received 100% breast milk than for those who received 50% breast milk, after accounting for the other explanatory variables? (ii) What is the importance of the percentage of breast milk variable in dealing with confounding variables?

11. Glasgow Graveyards. Do persons of higher socioeconomic standing tend to live longer? This was addressed by George Davey Smith and colleagues through the relationship of the heights of commemoration obelisks and the life lengths of the corresponding grave site occupants. In burial grounds in Glasgow a certain design of obelisk is quite prevalent, but the heights vary greatly. Since the height would influence the cost of the obelisk, it is reasonable to believe that height is related to socioeconomic status. The researchers recorded obelisk height, year of death, age at death, and gender for 1,349 individuals who died prior to 1921. Although they were interested in the relationship between mean life length and obelisk height, it is important that they included year of construction as an explanatory variable since life lengths tended to increase over the years represented (1801 to 1920). For males, they fit the regression of life length on obelisk height (in meters) and year of obelisk construction and found the coefficient of obelisk height to be 1.93. For females they fit the same regression and found the coefficient of obelisk height to be 2.92. (Data from Smith *et al.*, "Socioeconomic Differentials in Mortality: Evidence from Glasgow Graveyards," *British Medical Journal* 305 (1992): 1557–60.)

(a) After accounting for year of obelisk construction, each extra meter in obelisk height is associated with Z extra years in mean lifetime. What is the estimated Z for males? What is the estimated Z for females?

(b) Since the coefficients differ significantly from zero, would it be wise for an individual to build an extremely tall obelisk, to ensure a long life time?

(c) The data were collected from eight different graveyards in Glasgow. Since there is a potential blocking effect due to the different graveyards, it might be appropriate to include a graveyard effect in the model. How can this be done?

Computational Exercises

12. Mammal Brain Weights. (a) Draw a matrix of scatterplots for the mammal brain weight data (Display 9.4) with all variables transformed to their logarithms (to reproduce Display 9.11). (b) Fit the multiple linear regression of log brain weight on log body weight, log gestation, and log litter size, to confirm the estimates in Display 9.15. (c) Draw a matrix of scatterplots as in (a) but with litter size on its natural scale (untransformed). Does the relationship between log brain weight and litter size appear to be any better or any worse (more like a straight line) than the relationship between log brain weight and log litter size?

13. Meat Processing. One way to check on the adequacy of a linear regression is to try to include an X-squared term in the model to see if there is significant curvature. Use this technique on the meat processing data of Section 7.1.2. (a) Fit the multiple regression of pH on hour and hour-squared. Is the coefficient of hour-squared significantly different from zero? What is the p-value? (b) Fit the multiple regression of pH on log(hour) and the square of log(hour). Is the coefficient of the squared-term significantly different from zero? What is the p-value? (c) Does this exercise suggest a potential way of checking the appropriateness of taking the logarithm of X or of leaving it untransformed?

14. Pace of Life and Heart Disease. Some believe that individuals with a constant sense of time urgency (often called type-A behavior) are more susceptible to heart disease than are more relaxed individuals. Although most studies of this issue have focused on individuals, some psychologists have investigated geographical areas. They considered the relationship of city-wide heart disease rates and general measures of the pace of life in the city.

For each region of the United States (Northeast, Midwest, South, and West) they selected three large metropolitan areas, three medium-size cities, and three smaller cities. In each city they measured three

20 adults from each group. Display 9.19 shows average wing size in millimeters on a logarithmic scale, and average ratios of basal lengths to wing size.

(a) Construct a scatter plot of average wing size against latitude, in which the four groups defined by continent and sex are coded differently. Do these suggest that the wing sizes of the NA flies have evolved toward the same cline as in EU?

(b) Construct a multiple linear regression model with wing size as the response, with latitude as a linear explanatory variable, and with indicator variables to distinguish the sexes and continents. As there are four groups, you will want to have three indicator variables: the continent indicator, the sex indicator, and the product of the two. Construct the model in such a way that one parameter measures the difference between the slopes of the wing size versus latitude regressions of NA and EU for males, one measures the difference between the NA–EU slope difference for females and that for males, one measures the difference between the intercepts of the regressions of NA and EU for males, and one measures the difference between the NA–EU intercepts' difference for females and that for males.

19. Depression and Education. Has homework got you depressed? It could be worse. Depression, like other illnesses, is more prevalent among adults with less education than you have.

R. A. Miech and M. J. Shanahan investigated the association of depression with age and education, based on a 1990 nationwide (U.S.) telephone survey of 2,031 adults aged 18 to 90. Of particular interest was their finding that the association of depression with education strengthens with increasing age—a phenomenon they called the "divergence hypothesis."

They constructed a depression score from responses to several related questions. Education was categorized as (i) college degree, (ii) high school degree plus some college, or (iii) high school degree only. (See "Socioeconomic Status and Depression over the Life Course," *Journal of Health and Social Behaviour* 41(2) (June, 2000): 162–74.)

(a) Construct a multiple linear regression model in which the mean depression score changes linearly with age in all three education categories, with possibly unequal slopes and intercepts. Identify a single parameter that measures the diverging gap between categories (iii) and (i) with age.

(b) Modify the model to specify that the slopes of the regression lines with age are equal in categories (i) and (ii) but possibly different in category (iii). Again identify a single parameter measuring divergence.

This and other studies found evidence that the mean depression is high in the late teens, declines toward middle age, and then increases towards old age. Construct a multiple linear regression model in which the association has these characteristics, with possibly different structures in the three education categories. Can this be done in such a way that a single parameter characterizes the divergence hypothesis?

Data Problems

20. Winning Speeds at the Kentucky Derby. The Kentucky Derby is a 1.25 mile horse race held annually at the Churchill Downs race track in Louisville, Kentucky. Shown in Display 9.20 are some sample rows of a data set containing the year of the race, the winning horse, the condition of the track, and the average speed (in feet per second) of the winner, for years 1896–2000. The track conditions have been grouped into three categories: fast, good (which includes the official designations "good"and "dusty"), and slow (which includes the designations "slow," "heavy," "muddy," and "sloppy"). Use a statistical computer program to fit a model for the mean winning speed as a function of year and the track condition factor. The data are from www.kentuckyderby.com.

Display 9.20 Sample rows of the Kentucky Derby winning speeds data set

Year	Winner	Condition	Speed
1896	Ben Brush	good	51.66
1897	Typhoon II	slow	49.81
1898	Plaudit	good	51.16
1899	Manuel	fast	50.00
1900	Lieut. Gibson	fast	52.28
1901	His Eminence	fast	51.66
...			
2000	Fusaichi Pegasus	fast	54.49

Answers to Conceptual Exercises

1. (a) Let *early* = 1 if *time* = 24 and 0 if *time* = 0. Then $\mu\{flowers \mid light, early\} = \beta_0 + \beta_1 light + \beta_2 early$. (b) $\beta_0 + \beta_1 light + \beta_2 early + \beta_3 (light \times early)$.

2. The difference is 300 μmol/m^2/sec times the coefficient of *light*, or about -12.15 flowers.

3. (a) The principal reason is that the 10 plants were all treated together and grown together in the same chamber. The experimental unit is always defined as the unit that receives the treatment, here, plants in the same chamber. (b) The assumption of normality is assisted. Averages tend to have normal distributions, so the averaging may alleviate some distributional problems that could arise from looking at separate numbers of flowers.

4. No. The difficulty with interpreting regression coefficients individually, as in a controlled experiment, is that explanatory variables cannot be manipulated individually. In this instance, the sloth and the fruit bat also have different body weights—the sloth weighs 50 times what the fruit bat weighs. (The full model estimates the brain weight of the fruit bat to be only about 35% of the brain weight of the sloth.) One might attempt to envision a fruit bat having the same weight (0.9 kg) as the sloth and the same litter size (1.0), but having a gestation period of 165 instead of 145 days. This approach, however, is generally unsatisfactory because it extrapolates beyond the experience of the data set (resulting in animals like a fish-eating kangaroo with wings).

5. Yes. A common way to explore lack of fit is to introduce curvature and interaction terms to see if measured effects change as the configuration of explanatory variables changes.

6. Yes.

7. Keep your eye on the parameters. If the mean is linear in the parameters, the model is *linear*.

- **(a)** Yes, even though it is not linear in X.
- **(b)** Yes.
- **(c)** No. Both numerator and denominator are linear in parameters and X, but the whole is not.
- **(d)** No. This is a very useful model, however.

8. In both, σ is a measure of the magnitude of the difference between a response and the mean in the population from which the response was drawn. In the meadowfoam problem, σ measures the typical size of differences between seedling flowers (averaged from 10 plants) and the mean seedling flowers (averaged from 10 plants) treated similarly (same intensity and timing potential). In the brain weight problem, it is more difficult to describe what σ measures because the theoretical model mass for (i) non-echolocating bats, (ii) non-echolocating birds, and (iii) echolocating bats? How 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?

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	1125	110	112.5	1125	115
0	1500	115	112.5	1500	108
37.5	0	202	150	0	133
37.5	375	161	150	375	125
37.5	750	172	150	750	184
37.5	1125	138	150	1125	135
37.5	1500	133	150	1500	114
75	0	204			
75	375	165			
75	750	148			
75	1125	143			
75	1500	123			

15. **Old Faithful.** Reconsider the Old Faithful eruption durations and intervals in Display 7.14. Fit the regression of Interval on Duration and Day treated as a factor (include seven indicator variables to distinguish the eight days). Obtain the analysis of variance table, and then fit the regression of interval on duration alone to obtain the analysis of variance table for this reduced model. Use the quantities listed in the tables to construct an F-statistic for the test of whether any difference in mean intervals is due to the particular day of recording. Find the p-value (or, at least, say whether the p-value is bigger than .05, between .05 and .01, or less than .01).

16. **Galileo's Data.** Use Galileo's data in Display 10.1 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.

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. 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 25,632 males between the age of 18 and 70 who worked full-time, along with their years of education, years of experience, an indicator variable for whether they were black, an indicator variable for 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 were discussed in the paper by H. J. Bierens and D. K. Ginther "Integrated Conditional Moment Testing of Quantile Regression Models," 2000, to appear in a special issue of *Empirical Economics* on Economic Applications of Quantile Regression; and made available at the web site http://econ.la.psu.edu/ hbierens/MEDIAN.HTM associated with the software EasyReg.)

Display 10.25 Data on the first 6 indivituals (out of 25,632) in the wage and race data set

Weekly wage ($)	Education (years)	Experience (years)	Indicator for Black	Indicator for SMSA	Region (ne,mw,s,w)
354.94	7	45	0	1	ne
123.46	12	1	0	1	ne
370.37	9	9	0	1	ne
754.94	11	46	0	1	ne
593.54	12	36	0	1	ne
377.23	16	22	0	1	ne
. . .					

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.