this study. (*Hint*: Consider looking at case-influence statistics.)

24. Natal Dispersal Distances of Mammals. Natal dispersal distances are the distances that juvenile animals travel from their birthplace to their adult home. An assessment of the factors affecting dispersal distances is important for understanding population spread, recolonization, and gene flow—which are central issues for conservation of many vertebrate species. For example, an understanding of dispersal distances will help to identify which species in a community are vulnerable to the loss of connectedness of habitat. To further the understanding of determinants of natal dispersal distances, researchers gathered data on body weight, diet type, and maximum natal dispersal distance for various animals. Shown in Display 11.24 are the first 6 of 64 rows of data on mammals. (Data from G. D. Sutherland et al., "Scaling of Natal Dispersal Distances in Terrestrial Birds and Mammals," *Conservation Ecology* 4(1) (2000): 16.) Analyze the data to describe the distribution of maximum dispersal distance as a function of body mass and diet type. Write a summary of statistical findings.

DISPLAY 11.24 Natal dispersal distances and explanatory variables for 64 mammals; first 6 of 64 rows

Species	Body mass (kg)	Diet type	Maximum dispersal distance (km)
1. *Didellphis virginianus*	2.41	Omnivore	5.15
2. *Phascogale tapotafa*	0.17	Carnivore	6.80
3. *Trichosurus vulpecula*	2.93	Carnivore	12.80
4. *Sorex araneus*	0.004	Carnivore	0.87
5. *Scapanus townsendii*	0.15	Omnivore	0.86
6. *Ursus americanus*	104.45	Omnivore	225.00

25. Ingestion Rates of Deposit Feeders. The ingestion rates and organic consumption percentages of deposit feeders were considered in Exercise 9.21. The data set ex1125 repeats these data, but this time with three additional bivalve species included (the last three). The researcher wished to see if ingestion rate is associated with the percentage of organic matter in food, after accounting for animal weight, but was unsure about whether bivalves should be included in the analysis. Analyze the data to address this question of interest. (Data from L. M. Cammen, "Ingestion Rate: An Empirical Model for Aquatic Deposit Feeders and Detritivores," *Oecologia* 44 (1980): 303–10.)

26. Metabolism and Lifespan of Mammals. Use the data from Exercise 8.26 to describe the distribution of mammal lifespan as a function of metabolism rate, after accounting for the effect of body mass. One theory holds that for a given body size, animals that expend less energy per day (lower metabolic rate) will tend to live longer.

Answers to Conceptual Exercises

1. By randomly determining order, the researchers avoid bias in determining first-pass metabolism that would occur if an order effect existed.

2. (a) Neither of the models is a subset of the others, so it is impossible to test a term in a "full model" to see whether the "reduced model" does just as well. (Incidentally, the models explain the data about equally well.) (b) (i) β_2 (ii) $\beta_2 gast$.

3. It would mean that the effect of gastric AD activity on first-pass metabolism differed between males and females, and that the amount of the sex difference differed between alcoholics and

nonalcoholics. (Two-factor interactions are hard enough to describe in words. Three- and higher-factor interactions typically involve very long and confusing sentences. A theory without interactions is obviously simpler than one with interactions.)

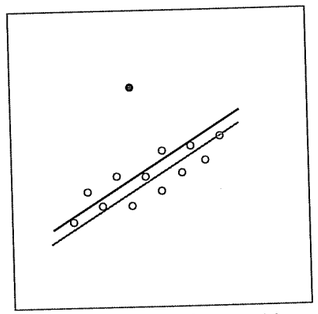
4. (a) Rats should have been randomly assigned to one of eight groups, corresponding to the eight combinations of treatment and sacrifice time. This is a completely randomized design with factorial (2×4) treatment structure. (b) If the response is suspected to be related to the sex of the rat, a randomized block experiment should be performed. The procedure in part (a) can be followed separately for male and female rats.

5. (a) The fitted values are noticeably larger for the rats in the groups with longer sacrifice times. (b) No.

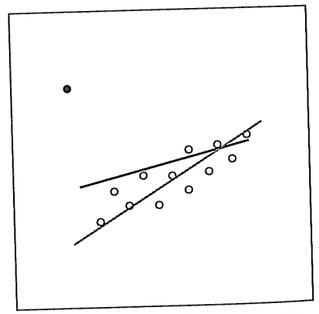
6. (a) *Robustness* describes the extent to which inferential statements are correct when specific assumptions are violated. *Resistance* describes the extent to which the results remain unchanged when a small portion of the data is changed, perhaps drastically (and therefore describes the extent to which individual observations can be influential). (b) When the distribution is long-tailed, the robustness against departures from normality cannot be guaranteed. Put another way, there are likely to be outliers, which can have undue influence on the results, since the least squares method is not resistant.

7. In the absence of further knowledge, it is safest to exclude the very experienced male from the data set and restrict conclusions about the differences between male and female salaries to individuals with 10 years of experience or fewer. It may be that males and females have different slopes over the wider range of experience, or it may be that the straight line is not an adequate model over the wider range of experiences. There is certainly insufficient data in the more-than-10-years range to resolve this issue.

8. (a) A large leverage indicates that a case occupies a position in the "X-space" that is not densely populated. It therefore plays a large role in shaping the estimated regression model in that region. Since it does not share its role (much) with other nearby points, it must draw the regression surface close to it. For this reason it has a high potential for influence. If, however, the fit of the model without that point is about the same as the fit of the model with it, it is not influential.



(b) The case with high leverage (●) exerts no influence on the slope. The line without the case (–) only has its intercept changed when the case is included (–).



(c) When the case with high leverage has its explanatory variable value outside those of the other cases, the slope of the regression can be changed dramatically.

9. Since $res_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}$, the alternative calculating formula is $Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} + \hat{\beta}_2 X_{2i}$. The last two terms cancel, leaving the original definition.

CHAPTER 12

Strategies for Variable Selection

There are two good reasons for paring down a large number of explanatory variables to a smaller set. The first reason is somewhat philosophical: Simplicity is preferable to complexity. Thus, redundant and unnecessary explanatory variables should be excluded on principle. The second reason is more concrete: Unnecessary terms in the model yield less precise inferences.

Various statistical tools are available for choosing a good subset from a large pool of explanatory variables. Discussed in this chapter are sequential variable-selection techniques, and comparisons among all possible subsets through examination of Cp and the Bayesian Information Criterion. The most important practical lessons are that the variable selection process should be sensitive to the objectives of the study and that the particular subset chosen is relatively unimportant.