DISPLAY 9.14 Estimates of regression coefficients in the multiple regression of *flowers* on *light*, *early*, and *light* × *early*—the meadowfoam study

Variable	Coefficient	Standard error	*t*-statistic	*p*-value
Constant	71.6233	4.3433	16.4905	<0.0001
light	−0.0411	0.0074	5.5247	<0.0001
early	11.5233	6.1424	1.8760	0.0753
light × *early*	0.0012	0.0105	0.1150	0.9096

DISPLAY 9.15 Estimates of regression coefficients in the multiple regression of log brain weight on log body weight, log gestation, and log litter size—brain weight data

Variable	Coefficient	Standard error	*t*-statistic	*p*-value
Constant	0.8548	0.6617	1.2919	0.1996
lbody	0.5751	0.0326	17.6468	<0.0001
lgest	0.4179	0.1408	2.9687	0.0038
llitter	−0.3101	0.1159	2.6747	0.0089

more efficient use of the available experimental units with the multifactor arrangement (resulting in smaller standard errors for parameters of interest); and the results for the multifactor experiment are more general, since each treatment is investigated at several levels of the other treatment.

9.7 SUMMARY

Multiple regression analysis refers to a large set of tools associated with developing and using regression models for answering questions of interest. Multiple regression models describe the mean of a single response variable as a function of several explanatory variables. Transformations, indicator variables for grouped data (factors), squared explanatory variables for curvature, and product terms for interaction greatly enhance the usefulness of this model.

Substantial exploratory analysis is recommended for gaining initial insight into what the data have to say in answer to the questions of interest and for suggesting possible regression models for answering them more formally. Some standard graphical procedures are presented here. The data analyst should be prepared to be creative in using the available computer tools to best display the data, while keeping in mind the questions of interest and the statistical tools that might be useful for answering them.

Brain Weight Study

Is brain weight associated with gestation period and/or litter size after accounting for the effect of body weight? This is exactly the type of question for which multiple regression is useful. The regression coefficient of gestation in the regression of brain weight on body weight and gestation describes the effect of gestation for species of roughly the same body weight. With multiple linear regression, the coefficients can be estimated from all the animals without a need for grouping into subsets of similar body weight. Initial scatterplots (or initial inspection of the data) indicate that the regression model should be formed after transforming all the variables to their logarithms.

Meadowfoam Study

A starting point for the analysis is the coded scatterplot of number of flowers per plant versus light intensity, with different codes to represent the two levels of the timing factor (Display 9.3). The plot suggests that the mean number of flowers per plant decreases with increasing light intensity, that the rate of decrease does not depend on timing, and that (for any light intensity value) a larger mean number of flowers is associated with the "before PFI" level of the timing factor. Using an indicator variable for one of the timing levels permits fitting the parallel regression lines model. Further inclusion of the interaction of timing and intensity produces a model that fits separate regression lines for each level of timing. This model permits a check on whether the regression lines are indeed parallel.

9.8 EXERCISES

Conceptual Exercises

1. Meadowfoam. (a) Write down a multiple regression model with parallel regression lines of *flowers* on *light* for the two separate levels of *time* (using an indicator variable). (b) Add a term to the model in (a) so that the regression lines are not parallel.

2. Meadowfoam. A model (without interaction) for the mean *flowers* is estimated to be $71.3058 - 0.0405light + 12.1583early$. For a fixed level of timing, what is the estimated difference between the mean *flowers* at 600 and 300 μmol/m^2/sec of *light* intensity?

3. Meadowfoam. (a) Why were the numbers of flowers from 10 plants averaged to make a response, rather than representing them as 10 different responses? (b) What assumption is assisted by averaging the numbers from the 10 plants?

4. Mammal Brain Weights. The three-toed sloth has a gestation period of 165 days. The Indian fruit bat has a gestation period of 145 days. From Display 9.14 the estimated model for the mean of log brain weight is $0.8548 + 0.5751lbody + 0.4179lgest - 0.3101llitter$. Since *lgest* for the sloth is 0.1292 more than *lgest* for the fruit bat, does this imply that an estimate of the mean log brain weight for the sloth is (0.4179)(0.1292) more than the mean log brain weight for the bat (i.e., the median is 5.5% higher)? Why? Why not?

5. Insulating Fluid (Section 8.1.2). Would it be possible to test for lack of fit to the straight line model for the regression of log breakdown time on voltage by including a voltage-squared term in the model, and testing whether the coefficient of the squared term is zero?

6. Island Area and Species. For the island area and number of species data in Section 8.1.1, would it be possible to test for lack of fit to the straight line model for the regression of log number

of species on log island area by including the square of log area in the model and testing whether its coefficient is zero?

7. Which of the following regression models are *linear*?

(a) $\mu\{Y|X\} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$
(b) $\mu\{Y|X\} = \beta_0 + \beta_1 10^X$
(c) $\mu\{Y|X\} = (\beta_0 + \beta_1 X)/(\beta_0 + \beta_2 X)$
(d) $\mu\{Y|X\} = \beta_0 \exp(\beta_1 X)$.

8. Describe what σ measures in the meadowfoam problem and in the brain weight problem.

9. Pollen Removal. Reconsider the data on proportion of pollen removed and duration of visit to the flower for bumblebee queens and honeybee workers, in Exercise 3.28. (a) Write down a model that describes the mean proportion of pollen removed as a straight-line function of duration of visit, with separate intercepts and separate slopes for bumblebee queens and honeybee workers. (b) How would you test whether the effect of duration of visit on proportion removed is the same for queens as for workers?

10. Breast Milk and IQ. In a study, intelligence quotient (IQ) test scores were obtained for 300 8-year-old children who had been part of a study of premature babies in the early 1980s. Because they were premature, all the babies were fed milk by a tube. Some of them received breast milk entirely, some received a prepared formula entirely, and some received some combination of breast milk and formula. The proportion of breast milk in the diet depended on whether the mother elected to provide breast milk and to what extent she was successful in expressing any, or enough, for the baby's diet. The researchers reported the results of the regression of the response variable—IQ at age 8—on social class (ordered from 1, the highest, to 5), mother's education (ordered from 1, the lowest, to 5), an indicator variable taking the value 1 if the child was female and 0 if male, the number of days of ventilation of the baby after birth, and an indicator variable taking the value 1 if there was any breast milk in the baby's diet and 0 if there was none. The estimates are reported in Display 9.16 along with the p-values for the tests that each coefficient is zero. (Data from Lucas et al., "Breast Milk and Subsequent Intelligence Quotient in Children Born Preterm," *Lancet* 339 (1992): 261–64).

DISPLAY 9.16 Breast milk and intelligence data

Explanatory variable	Estimated coefficient	p-value
Social class	−3.5	0.0004
Mother's education	2.0	0.01
Female indicator	4.2	0.01
Days of ventilation	−2.6	0.02
Breast milk indicator	8.3	<0.0001

(a) After accounting for the effects of social class, mother's education, whether the child was a female, and days after birth of ventilation, how much higher is the estimated mean IQ for those children who received breast milk than for those who did not?
(b) Is it appropriate to use the variables "Social class" and "Mother's education" in the regression even though in both instances the numbers 1 to 5 do not correspond to anything real but are merely ordered categories?
(c) Does it seem appropriate for the authors to simply report < 0.0001 for the p-value of the breast milk coefficient rather than the actual p-value?

(d) Previous studies on breast milk and intelligence could not separate out the effects of breast milk and the act of breast feeding (the bonding from which might encourage intellectual development of the child). How is the important confounding variable of whether a child is breast fed dealt with in this study?
(e) Why is it important to have social class and mother's education as explanatory variables?
(f) In a subsidiary analysis the researchers fit the same regression model as above except with the indicator variable for whether the child received breast milk replaced by the percentage of breast milk in the diet (between 0 and 100%). The coefficient of that variable turned out to be 0.09. (i) From this model, how much larger is the estimated mean IQ for children who 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 indicators of the pace of life. The variable *walk* is the walking speed of pedestrians over a distance of 60 feet during business hours on a clear summer day along a main downtown street. *Bank* is the average time a sample of bank clerks takes to make change for two $20 bills or to give $20 bills for change. The variable *talk* was obtained by recording responses of postal clerks explaining the difference between regular, certified, and insured mail and by dividing the total number of syllables by the time of their response. The researchers also obtained the age-adjusted death rates from ischemic heart disease (a decreased flow of blood to the heart) for each city (*heart*). The data in Display 9.17 were read from a graph in the published paper. (Data from R. V. Levine, "The Pace of Life," *American Scientist* 78 (1990): 450–9.) The variables have been standardized, so there are no units of measurement involved.

DISPLAY 9.17 First five rows of the pace-of-life data set with bank clerk speed, pedestrian walking speed, postal clerk talking speed, and age-adjusted death rates due to heart disease, in 36 cities

City	Bank	Walk	Talk	Heart
Atlanta, GA	25	27	27	19
Bakersfield, CA	29	18	25	11
Boston, MA	31	28	24	24
Buffalo, NY	30	23	23	29
Canton, OH	28	20	18	19

(a) Draw a matrix of scatterplots of the four variables. Construct it so that the bottom row of plots all have *heart* on the vertical axis. If you do not have this facility, draw scatterplots of *heart* versus each of the other variables individually.
(b) Obtain the least squares fit to the linear regression of *heart* on *bank*, *walk*, and *talk*.
(c) Plot the residuals versus the fitted values. Is there evidence that the variance of the residuals increases with increasing fitted values or that there are any outliers?
(d) Report a summary of the least squares fit. Write down the estimated equation with standard errors below each estimated coefficient.

15. Rainfall and Corn Yield. The data on corn yields and rainfall, discussed in Section 9.3.1, appear in Display 9.18. (Data from M. Ezekiel and K. A. Fox, *Methods of Correlation and Regression Analysis*, New York: John Wiley & Sons, 1959; originally from E. G. Misner, "Studies of the Relationship of Weather to the Production and Price of Farm Products, I. Corn" [mimeographed publication, Cornell University, March 1928].)

(a) Plot corn yield versus rainfall.
(b) Fit the multiple regression of corn yield on *rain* and *rain*^{2}.

DISPLAY 9.18 Partial listing of a data set with average corn yield in a year (bushels) and total rainfall (inches) in six U.S. states (1890–1927).

Year	Yield	Rainfall
1890	24.5	9.6
1891	33.7	12.9
1892	27.9	9.9
1893	27.5	8.7
1894	21.7	6.8
...		
1927	32.6	10.4

(c) Plot the residuals versus year. Is there any pattern evident in this plot? What does it mean? (Anything to do, possibly, with advances in technology?)
(d) Fit the multiple regression of corn yield on *rain*, *rain*^{2}, and *year*. Write the estimated model and report standard errors, in parentheses, below estimated coefficients. How do the coefficients of *rain* and *rain*^{2} differ from those in the estimated model in (b)? How does the estimate of σ differ? (larger or smaller?) How do the standard errors of the coefficients differ? (larger or smaller?) Describe the effect of an increase of one inch of rainfall on the mean yield over the range of rainfalls and years.
(e) Fit the multiple regression of corn yield on *rain*, *rain*^{2}, *year*, and *year* × *rain*. Is the coefficient of the interaction term significantly different from zero? Could this term be used to say something about technological improvements regarding irrigation?

16. Pollen Removal. The data in Exercise 3.27 are the proportions of pollen removed and the duration of visits on a flower for 35 bumblebee queens and 12 honeybee workers. It is of interest to understand the relationship between the proportion removed and duration and the relative pollen removal efficiency of queens and workers. (a) Draw a coded scatterplot of proportion of pollen removed versus duration of visit; use different symbols or letters as the plotting codes for queens and workers. Does it appear that the relationship between proportion removed and duration is a straight line? (b) The logit transformation is often useful for proportions between 0 and 1. If p is the proportion then the logit is $\log[p/(1-p)]$. This is the log of the ratio of the amount of pollen removed to the amount not removed. Draw a coded scatterplot of the logit versus duration. (c) Draw a coded scatterplot of the logit versus log duration. From the three plots, which transformations appear to be worthy of pursuing with a regression model? (d) Fit the multiple linear regression of the logit of the proportion of pollen removed on (i) log duration, (ii) an indicator variable for whether the bee is a queen or a worker, and (iii) a product term for the interaction of the first two explanatory variables. By examining the p-value of the interaction term, determine whether there is any evidence that the proportion of pollen depends on duration of visit differently for queens than for workers. (e) Refit the multiple regression but without the interaction term. Is there evidence that, after accounting for the amount of time on the flower, queens tend to remove a smaller proportion of pollen than workers? Why is the p-value for the significance of the indicator variable so different in this model than in the one with the interaction term?

17. Crab Claw and Force. Using the crab data from Exercise 7.22, (a) draw a scatterplot of claw closing force versus propodus height (both on a log scale), with different plotting symbols to distinguish the three different crab species, and (b) fit the multiple regression of log force on log height and species (as a factor). Provide the estimated model including standard errors of estimated

DISPLAY 9.19 First five rows of a data set with average wing sizes of male and female flies, on logarithmic scale, with standard errors; and average basal length to wing size ratios of the females, at 11 locations in North America and 10 locations in Europe

		Wing size (10^3×log mm)				Basal length to wing size (females)	
Continent	Latitude (N)	Females	SE	Males	SE	Ratio (av)	SE
NA	35.5	901	2.5	797	3.8	0.831	0.010
NA	37.0	896	3.5	806	3.0	0.834	0.014
NA	38.6	906	3.0	812	3.2	0.836	0.012
NA	40.7	907	3.5	807	3.2	0.833	0.013
NA	40.9	898	3.6	818	2.7	0.830	0.012

regression coefficients. (See Exercises 10.9 and 10.10 for analyses that explore a more sophisticated model for these data.)

18. Speed of Evolution. How fast can evolution occur in nature? Are evolutionary trajectories predictable or idiosyncratic? To answer these questions R. B. Huey et al. ("Rapid Evolution of a Geographic Cline in Size in an Introduced Fly," *Science* 287 (2000): 308–9) studied the development of a fly—*Drosophila subobscura*—that had accidentally been introduced from the Old World into North America (NA) around 1980. In Europe (EU), characteristics of the flies' wings follow a "cline"—a steady change with latitude. One decade after introduction, the NA population had spread throughout the continent, but no such cline could be found. After two decades, Huey and his team collected flies from 11 locations in western NA and native flies from 10 locations in EU at latitudes ranging from 35–55 degrees N. They maintained all samples in uniform conditions through several generations to isolate genetic differences from environmental differences. Then they measured about 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 scatterplot 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."

DISPLAY 9.20 Kentucky Derby winners, 1896–2011. *Starters* is the number of horses that started the race, *NetToWinner* is the net winnings in dollars; *Time* is the winning time in seconds; *Speed* is the winning average speed, in miles per hour; *Track* is a categorical variable describing track conditions, with seven categories: Fast, Good, Dusty, Slow, Heavy, Muddy, and Sloppy; *Conditions* is a two-category version of *Track*, with categories Fast (which combines categories Fast and Good from *Track*) and Slow (which combines all other categories of *Track*); partial listing.

Year	Winner	Starters	NetToWinner	Time	Speed	Track	Conditions
1896	Ben Brush	8	4,850	127.75	35.23	Dusty	Fast
1897	Typhoon II	6	4,850	132.50	33.96	Heavy	Slow
1898	Plaudit	4	4,850	129.00	34.88	Good	Fast
1899	Manuel	5	4,850	132.00	34.09	Fast	Fast
1900	Lieut. Gibson	7	4,850	126.25	35.64	Fast	Fast
...							
2011	Animal Kingdom	20	2,000,000	122.04	36.87	Fast	Fast

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. Kentucky Derby. Display 9.20 is a partial listing of data in file ex0920 on Kentucky Derby horse race winners from 1896 to 2011. In all those years the race was 1.25 miles in length so that winning time and speed are exactly inversely related. Nevertheless, a simple regression model for changes over time—such as a straight line model that includes *Year* or a quadratic curve that includes *Year* and $Year^2$—might work better for one of these response variables than the other. (a) Find a model for describing the mean of either winning time or winning speed as a function of year, whichever works better. (b) Quantify the amount (in seconds or miles per hour) by which the mean winning time or speed on fast tracks exceeds the mean on slow tracks (using the two-category variable *Conditions*), after accounting for the effect of year. (c) After accounting for the effects of year and track conditions, is there any evidence that the mean winning time or speed depends on number of horses in the race (*Starters*)? Is there any evidence of an interactive effect of *Starters* and *Conditions*;

DISPLAY 9.21 Dry weight (mg), ingestion rates (mg per day), and percentage of organic matter in the food, for 22 species of aquatic deposit feeders

Species	Weight	Ingestion	Organic
Hydrobia neglecta	0.20	0.57	18.0
Hydrobia ventrosa	0.20	0.86	17.0
Tubifex tubifex	0.27	0.43	29.7
Hyalella azteca	0.32	0.43	50.0
Potamopyrgus jenkinsi	0.46	2.70	14.4
Hydrobia ulvae	0.90	0.67	13.0
Nereis succinea	5.80	20.20	6.8
Pteronarcys scotti	8.40	1.49	93.0
Orchestia grillus	12.40	4.40	88.0
Arenicola grubii	20.40	240.00	2.2
Thoracophelia mucronata	40.00	230.00	1.0
Ilypolax pusilla	53.00	300.00	4.2
Uca pubnax	63.30	19.90	51.0
Scopimera globosa	65.00	50.00	23.6
Pectinaria gouldii	80.00	1,667.00	0.7
Abarenicola pacifica	380.00	3,400.00	1.2
Abarenicola claparedi	380.00	9,400.00	0.4
Arenicola marina	930.00	4,700.00	0.6
Macrophthalmus japonicus	2,050.00	4,680.00	2.1

that is, does the effect of number of horses on the response depend on whether the track was fast or slow? Describe the effect of number of horses on mean winning time or speed. (Data from Kentucky Derby: Kentucky Derby Racing Results, http://www.kentuckyderby.info/kentuckyderby-results.php (July 21, 2011).)

21. Ingestion Rates of Deposit Feeders. Aquatic deposit feeders are organisms that feed on organic material in the bottoms of lakes and oceans. Display 9.21 shows the typical dry weight in mg, the typical ingestion rate (weight of food intake per day for one animal) in mg/day, and the percentage of the food that is composed of organic matter for 19 species of deposit feeders. The organic matter is considered the "quality" part of their food. Zoologist Leon Cammen wondered if, after accounting for the size of the species as measured by the average dry weight, the amount of food intake is associated with the percentage of organic matter in the sediment. If so, that would suggest that either the animals have the ability to adjust their feeding rates according to some perception of food "quality" or that species' ingestion rates have evolved in their particular environments. Analyze the data to see whether the distribution of species' ingestion rates depends on the percentage of organic matter in the food, after accounting for the effect of species weight. Also describe the association. Notice that the values of all three variables differ by orders of magnitude among these species, so that log transformations are probably useful. (Data from L. M. Cammen, "Ingestion Rate: An Empirical Model for Aquatic Deposit Feeders and Detritivores," *Oecologia*, 44 (1980): 303–10.)

22. Mammal Lifespans. The Exercise 8.26 data set contains the mass (in kilograms), average basal metabolic rate (in kilojoules per day), and lifespan (in years) for 95 mammal species. Is metabolic rate a useful predictor of lifespan after accounting for the effect of body mass? What

DISPLAY 9.22 First five rows of a data set with AFQT intelligence test score percentile, years of education achieved by time of interview in 2006, and 2005 annual income in dollars for 1,306 males and 1,278 females in the NLSY survey

Subject	Gender	AFQT	Educ2006	Income2005
2	female	6.841	12	5,500
6	male	99.393	16	65,000
7	male	47.412	12	19,000
8	female	44.022	14	36,000
9	male	59.683	14	65,000

percentage of variation in lifespans (on the log scale) is explained by the regression on log mass alone? What additional percentage of variation is explained by metabolism? Describe the dependence of the distribution of lifespan on metabolic rate for species of similar body mass.

23. Comparing Male and Female Incomes, After Accounting for Education and IQ. Display 9.22 shows the first five rows of a subset of the National Longitudinal Study of Youth data (see Exercise 2.22) with annual incomes in 2005, intelligence test scores (AFQT) measured in 1981, and years of education completed by 2006 for 1,306 males and 1,278 females who were between the ages of 14 and 22 when selected for the survey in 1979, who were available for re-interview in 2006, and who had paying jobs in 2005. Is there any evidence that the mean salary for males exceeds the mean salary for females with the same years of education and AFQT scores? By how many dollars or by what percent is the male mean larger?

Answers to Conceptual Exercises

1. (a) Let *early* $= 1$ if *time* $= 24$ and 0 if *time* $= 0$. Then $\mu\{\dot{f}lowers \mid light, early\} = \beta_0 + \beta_1 light + \beta_2 early$. (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 invents a hypothetical subpopulation of animal species all having the same body weight, gestation, and litter size.

9. (a) $\mu\{pollen \mid duration, queen\} = \beta_0 + \beta_1 duration + \beta_2 queen + \beta_3 (duration \times queen)$, where *queen* is 1 for queens and 0 for workers (or the other way around). (b) H_0: $\beta_3 = 0$.

10. (a) 8.3 points. (b) As long as the effects are indeed linear in these coded variables, it is a useful way to include them in the multiple regression (and more powerful than considering them to be factors). (c) Yes. The evidence is overwhelming that this coefficient is not zero. (d) The use of premature babies who all had to be fed by tube makes it so that some babies received breast milk but all babies were administered their milk in exactly the same way. (e) It is possible that the decision to provide breast milk and the ability to express breast milk are related to social class and mother's education, which are likely to be related to child's IQ score. It is desired to examine the effect of breast milk that is separate from the association with these potentially confounding variables. (f) (i) 4.5 points. (ii) An important confounding variable in the previous result is the mother's decision to provide breast milk, which may be associated with good mothering skills, which may be associated with better intelligence development in the child. Using the proportion of breast milk as an explanatory variable allows the dose–response assessment of breast milk, which indicates that children of mothers who provided breast milk for 100% of the diet tended to score higher on the IQ test than children of mothers who also decided to provide breast milk but were only capable of supplying a smaller proportion of the diet.

11. (a) 1.93 years. 2.92 years. (b) No. No cause-and-effect is implied by the analysis of these observational data. (c) Seven indicator variables can be included to distinguish the eight graveyards.

CHAPTER 10

Inferential Tools for Multiple Regression

Data analysis involves finding a good-fitting model whose parameters relate to the questions of interest. Once the model has been established, the questions of interest can be investigated through the parameter estimates, with uncertainty expressed through p-values, confidence intervals, or prediction intervals, depending on the nature of the questions asked.

The primary inferential tools associated with regression analysis—t-tests and confidence intervals for single coefficients and linear combinations of coefficients, F-tests for several coefficients, and prediction intervals—are described in this chapter. As usual, the numerical calculations for these tools are performed with the help of a computer. The most difficult parts of the task are knowing what model and what inferential tool best suit the need, and knowing how to interpret and communicate the results.

The inferential tools are illustrated in this chapter on models that incorporate special explanatory variables from the previous chapter: indicator variables, quadratic terms, and interaction terms. The tests and confidence statements found in the examples, and their interpretations, are typical of the ones used in many fields and for many different kinds of data.