10.4.6 The Principle of Occam's Razor

The principle of Occam's Razor is that simple models are to be preferred over complicated ones. Named after the 14th-century English philosopher, William of Occam, this principle has guided scientific research ever since its formulation. It has no underlying theoretical or logical basis; rather, it is founded in common sense and successful experience. It is often called the Principle of Parsimony.

In statistical applications, the idea translates into a preference for the more simple of two models that fit data equally well. One should seek a parsimonious model that is as simple as possible and yet adequately explains all that can be explained. Methods for paring down large sets of explanatory variables are discussed in Chapter 12.

10.4.7 Informal Tests in Model Fitting

Tests for hypotheses about regression coefficients—t-tests and extra-sum-of-squares F-tests—are valuable for two purposes: for formally providing evidence regarding questions of interest in a final model and for exploring models by testing potential terms at the exploratory stage. The attitude toward the p-values is somewhat different for these two purposes.

In answering questions of interest, p-values are given their formal interpretation. A small p-value provides evidence against the null hypothesis and in favor of the alternative. The null hypothesis may be true and the particular sample may have happened by chance to be an unusual one; the p-value provides a measure of just how unusual it would be. A large p-value means either that the null hypothesis is correct or that the study was not powerful enough to detect a departure from it.

For example, the adequacy of a straight line regression model may be examined through casual testing of an additional quadratic term. The statistical significance of the quadratic term helps to clarify possible curvature. In this usage of testing, some decision—in this case about a suitable model—is made. It should be realized that this decision is sample size dependent; one is more likely to find evidence of curvature in a large data set than in a small one. Nevertheless, the device of informal testing is useful in conjunction with other exploratory tools.

For answering questions of interest, tests should not be overemphasized. Even if the question of interest calls for a test, reporting a confidence interval to indicate the possible sizes of the effects of interest remains important. This is true whether the p-value for the test is small or large.

10.5 SUMMARY

Galileo's Study

Galileo's data are used to find a polynomial describing the mean distance as a function of height. The scatterplot shows that the relationship is not a straight line. The coefficient of a height-squared term, when added to the simple linear regression model, significantly differs from zero. R^2 is a useful summary here: the quadratic

regression model explains 99.03% of the variation in the horizontal distances. The large R^2 does not mean that all other terms are insignificant. A test of hypothesis is used to resolve that matter. In fact, when a height-cubed term is added to the model, the p-value for testing whether its coefficient is zero is 0.007, and the value of R^2 increases to 99.94%. This example demonstrates the benefit of R^2 for summarizing a fit and for indicating the degree of relevance of a significant term.

Echolocation by Bats

The goal is to see whether the in-flight energy expended by echolocating bats differs from the energy used by non-echolocating bats of similar body mass. There is no major question involving the birds, but their inclusion helps to clarify a common relationship between energy and body mass. A useful starting point in the analysis is a coded scatterplot of energy versus body mass. It is apparent by inspection that both energy and body mass should be transformed to their logarithms. A plot on this scale (Display 10.4) reveals a straight line relationship but very little additional difference in energy expenditure among the three types of flying vertebrates. Comparing the in-flight energy expenditures of echolocating bats and non-echolocating bats is difficult because the former are all small bats and the latter all big bats. Multiple linear regression permits a comparison after accounting for body mass, but the comparison must be made cautiously, since the ranges of body mass for the two types do not overlap. In light of this limitation, the comparison is made on the basis of an indicator variable in a parallel regression lines model. The data are consistent with the hypothesis that echolocating bats pay no extra energy price for their echolocating skills.

10.6 EXERCISES

Conceptual Exercises

1. **Galileo's Data.** Why is horizontal distance, rather than height, the response variable?

2. **Brain Weight.** Display 10.16 shows two possible *influence diagrams* relating the variables in the brain weight study of Section 9.1.2. If brain weight is directly associated with gestation period and litter size, as in part (a) of the figure, then animals that have approximately the same body size but different gestation period and different litter size should have different brain sizes, and scatterplots of brain size versus gestation and litter size individually should show some association. But the scatterplots can also show indirect associations, as in part (b) of the figure. If brain weight, gestation period, and litter size are all driven by body size, they should show mutual associations, whether or not a direct association exists. Can a statistical analysis distinguish between direct and indirect association? Explain how or, if not, why not.

3. **Brain Weight.** Consider the mammal brain weight data from Section 9.1.2, the model

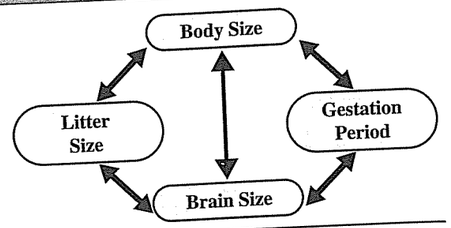
$$\mu\{lbrain \mid lbody, lgest, llitter\} = \beta_0 + \beta_1 lbody + \beta_2 lgest + \beta_3 llitter,$$

and the hypothesis $H\colon \beta_2 = 0$ and $\beta_3 = 0$. (a) Why can this not be tested by the two t-tests reported in the standard output? (b) Why can this not be tested by the two t-tests along with an adjustment for multiple comparisons?

DISPLAY 10.16 Associations of brain weight with gestation period and litter size: direct or indirect?

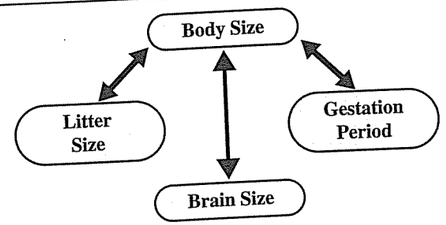
(a) DIRECT ASSOCIATION

Animals with about the same body size will have different brain sizes when their litter sizes and gestation periods are different.



(b) INDIRECT ASSOCIATION

Animals with about the same body size will have about the same brain size even if their gestation periods and litter sizes are different. Any apparent association between brain size and litter size, say, is a result of their both being related to body size.



4. Stocks and Jocks. Based on data from nine days in June 1994, a multiple regression equation was fit to the Dow Jones Index on the following seven explanatory variables: the high temperature in New York City on the previous day; the low temperature on the previous day; an indicator variable taking on the value 1 if the forecast for the day was sunny and 0 otherwise; an indicator variable taking on the value 1 if the New York Yankees won their baseball game of the previous day and 0 if not; the number of runs the Yankees scored; an indicator variable taking on the value of 1 if the New York Mets won their baseball game of the previous day and 0 if not; and the number of runs the Mets scored. As the chart in Display 10.17 shows, the predicted values of the stock market index were strikingly close to the actual values. R^2 was 89.6%. Why is this unremarkable?

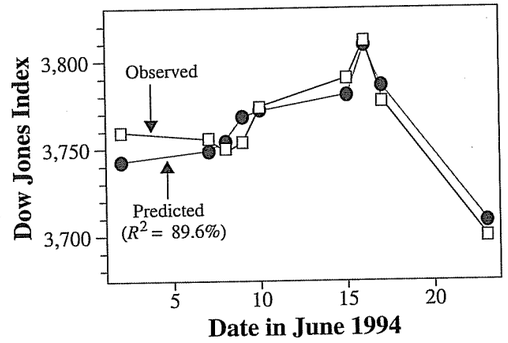
5. Bat Echolocation. Consider these three models:

$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass$$
$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$$
$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat + \beta_4 (lmass) \times (bird) + \beta_5 (lmass \times (ebat)).$$

(a) Explain why they can be described as representing the single line, parallel lines, and separate lines models, respectively. (b) Explain why the second model can be the "reduced model" in one F-test but the "full model" in another.

6. Bat Echolocation. A possible statement in conclusion of the analysis is this: "It is estimated that the median in-flight energy expenditure for echolocating bats is 1.08 times as large as the median in-flight energy expenditure for non-echolocating bats of similar body size." Referring to Display 10.4, explain why including the phrase, "of similar body size" in this statement is suspect. What alternative wording is available?

DISPLAY 10.17 Actual values and predicted values of Dow Jones Index



7. Life Is a Rocky Road. A regression of the number of crimes committed in a day on volume of ice cream sales in the same day showed that the coefficient of ice cream sales was positive and significantly differed from zero. Which of the following is the most likely explanation? (a) The content of ice cream (probably the sugar) encourages people to commit crimes. (b) Successful criminals celebrate by eating ice cream. (c) A pathological desire for ice cream is triggered in a certain percentage of individuals by certain environmental conditions (such as warm days), and these individuals will stop at nothing to satisfy their craving. (d) Another variable, such as temperature, is associated with both crime and ice cream sales.

8. In the terminology of extra-sums-of-squares F-tests, does the reduced model correspond to the case where the null hypothesis is true? Is the full model the one that corresponds to the alternative hypothesis's being true?

Computational Exercises

9. Crab Claws. Reconsider the data on claw closing force and claw size for three species of crabs, shown in Exercise 7.22. Display 10.18 shows output from the least squares fit to the separate lines model for the regression of log force on log height. The regression model for log force was

$$\mu\{lforce \mid lheight, SPECIES\} = \beta_0 + \beta_1 lheight + \beta_2 lb + \beta_3 cp + \beta_4 (lheight \times lb) + \beta_5 (lheight \times cp),$$

where *lheight* represents log height, *lb* represents an indicator variable for the second species, and *cp* represents an indicator variable for the third species. The sample size was 38.

(a) How many degrees of freedom are there in the estimate of σ?
(b) What is the p-value for the test of the hypothesis that the slope in the regression of log force on log height is the same for species 2 as it is for species 1?
(c) What is a 95% confidence interval for the amount by which the slope for species 3 exceeds the slope for species 1?

10. Crab Claws. The sum of squared residuals from the fit described in Exercise 9 is 5.99713, based on 32 degrees of freedom. The sum of squared residuals from the fit without the last two terms

DISPLAY 10.18 Least squares output for Exercise 9

Variable	Estimate	SE	*t*-stat	*p*-value
Constant	0.5191	1.0000	0.5191	0.6073
lheight	0.4083	0.4868	0.8387	0.4079
lb	−4.2992	1.5283	2.8131	0.0083
cp	−2.4864	1.7606	1.4123	0.1675
lheight × *lb*	2.5653	0.7354	3.4885	0.0014
lheight × *cp*	1.6601	0.7889	2.1043	0.0433

DISPLAY 10.19 Regression output data for Exercise 11

Variable	Estimate	SE	*t*-stat	*p*-value
Constant	3.775	0.3881	9.7321	<0.0000
lsize	0.0809	0.1131	0.7139	0.2443
days	0.0774	0.1447	0.5346	0.5104

Estimated SD about the regression is 0.8234 on 13 degrees of freedom; $R^2 = 11.41\%$.

is 8.38155, based on 34 degrees of freedom. Form an F-statistic and find the p-value for the test that the slopes are the same for the three species.

11. Butterfly Occurrences. Display 10.19 summarizes results from the regression of the log of the number of butterfly species observed on the log of the size of the reserve and the number of days of observations, from 16 reserves in the Amazon River Basin.

(a) What is the two-sided p-value for the test of whether size of reserve has any effect on number of species, after accounting for the days of observation? What is the one-sided p-value if the alternative is that size has a positive effect? Does this imply that there is no evidence that the median number of species is related to reserve size? The researchers tended to spend more days searching for butterflies in the larger reserves. How might this affect the interpretation of the results?

(b) What is a two-sided p-value for the test that the coefficient of *lsize* is 1? (This is simply a computational exercise; there is no obvious reason to conduct this test with these data.)

(c) What is a 95% confidence interval for the coefficient of *lsize*?

(d) What proportion of the variation in log number of species remains unexplained by log size and days of observations?

12. Brain Weights. With the data described in Section 9.1.2, construct an extra sum of squares F-test for determining whether gestation period and litter size are associated with brain weight after body weight is accounted for.

13. Bat Echolocation. (a) Fit the parallel regression lines model to duplicate the results in Display 10.6. (b) From these results, what are the estimated intercept and estimated slope for the regression of log energy on log mass for (i) non-echolocating bats, (ii) non-echolocating birds, and (iii) echolocating bats? (c) Refit the model using, instead, the indicator variables *bird* and *nbat*, where *nbat* takes on the value 1 for species of non-echolocating bats and 0 for other species. (d) Based on the results in part (c), what are the estimated intercept and estimated slope for the regression of log energy on log mass for (i) non-echolocating bats, (ii) non-echolocating birds, and (iii) echolocating bats? How

DISPLAY 10.20 Protein in minnow larvae exposed to copper and zinc

Copper (ppm)	Zinc (ppm)	Protein (μg/larva)	Copper (ppm)	Zinc (ppm)	Protein (μg/larva)
0	0	201	112.5	0	188
0	375	186	112.5	375	172
0	750	173	112.5	750	157
0	1,125	110	112.5	1,125	115
0	1,500	115	112.5	1,500	108
37.5	0	202	150	0	133
37.5	375	161	150	375	125
37.5	750	172	150	750	184
37.5	1,125	138	150	1,125	135
37.5	1,500	133	150	1,500	114
75	0	204			
75	375	165			
75	750	148			
75	1,125	143			
75	1,500	123			

do these compare to the estimates obtained in part (b)? (e) With the results of (c), test whether the lines for the echolocating bats and the non-echolocating birds coincide.

14. Toxic Effects of Copper and Zinc. In a study of the joint toxicity of copper and zinc, researchers randomly allocated 25 beakers containing minnow larvae to receive one of 25 treatment combinations. The treatment levels were all combinations of 5 levels of zinc and 5 levels of copper added to a beaker. Following a four-day exposure, a sample of the minnow larvae were homogenized and analyzed for protein. The results are shown in Display 10.20. (Data from D. A. J. Ryan, J. J. Hubert, J. B. Sprague, and J. Parrott, "A Reduced-Rank Multivariate Regression Approach to Aquatic Joint Toxicity Experiments," *Biometrics* 48 (1992): 155–62.) Fit a full second-order model for the regression of protein on copper and zinc, and examine the plot of residuals versus fitted values. Repeat after taking the log of protein. Which model is preferable?

15. Kentucky Derby. Reconsider the Kentucky Derby winning times and speeds from Exercise 9.20. Test whether there is any effect of the categorical factor "Track" (with seven categories) on winning speed, after accounting for year. The full model will have *Year* and the categorical factor *Track*; the reduced model will have only *Year*.

16. Galileo's Data. Use Galileo's data in Display 10.1 (data file: case1001) to perform the following operations.

(a) Fit the regression of distance on height and height-squared. Obtain the estimates, their standard errors, the estimate of σ^2, and the variance–covariance matrix of the estimated coefficients.

(b) Verify that the square roots of the diagonal elements are equal to the standard errors reported with the estimated coefficients.

(c) Compute the estimated mean distance when the initial height is 500 punti.

(d) Calculate the standard error for the estimated mean in part (c).

(e) Use the answer to parts (a) and (d) and the relationship between the variance of the estimated mean and the variance of prediction to obtain the standard error of prediction at an initial height of 500 punti.

17. Galileo's Data. Use Galileo's data in Display 10.1 (data file case1001) to fit the regression of distance on (a) height; (b) height and height^2; (c) height, height^2, and height^3; (d) height, height^2, height^3, and height^4; (e) height, height^2, height^3, height^4, and height^5; (f) height, height^2, height^3, height^4, height^5, and height^6. For each part, find R^2 and R^2_{adj}.

18. Corn Yield and Rainfall. Reconsider the corn yield and rainfall data (Display 9.18; data file ex0915). Fit the regression of yield on rainfall, rainfall-squared, and year. Use the approach of Section 10.4.5 to find the rainfall that maximizes mean yield.

19. Meadowfoam. Carry out a *lack-of-fit* F-test for the regression of number of flowers on light intensity and an indicator variable for time, using the data in Display 9.2 (data file case0901): (a) Fit the regression of *flowers* on *light* and an indicator variable for *time* = 24, and obtain the analysis of variance table. (b) Fit the same regression except with *light* treated as a factor (using 5 indicator variables to distinguish the 6 groups), and with the interaction of these two factors, and obtain the analysis of variance table. (c) Perform an extra-sum-of-squares F-test comparing the full model in part (b) to the reduced model in part (a). (*Note*: The full model contains 12 parameters, which is equivalent to the model in which a separate mean exists for each of the 12 groups. No pattern is implied in this model. See Displays 9.7 and 9.8 for help.)

20. Calculus Problem. The least squares problem in multiple linear regression is to find the parameter values that minimize the sum of squared differences between responses and fitted values,

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

Set the partial derivatives of SS with respect to each of the unknowns equal to zero. Show that the solutions must satisfy the set of *normal equations*, as follows:

$$\begin{aligned}
&\beta_0 n + \beta_1 \Sigma X_{1i} + \beta_2 \Sigma X_{2i} + \cdots + \beta_p \Sigma X_{pi} = \Sigma Y_i \\
&\beta_0 \Sigma X_{1i} + \beta_1 \Sigma X_{1i}^2 + \beta_2 \Sigma X_{1i} X_{2i} + \cdots + \beta_p \Sigma X_{1i} X_{pi} = \Sigma X_{1i} Y_i \\
&\beta_0 \Sigma X_{2i} + \beta_1 \Sigma X_{2i} X_{1i} + \beta_2 \Sigma X_{2i}^2 + \cdots + \beta_p \Sigma X_{2i} X_{pi} = \Sigma X_{2i} Y_i \\
&\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\
&\beta_0 \Sigma X_{pi} + \beta_1 \Sigma X_{pi} X_{1i} + \beta_2 \Sigma X_{pi} X_{2i} + \cdots + \beta_p \Sigma X_{pi}^2 = \Sigma X_{pi} Y_i,
\end{aligned}$$

where each Σ indicates summation over all cases ($i = 1, 2, \ldots, n$). Show, too, that solutions to the normal equations minimize SS.

21. Matrix Algebra Problem. Let $\mathbf{Y}$ be the $n \times 1$ column vector containing the responses, let $\mathbf{X}$ be the $n \times (p+1)$ array whose first column consists entirely of ones and whose other columns are the explanatory variable values, and let $\mathbf{b}$ be the $(p+1) \times 1$ column containing the resulting parameter estimates. Show that the normal equations in Exercise 20 can be written in the form

$$(\mathbf{X}^T\mathbf{X})\mathbf{b} = \mathbf{X}^T\mathbf{Y}.$$

Therefore, as long as the matrix inversion is possible, the least squares solution is

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

When is the inversion possible?

22. Continuing Exercise 21, statistical theory says that the means of the estimates in the vector $\mathbf{AY}$, where $\mathbf{A}$ is a matrix, are the elements of the vector $\mathbf{A}\mu\{\mathbf{Y}\}$; and the matrix of covariances of these estimates is $\mathbf{A}\text{Cov}(\mathbf{Y})\mathbf{A}^T$. Use the theory and the model $\mu\{\mathbf{Y}\} = \mathbf{X}\beta, \text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$ (where $\mathbf{I}$ is an $n \times n$ identity matrix) to show that the mean in the sampling distributions of the least squares estimate $\mathbf{b}$ is β. Then show that the matrix of covariances is $\text{Cov}\{\mathbf{b}\} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

23. Speed of Evolution. Refer back to Exercise 9.18. The authors of that study concluded that although the wing size of North American flies was converging rapidly to the same cline as exhibited by the European flies, the means by which the cline is achieved is different in the North American population.

(a) As evidence that the means of convergence is different, the authors concluded that there was a marked difference between the NA and the EU patterns of the basal length-to-wing size ratios versus latitude (in females). Fit a multiple linear regression that allows for different slopes and different intercepts. In a single F-test, evaluate the evidence against there being a single straight line that describes the cline on both continents. If you conclude there is a difference, is the difference one of slope alone? of intercept alone? or of both?

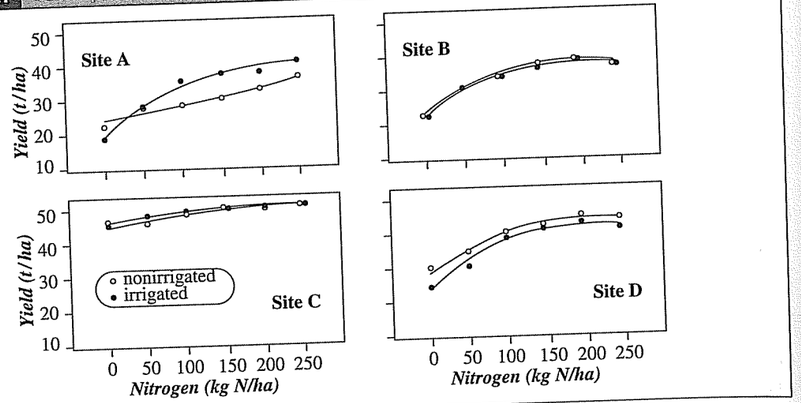
(b) Return to the basic question of whether the wing sizes in NA flies have established a cline similar to their EU ancestors. Using the model developed in Exercise 9.18, answer these questions: (i) Is there a nonzero slope to the cline of NA females? (ii) Is there a nonzero slope to the cline of NA males? (iii) Is there a difference between the clines of NA and EU females, and if so, what is its nature? (iv) Repeat (iii) for males.

24. Speed of Evolution. (Refer again to Exercise 9.18 and also to Exercise 10.23.) Many software systems allow the user to perform weighted regression, in which different squared residuals from regression receive different weights in deciding which set of parameter estimates provide the smallest sum of squared residuals. If each individual response has an independent estimate of its likely error, the weight given to each residual is usually taken to be the reciprocal of the square of that likely error. The standard error of wing sizes are standard errors of the averages around 2 individual (log) wing sizes. If your software allows for weights, construct a weight variable as the inverse square of the standard errors. Then repeat both parts of Exercise 10.23 using weighted regression. Do the results differ? Why is this preferable to using each fly as a separate case?

25. Potato Yields. Nitrogen and water are important factors influencing potato production. One study of their roles was conducted at sites in the St. John River Valley of New Brunswick. (G. Belanger et al., "Yield Response of Two Potato Cultivars to Supplemental Irrigation and N fertilization in New Brunswick," *American Journal of Potato Research* 77 (2000): 11–21.) Nitrogen fertilizer was applied at six different levels in combination with two water conditions: irrigated or nonirrigated. This design was repeated at four different sites in 1996, with the resulting yields depicted in Display 10.21. Notice that the patterns of responses against nitrogen level are fit reasonably well by quadratic curves.

Each quadratic requires 3 parameters, so a model that would allow for separate quadratic curves for each site-by-irrigation combination would have 24 parameters. (a) Using indicator functions for sites and for irrigation, construct a multiple linear regression model with 23 variables that will allow for completely different quadratic curves. Interpret the parameters in this model, if possible. (b) Describe how you would answer the following questions: (i) Is there evidence that the manner in which the quadratic terms differ by water condition changes from site to site (or is the difference the same at all four sites)? (ii) If the quadratic term differences are the same at all sites, is there strong evidence of a difference by water condition? (iii) If there is no difference between quadratic terms by water or by site, is there evidence of any quadratic term at all? (iv), (v), and (vi): Repeat (i), (ii), and (iii) for the linear terms, if there is no evidence of any quadratic terms. (c) Why are the questions in (b) ordered as they are?

DISPLAY 10.21 Potato yield versus nitrogen at four sites



Data Problems

26. Thinning of Ozone Layer. Thinning of the protective layer of ozone surrounding the earth may have catastrophic consequences. A team of University of California scientists estimated that increased solar radiation through the hole in the ozone layer over Antarctica altered processes to such an extent that primary production of phytoplankton was reduced 6 to 12%.

Depletion of the ozone layer allows the most damaging ultraviolet radiation—UVB (280–320 nm)—to reach the earth's surface. An important consequence is the degree to which oceanic phytoplankton production is inhibited by exposure to UVB, both near the ocean surface (where the effect should be slight) and below the surface (where the effect could be considerable).

To measure this relationship, the researchers sampled from the ocean column at various depths at 17 locations around Antarctica during the austral spring of 1990. To account for shifting of the ozone hole's positioning, they constructed a measure of UVB exposure integrated over exposure time. The exposure measurements and the percentages of inhibition of normal phytoplankton production were extracted from their graph to produce Display 10.22. (Data from R. C. Smith et al., "Ozone Depletion: Ultraviolet Radiation and Phytoplankton Biology in Antarctic Waters," *Science* 255 (1992): 952–57.) Does the effect of UVB exposure on the distribution of percentage inhibition differ at the surface and in the deep? How much difference is there? Analyze the data, and write a summary of statistical findings and a section of details documenting those findings. (*Suggestion*: Fit the model with different intercepts and different slopes, even if some terms are not significantly different from zero.)

27. Factors Affecting Extinction. The data in Display 10.23 are measurements on breeding pairs of land-bird species collected from 16 islands around Britain over the course of several decades. For each species, the data set contains an average time of extinction on those islands where it appeared (this is actually the reciprocal of the average of $1/T$, where T is the length of time the species remained on the island, and $1/T$ is taken to be zero if the species did not become extinct on the island); the average number of nesting pairs (the average, over all islands where the birds appeared, of the number of nesting pairs per year); the size of the species (categorized as large or small); and the migratory status of the species (migrant or resident). (Data from S. L. Pimm, H. L. Jones, and J. Diamond, "On the Risk of Extinction," *American Naturalist* 132 (1988): 757–85.) It is expected that species with larger numbers of nesting pairs will tend to remain longer before becoming extinct. Of interest is whether, after accounting for number of nesting pairs, size or migratory status has any effect. There is also some interest in whether the effect of size differs depending on the number of nesting pairs. If any species have unusually small or large extinction times compared to other species with similar values of the explanatory variables, it would be useful to point them out. Analyze the data. Write a summary of statistical findings and a section of details documenting the findings.

DISPLAY 10.22 First 5 of 17 rows of a data set with exposure to ultraviolet B radiation and percentage inhibition of primary phytoplankton production in Antarctic water

Location	Percent inhibition	UVB exposure	Surface (S) or Deep (D)
1	0.0	0.0000	D
2	1.0	0.0000	D
3	6.0	0.0100	D
4	7.0	0.0150	S
5	7.0	0.0185	S

28. El Niño and Hurricanes. Shown in Display 10.24 are the first few rows of a data set with the numbers of Atlantic Basin tropical storms and hurricanes for each year from 1950 to 1997. The variable *storm index* is an index of overall intensity of the hurricane season. (It is the average of number of tropical storms, number of hurricanes, the number of days of tropical storms, the number of days of hurricanes, the total number of intense hurricanes, and the number of days they last—when each of these is expressed as a percentage of the average value for that variable. A *storm index* score of 100, therefore, represents, essentially, an average hurricane year.) Also listed are whether the year was a cold, warm, or neutral El Niño year, a constructed numerical variable *temperature* that takes on the values −1, 0, and 1 according to whether the El Niño temperature is cold, neutral, or warm; and a variable indicating whether West Africa was wet or dry that year. It is thought that the warm phase of El Niño suppresses hurricanes while a cold phase encourages them. It is also thought that wet years in West Africa often bring more hurricanes. Analyze the data to describe the effect of El Niño on (a) the number of tropical storms, (b) the number of hurricanes, and (c) the storm index after accounting for the effects of West African wetness and for any time trends, if appropriate. (These data were gathered by William Gray of Colorado State University, and reported on the *USA Today* weather page: www.usatoday.com/weather/whurnum.htm)

29. Wage and Race 1987. Shown in Display 10.25 are the first few rows of a data set from the 1988 March U.S. Current Population Survey. The set contains weekly wages in 1987 (in 1992 dollars) for a sample of 25,437 males between the age of 18 and 70 who worked full-time, their years of education, years of experience, whether they were black, whether they worked in a standard metropolitan statistical area (i.e., in or near a city), and a code for the region in the U.S. where they worked (Northeast, Midwest, South, and West). Analyze the data and write a brief statistical report to see whether and to what extent black males were paid less than nonblack males in the same region and with the same levels of education and experience. Realize that the extent to which blacks were paid differently than nonblacks may depend on region. (Suggestion: Refrain from looking at interactive effects, except for the one implied by the previous sentence.) (These data, from the Current Population Survey (CPS), were discussed in the paper by H. J. Bierens and D. K. Ginther, "Integrated Conditional Moment Testing of Quantile Regression Models," *Empirical Economics* 26 (2001): 307–24; and made available at the Web site http://econ.la.psu.edu/~hbierens/MEDIAN.HTM (April, 2008).)

DISPLAY 10.23 Bird extinction data

Species	Ave. extinction time (years)	Ave. number of nesting pairs	Size (large or small)	Migratory status (resident or migrant)
Sparrowhawk	3.030	1.000	L	R
Buzzard	5.464	2.000	L	R
Kestrel	4.098	1.210	L	R
Peregrine	1.681	1.125	L	R
Grey partridge	8.850	5.167	L	R
Quail	1.493	1.000	L	M
Red-legged partridge	7.692	2.750	L	R
Pheasant	3.846	5.630	L	R
Water rail	16.667	3.000	L	R
Corncrake	4.219	4.670	L	M
Moorhen	8.130	4.056	L	R
Coot	5.000	1.000	L	R
Lapwing	7.299	6.960	L	M
Golden plover	1.000	1.670	L	M
Ringed plover	27.027	5.560	L	R
Curlew	3.106	2.830	L	M
Redshank	4.000	4.375	L	M
Snipe	16.129	4.125	L	M
Stock dove	3.484	3.670	L	R
Rock dove	37.037	8.330	L	R
Wood pigeon	7.299	2.750	L	R
Cuckoo	2.525	1.430	L	M
Short-eared owl	4.132	2.000	L	R
Little owl	2.000	2.750	L	R
Magpie	10.000	4.500	L	R
Jackdaw	2.667	7.120	L	R
Carrion crow	4.587	4.580	L	R
Raven	58.824	2.350	L	R
Skylark	32.258	6.870	S	R
Swallow	2.571	3.830	S	M
House martin	2.160	5.000	S	M
Yellow wagtail	1.000	1.250	S	M
Pied wagtail	2.967	2.270	S	R
Meadow pipit	9.524	5.350	S	R
Wren	11.111	8.700	S	R
Dunnock	7.299	6.100	S	R
Robin	4.000	3.330	S	R
Stonechat	2.381	3.640	S	R
Wheatear	2.611	4.830	S	M
Blackbird	3.257	4.670	S	R
Song thrush	1.701	1.700	S	R
Mistle thrush	1.795	1.330	S	R
Grasshopper warbler	1.198	1.000	S	M
Sedge warbler	3.185	1.900	S	M
Whitethroat	2.273	4.420	S	M
Willow warbler	1.111	1.250	S	M
Chiffchaff	1.000	1.000	S	M
Goldcrest	1.000	1.000	S	R
Spotted flycatcher	1.230	1.000	S	M
Great tit	6.061	2.500	S	R
Blue tit	3.175	1.500	S	R
Yellowhammer	2.000	2.500	S	R
Reed bunting	5.076	5.630	S	R
Chaffinch	1.934	2.370	S	R
Goldfinch	1.493	1.500	S	R
Redpoll	1.000	1.000	S	R
Linnet	5.102	6.500	S	R
House sparrow	3.003	4.500	S	R
Tree sparrow	1.898	2.170	S	R
Starling	41.667	11.620	S	R
Pied flycatcher	1.000	1.000	S	M
Siskin	1.000	1.000	S	R

30. Wages and Race 2011. Display 10.26 is a partial listing of a data set with weekly earnings for 4,952 males between the age of 18 and 70 sampled in the March 2011 Current Population Survey (CPS). These males are a subset who had reported earnings and who responded as having race as either "Only White" or "Only Black." Also recorded are the region of the country (with four categories: Northeast, Midwest, South, and West), the metropolitan status of the men's employment (with three categories: Metropolitan, Not Metropolitan, and Not Identified), age, education category (with 16 categories ranging from "Less than first grade" to "Doctorate Degree"), and education code, which is a numerical value that corresponds roughly to increasing levels of education (and so may be useful for plotting). What evidence do the data provide that the distributions of weekly earnings differ in the populations of white and black workers after accounting for the other variables? By how many dollars or by what percent does the White population mean (or median) exceed the Black population mean (or median)? (Data from U.S. Bureau of Labor Statistics and U.S. Bureau of the Census: Current Population Survey, March 2011 http://www.bls.census.gov/cps_ftp.html#cpsbasic; accessed July 25, 2011.)

DISPLAY 10.24 Atlantic Basin hurricane and El Niño data for 1950–1997, partial listing

Year	El Niño	Temperature	West Africa	Storms	Hurricanes	Storm index
1950	cold	−1	1	13	11	243
1951	warm	1	0	10	8	121
1952	neutral	0	1	7	6	97
1953	warm	1	1	14	6	121
...						

DISPLAY 10.25 Data on the first 5 individuals (out of 25,437) in the 1987 wage and race data set

Region	MetropolitanStatus	Exper	Educ	Race	WeeklyEarnings
South	NotMetropolitanArea	8	12	NotBlack	859.71
Midwest	MetropolitanArea	30	12	NotBlack	786.73
West	MetropolitanArea	31	14	NotBlack	1424.5
West	MetropolitanArea	17	16	NotBlack	959.16
West	MetropolitanArea	6	12	NotBlack	154.32

DISPLAY 10.26 First five rows of a data set with weekly earnings (in U.S. dollars) in 2011 for 4,952 males who specified either "White Only" or "Black Only" as their race

Region	MetropolitanStatus	Age	EducCat	EducCode	Race	WeeklyEarnings
West	Not Metropolitan	64	SomeCollegeButNoDegree	40	White	1418.84
Midwest	Metropolitan	51	AssocDegAcadem	42	White	1000.00
South	Metropolitan	25	NinthGrade	35	White	420.00
West	Not Metropolitan	46	MastersDegree	44	White	1980.00
Northeast	Metropolitan	31	BachelorsDegree	43	White	1750.00

31. Who Looks After the Kids? Different bird species have different strategies: Maternal, Paternal, and BiParental care. In 1984 J. Van Rhijn argued in the *Netherlands Journal of Zoology* that parental care was the ancestral condition, going back to the dinosaur predecessors of birds. D. J. Varricchio et al. (*Science*, Vol. 322, Dec. 19, 2008, pp. 1826–28) tested that argument by comparing the relationships between Clutch Volume and adult Body Mass in six different groups: modern maternal-care bird species (*Mat*; $n = 171$), modern paternal-care bird species (*Pat*; $n = 40$), modern biparental-care bird species (*BiP*; $n = 204$), modern maternal-care crocodiles (*Croc*; $n = 19$), non-avian maniraptoran dinosaurs thought to be ancestors of modern birds (*Mani*; $n = 3$), and other non-avian dinosaurs (*Othr*; $n = 6$). The question of interest was which group of modern creatures most closely matches the relationship in the maniraptoran dinosaurs. A partial listing of the data appears in Display 10.27.

Fit a single model, with clutch volume (possibly transformed) as the response, that allows for separate straight lines in each group, using indicator variables for the different groups. For example, by selecting *Mani* as the reference group, you can determine how close any other group is to *Mani* by dropping out the indicator for that group and its product with the body mass variable. Do this for all other groups. (a) Are there differences among the relationships in the groups other than the *Mani*

DISPLAY 10.27 Body mass (in kilograms) and clutch volume (in cubic millimeters) for 443 species of animals in 6 groups (*Mat*: modern maternal-care birds, *Pat*: modern paternal-care birds, *BiP*: modern biparental-care birds, *Croc*: modern maternal-care crocodiles, *Mani*: non-avian maniraptoran dinosaurs, and *Otin*: other non-avian dinosaurs; first 5 of 443 rows

CommonName	Genus	Species	Group	BodyMass	ClutchVolume
Mallard	*Anas*	*platyrhynchos*	Mat	1.14E+00	5.42E+05
Greylag Goose	*Anser*	*anser*	Mat	3.31E+00	7.97E+05
Mute Swan	*Cygnus*	*olor*	Mat	1.07E+01	1.75E+06
Common Eider	*Somateria*	*mollissima*	Mat	2.07E+00	4.75E+05
Southern Screamer	*Chauna*	*torquata*	BiP	4.40E+00	5.13E+05

DISPLAY 10.28 First five rows of a data set with component test scores in Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, and Mathematics Knowledge taken in 1981; the AFQT score, which is a linear combination of them; and annual income in 2005 for 12,139 Americans who were selected in the NLSY79 sample, who were available for re-interview in 2006 and who had complete values of variables used in this and related analyses

Subject	Arith	Word	Parag	Math	AFQT	Income2005
2	8	15	6	6	6.841	5,500
6	30	35	15	23	99.393	65,000
7	14	27	8	11	47.412	19,000
8	13	35	12	4	44.022	36,000
9	21	28	10	13	59.683	65,000

group? What if there were none? (b) What do you conclude about which other group is nearest the maniraptoran dinosaur group? (c) Is the model defensible? (d) Comment on the study design.

32. Galton's Height Data. Reconsider the data in Exercise 7.26 on heights of adult children and their parents. Ignore the possible dependence of observations on children from the same family for now to answer the following: (a) What is an equation for predicting a child's adult height from their mother's height, their father's height, and their gender? (b) By how much does the male mean height exceed the female mean height for children whose parents' heights are the same? (c) Find a 95% prediction interval for a female whose father's height is 72 inches and whose mother's height is 64 inches.

33. IQ Score and Income. Display 10.28 is a partial listing of the National Longitudinal Study of Youth (NLSY79) subset (see Exercise 2.22) with annual incomes in 2005 (in U.S. dollars, as recorded in a 2006 interview) and scores on the Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Mathematics Knowledge portions of the Armed Forces Vocational Aptitude Battery (ASVAB) of tests taken in 1981. The AFQT is the Armed Forces Qualifying Test percentile, which is based on a linear combination of these four components and which is sometimes used as a general intelligence test score. A previous exercise had to do with the possible dependence of 2005 income on AFQT score. An interesting question is whether there might be some better linear combination of the four components than AFQT for predicting income. Investigate this by seeing whether the component scores are useful predictors of Income2006 in addition to AFQT. Also see whether

AFQT is a useful predictor in addition to the four test scores. Which test scores seem to be the most important predictors of 2006 income? (Answer with a single p-value conclusion.)

Answers to Conceptual Exercises

1. The initial heights were controlled by Galileo. Distance is the only random quantity.

2. Yes. The extra-sum-of-squares F-test, which compares the model with all three explanatory variables to the model with *lbody* only, addresses precisely this issue.

3. (a) The test where β_2 and β_3 both equal zero can be cast in terms of comparing the full model ($\beta_0 + \beta_1 lbody + \beta_2 lgest + \beta_3 llitter$) to the reduced model ($\beta_0 + \beta_1 lbody$). The t-test where β_2 is zero, on the other hand, implies a reduced model of $\beta_0 + \beta_1 lbody + \beta_3 llitter$; and the t-test where β_3 is zero implies a reduced model of $\beta_0 + \beta_1 lbody + \beta_2 lgest$. Neither of these reduced models is the same as the one sought, nor can the results from them be combined in any way to give some answer. (b) Same reason. The t-tests consider models where only one parameter is zero, but the model with both β_1 and β_2 equal to zero does not enter the picture. Multiple comparison adjustment does nothing to resolve the fact that these tests are different.

4. The model has nearly as many free parameters (8) as it has observations (9). One should expect a good fit to the data at hand, even if the explanatory variables have little relationship to the response.

5. (a) Each of the models represents the relationship between *lenergy* and *lmass* as a straight line, within the groups. The first model says that the intercept and slope—and hence the full line—is the same in all groups. The second model says that the slope is the same in each model while the intercepts are different. The third model allows the slopes and the intercepts to differ among all groups. (b) The second model is a reduced model in a test for equal slopes (with possibly differing intercepts); it is the full model for a test of equal intercepts (given equal slopes).

6. No bats of comparable size could be found in both groups. A more correct wording here might be "after adjustment for body size," but the underlying difficulty is not avoided.

7. (d).

8. The reduced model takes the null hypothesis to be true. The full model, however, encompasses both the null hypothesis and the alternative hypothesis. Thus, the full model is also correct when the reduced model is correct. Put another way, the full model is thought to be adequate from the start; the reduced model is obtained by imposing the constraints of the null hypothesis on the full model.