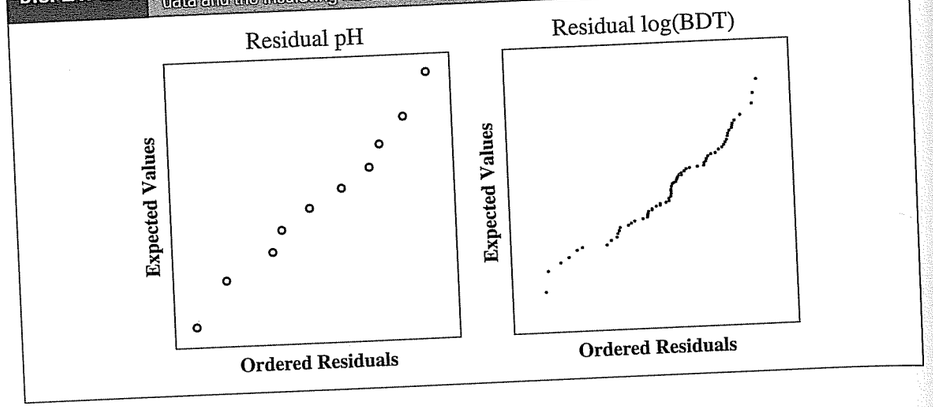
**DISPLAY 8.14** Normal probability plots of residuals from simple linear regression fits to the meat processing data and the insulating fluid data




observations in the first treatment group, five in the next, and so on. For the several-treatment experiment, balance is generally desirable in providing equal accuracy for all treatment comparisons, but it is not essential. The voltage experiment was designed as unbalanced presumably because of the much greater waiting time for breakdowns at low voltages and the primary interest in voltages between 30 and 36 kV. Balance will play a more important role when the data are cross-classified according to two factors since some simplifying formulas are only appropriate for balanced data and, more importantly, it allows unambiguous decomposition in the analysis of variance.

## 8.7 SUMMARY

Exploring statistical relationships begins with viewing scatterplots. Nonlinear regressions, nonconstant spread, and outliers can often be identified at this stage. In cases where problems are less apparent, a simple linear regression can be fit tentatively, and the decision about its appropriateness can be based on the residual plot.

When replicate response variables occur at some of the explanatory variable values, it is possible to conduct a formal lack-of-fit $F$-test. The test is a special case of the extra-sum-of-squares $F$-test for comparing two models. The models involved are the simple linear regression (reduced) model and the separate-means (full) model.



### *Insulating Fluid and Species–Area Studies*

Scatterplots and residual plots for the insulating fluid data and for the species–area data reveal nonconstant spread and nonlinear regressions, suggesting transformation of the response variable. In both cases, the spread increases as the mean level increases, indicating a logarithmic, square root, or reciprocal transformation. The scatterplot does not indicate which transformation is best. Several may be tried, with the final choice depending on what is appropriate to the scientific context of the study and to the statistical model assumptions.

After a logarithmic transformation of the times to breakdown, a simple linear regression model fits the insulating fluid data well. No evidence (from a residual plot and a lack-of-fit test) indicates lack of fit or (from the normal plot nonnormality) anything but a normal distribution of the residuals. Mean estimation and prediction can proceed from that model, with results back-transformed to the original scale. Other approaches are possible—another sensible analysis of these data assumes the Weibull distribution on the original scale and gives similar results.

To estimate the parameters in the species–area study, both response and explanatory variables are log-transformed. Here the logarithmic transformations to a simple linear regression model are indicated by theoretical model considerations. Weak evidence remains of increasing variability in the residual plot and of long-tailedness in the normal plot. These data should not, however, be used for predictions. With this small sample size, no further action is required, but confidence limits should be described as approximate.

## 8.8 EXERCISES

### Conceptual Exercises

**1. Island Area and Species Count.** The estimated regression line for the data of Section 8.1.1 is $\hat{\mu}\{\log \text{species} \mid \log \text{area}\} = 1.94 + 0.250 \log(\text{area})$. Show how this estimates that islands of area $0.5A$ have a median number of species that is 16% lower than the median number of species for islands of area $A$.

**2. Insulating Fluid.** For the insulating fluid data of Section 8.1.2 explain why the regression analysis allows for statements about the distribution of breakdown times at 27 kV while the one-way analysis of variance does not.

**3. Big Bang.** In the data set of Section 7.1.1 multiple distances were associated with a few recession velocities. Would it be possible to perform the lack-of-fit test for these data?

**4. Insulating Fluid.** If the sample correlation coefficient between the square root of breakdown time and voltage is $-0.648$, what is $R^2$ for the regression of square root of breakdown time on voltage?

**5.** Why can an $R^2$ close to 1 not be used as evidence that the simple linear regression model is appropriate?

**6.** A study is made of the stress response exhibited by a sample of 45 adults to rock music played at nine different volume levels (five adults at each level). What is the difference between using the

volume as an explanatory variable in a simple linear regression model and using the volume level as a group designator in a one-way classification model?

**7.** In a study where four levels of a magnetic resonance imaging (MRI) agent are each given to 3 cancer patients (so there are 12 patients in all), the response is a measure of the degree of seizure activity (an unpleasant side effect). The $F$-test for lack of fit to the simple linear regression model with $X$ = agent level has a $p$-value of 0.0082. The $t$-tools estimate that the effect of increasing the level of the MRI agent by 1 mg/cm$^2$ is to increase the level of seizure activity by 2.6 units (95% confidence interval from 1.8 to 3.4 units). (a) How should the latter inference be interpreted? (b) How many degrees of freedom are there for (i) the within-group variation? (ii) the lack-of-fit variation?

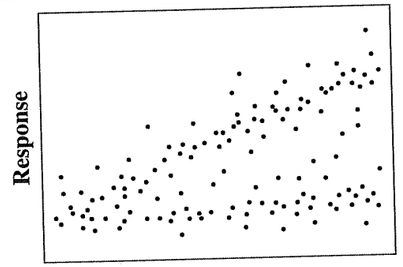
**8. Insulating Fluid.** Why would it be of interest to know whether batches of insulating fluid were *randomly* assigned to the different voltage levels?

**9.** Suppose the $(Y, X)$ pairs are: (5,1), (3,2), (4,3), (2,4), (3,5), and (1,6). Would the least squares fit to these data be much different from the least squares fit to the same data with the first pair replaced by (15,1)?

**10.** (a) What assumptions are used for exact justification of tests and confidence intervals for the slope and intercept in simple regression? (b) Are any of these assumptions relatively unimportant?

**11.** Suppose you had data on pairs $(Y, X)$ which gave the scatterplot shown in Display 8.15. How would you approach the analysis?

**DISPLAY 8.15** Scatterplot for Exercise 11




**12.** What is the technical difficulty with using the separate-means model as a basis for the lack-of-fit $F$-test when there are no replicate responses?

**13.** Researchers at a university wish to estimate the effect of class size on course comprehension. An intermediate course in statistics can be taught to classes of any size between 25 and 185 students, and four instructors are available. Suppose the researchers truly believe that the average course comprehension, measured by the average of student scores on a standardized test, is indeed a straight line in class sizes over the range from 25 to 185. What four class sizes should be used in the experiment? Why?

**14. Insulating Fluid.** Which would you use to predict the log breakdown time for a batch of insulating fluid which is to be put on test at 30 kV: the regression estimate of the mean at 30 kV or the average from the batches that were tested at 30 kV? Why?



## Computational Exercises

**15. Island Size and Species.** (a) Draw a scatterplot of the (untransformed) number of species on the (untransformed) area of the island (Display 8.2, top). (b) Fit the simple linear regression of number of species on area and obtain a residual plot. (c) What features in the two plots indicate a need for transformation?

**16. Meat Processing.** The data in Display 7.3 are a subset of the complete data on postmortum pH in 12 steer carcasses. (Data from J. R. Schwenke and G. A. Milliken, "On the Calibration Problem Extended to Nonlinear Models," *Biometrics* 47(2) (1991): 563–74). Once again, the purpose is to determine how much time after slaughter is needed to ensure that the pH reaches 6.0. In Chapter 7 the simple linear regression of pH on log(Hour) was fit to the first 10 carcasses only. Refit the model with all 12 carcasses (data given in Display 8.16). (a) Assess lack of fit using a residual plot. (b) Assess lack of fit using the lack-of-fit $F$-test. (c) The inappropriateness of the simple linear regression model can be remedied by dropping the last two carcasses. Is there any justification for doing so? (*Hint:* In order to answer the question of interest, what range of $X$'s appears to be important?)

**DISPLAY 8.16** pH of steer carcasses 1 to 24 hours after slaughter

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Animal Number:** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| **Processing Hour:** | 1 | 1 | 2 | 2 | 4 | 4 | 6 | 6 | 8 | 8 | 24 | 24 |
| **pH:** | 7.02 | 6.93 | 6.42 | 6.51 | 6.07 | 5.99 | 5.59 | 5.80 | 5.51 | 5.36 | 5.30 | 5.47 |

**17. Biological Pest Control.** In a study of the effectiveness of biological control of the exotic weed tansy ragwort, researchers manipulated the exposure to the ragwort flea beetle on 15 plots that had been planted with a high density of ragwort. Harvesting the plots the next season, they measured the average dry mass of ragwort remaining (grams/plant) and the flea beetle load (beetles/gram of ragwort dry mass) to see if the ragwort plants in plots with high flea beetle loads were smaller as a result of herbivory by the beetles. (Data from P. McEvoy and C. Cox, "Successful Biological Control of Ragwort, *Senecio jacobaea*, by Introduced Insects in Oregon," *Ecological Applications* 1(4) (1991): 430–42. The data in Display 8.17 were read from McEvoy and Cox, Figure #2.)

**DISPLAY 8.17** Dry mass of ragwort weed on 15 plots exposed to flea beetles

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Plot #:** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| **Flea beetle load:** | 12.2 | 14.6 | 15.8 | 25.3 | 38.6 | 76.4 | 163 | 182 | 415 | 446 | 628 | 377 | 770 | 1,446 | 1,012 |
| **Ragwort mass:** | 18.2 | 17.5 | 7.22 | 30.6 | 6.66 | 6.14 | 5.21 | 0.502 | 0.611 | 0.630 | 0.427 | 0.011 | 0.012 | 0.006 | 0.002 |

**(a)** Use scatterplots of the raw data, along with trial and error, to determine transformations of $Y$ = Ragwort dry mass and of $X$ = Flea beetle load that will produce an approximate linear relationship.
**(b)** Fit a linear regression model on the transformed scale; calculate residuals and fitted values.
**(c)** Look at the residual plot. Do you want to try other transformations? What do you suggest?

**18. Distance and Order from Sun.** Reconsider the planetary distance and order from sun data in Exercise 7.21. Fit a regression model to the data that includes the asteroid belt and fill in the blanks in this conclusion: Aside from some random variation, the distance to the sun increases by __% with each consecutive planet number (95% confidence interval: ____ to ____% increase).

**19. Pollen Removal.** Reconsider the pollen removal data of Exercise 3.27 and the regression of pollen removed on time spent on flower, for the bumblebee queens only. (a) What problems are evident in the residual plot? (b) Do log transformations of $Y$ or $X$ help any? (c) Try fitting the regression only for those times less than 31 seconds (i.e., excluding the two longest times). Does this fit better? (*Note*: If the linear regression fits for a restricted range of the $X$'s, it is acceptable to fit the model with all the other $X$'s excluded and to report the range of $X$'s for which the model holds.)

**20. Quantifying Evidence for Outlierness.** In a special election to fill a Pennsylvania State Senate seat in 1993, the Democratic candidate, William Stinson, received 19,127 machine-counted votes and the Republican, Bruce Marks, received 19,691 (i.e., 564 more votes than the Democrat). In addition, however, Stinson received 1,396 absentee ballots and Marks received only 371, so the total tally showed the Democrat, Stinson, winning by 461 votes. The large disparity between the machine-counted and absentee ratios, and the resulting reversal of the outcome due to the absentee ballots, sparked concern about possible illegal influence on the absentee votes. Investigators reviewed data on disparities between machine and absentee votes in prior Pennsylvania State Senate elections to see whether the 1993 disparity was larger than could be explained by historical variation. Display 8.18 shows the data in the form of percentage of absentee and machine-counted ballots cast for the Democratic candidate. The task is to clarify the unusualness of the Democratic absentee percentage in the disputed election. (a) Draw a scatterplot of Democratic percentage of absentee ballots versus Democratic percentage of machine-counted ballots. Use a separate plotting symbol to highlight the disputed election. (b) Fit the simple linear regression of absentee percentage on machine-count percentage, *excluding* the disputed election. Draw this line on the scatterplot. Also include a 95% prediction band. What does this plot reveal about the unusualness of the absentee percentage in the disputed election? (c) Find the prediction and standard error of prediction from this fit if the machine-count percentage is 49.3 (as it is for the disputed election). How many estimated standard deviations is the observed absentee percentage, 79.0, from this predicted value? Compare this answer to a $t$-distribution (with degrees of freedom equal to the residual degrees of freedom in the regression fit) to obtain a $p$-value. (d) **Outliers and data snooping.** The $p$-value in (c) makes sense if the investigation into the 1993 election was prompted by some other reason. Since it was prompted because the absentee percentage seemed too high, however, the $p$-value in (c) should be adjusted for data snooping. Adjust the $p$-value with a Bonferroni correction to account for all 22 residuals that could have been similarly considered. (Data from Orley Ashenfelter, 1994. Report on Expected Absentee Ballots. Typescript. Department of Economics, Princeton University. See also Simon Jackman (2011). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University. Stanford, California. R package version 1.03.10. URL http://pscl.stanford.edu/)

**DISPLAY 8.18** Partial listing of 22 Pennsylvania state senate election results, collected to explore the peculiarity noticed in the 1993 election between William Stinson and Bruce Marks (election number 22 in the data set). The full data set includes the year of the election, the senate district, the number of absentee ballots cast for the Democratic candidate and for the Republican candidate, and the number of machine-counted ballots cast for the Democratic candidate and for the Republican candidate. Also shown are the percentages of absentee and machine ballots cast for the Democratic candidate.

| Election | Year | District | DemPctOfAbsenteeVotes | DemPctOfMachineVotes | Disputed |
|---|---|---|---|---|---|
| 1 | 82 | D2 | 72.9 | 69.1 | no |
| 2 | 82 | D4 | 65.6 | 60.9 | no |
| 3 | 82 | D8 | 74.6 | 80.8 | no |
| 4 | 84 | D1 | 64.0 | 60.0 | no |
| 5 | 84 | D3 | 83.3 | 92.4 | no |
| ... | | | | | |
| 22 | 93 | D2 | 79.0 | 49.3 | yes |



**21. Fish Preferences.** Reconsider Case Study 2 in Chapter 6, the study of female preferences among platyfish. Fit the full model in which the mean preference for the yellowtailed male is possibly different for each male pair. Construct a normal probability plot of the residuals and a residual plot. If these suggest a transformation, make it and repeat the analysis including the linear contrast measuring the association of preference with male body size. If they suggest an outlier problem, use the inclusion/exclusion procedure to determine whether the outlying case(s) change(s) the answer to the questions of interest. Also, identify the outlying case(s) and suggest why it might be a true outlier.

## Data Problems

**22. Ecosystem Decay.** As an introduction to their study on the effect of Amazon forest clearing (data from T. E. Lovejoy, J. M. Rankin, R. O. Bierregaard, Jr., K. S. Brown, Jr., L. H. Emmons, and M. E. Van der Woot, "Ecosystem Decay of Amazon Forest Remnants," in M. H. Nitecki, ed., *Extinctions*, Chicago: University of Chicago Press, 1984) the researchers stated: "fragmentation of once continuous wild areas is a major way in which people are altering the landscape and biology of the planet." Their study takes advantage of a Brazilian requirement that 50% of the land in any development project remain in forest and tree cover. As a consequence of this requirement, "islands" of forest of various sizes remain in otherwise cleared areas. The data in Display 8.19 are the number of butterfly species in 16 such islands. Summarize the role of area in the distribution of number of butterfly species. Write a brief statistical report including a summary of statistical findings, a graphical display, and a section detailing the methods used to answer the questions of interest.

**DISPLAY 8.19** Forest patch area (hectares) and number of butterfly species found

| Reserve | Area | Species | Reserve | Area | Species |
|---|---|---|---|---|---|
| 1 | 1 | 14 | 9 | 10 | 33 |
| 2 | 1 | 50 | 10 | 10 | 53 |
| 3 | 1 | 55 | 11 | 10 | 50 |
| 4 | 1 | 34 | 12 | 100 | 110 |
| 5 | 1 | 40 | 13 | 100 | 70 |
| 6 | 1 | 57 | 14 | 100 | 119 |
| 7 | 10 | 43 | 15 | 100 | 60 |
| 8 | 10 | 103 | 16 | 1,000 | 145 |

**23. Wine Consumption and Heart Disease.** The data in Display 8.20 are the average wine consumption rates (in liters per person) and number of ischemic heart disease deaths (per 1,000 men aged 55 to 64 years old) for 18 industrialized countries. (Data from A. S. St. Leger, A. L. Cochrane, and F. Moore, "Factors Associated with Cardiac Mortality in Developed Countries with Particular Reference to the Consumption of Wine," *Lancet* (June 16, 1979): 1017–20.) Do these data suggest that the heart disease death rate is associated with average wine consumption? If so, how can that

**DISPLAY 8.20** First five rows of a data set with wine consumption (liters per person per year) and heart disease mortality rates (deaths per 1,000) in 18 countries

| Country | Wine consumption | Heart disease mortality |
|---|---|---|
| Norway | 2.8 | 6.2 |
| Scotland | 3.2 | 9.0 |
| England | 3.2 | 7.1 |
| Ireland | 3.4 | 6.8 |
| Finland | 4.3 | 10.2 |

relationship be described? Do any countries have substantially higher or lower death rates than others with similar wine consumption rates? Analyze the data and write a brief statistical report that includes a summary of statistical findings, a graphical display, and a section detailing the methods used to answer the questions of interest.

**24. Respiratory Rates for Children.** A high respiratory rate is a potential diagnostic indicator of respiratory infection in children. To judge whether a respiratory rate is truly "high," however, a physician must have a clear picture of the distribution of *normal* respiratory rates. To this end, Italian researchers measured the respiratory rates of 618 children between the ages of 15 days and 3 years. Display 8.21 shows a few rows of the data set. Analyze the data and provide a statistical summary. Include a useful plot or chart that a physician could use to assess a normal range of respiratory rate for children of any age between 0 and 3. (Data read from a graph in Rusconi et al., "Reference Values for Respiratory Rate in the First 3 Years of Life," *Pediatrics*, 94 (1994): 350–55.)

**DISPLAY 8.21** Partial listing of data on ages (months) and respiratory rates (breaths per minute) for 618 children

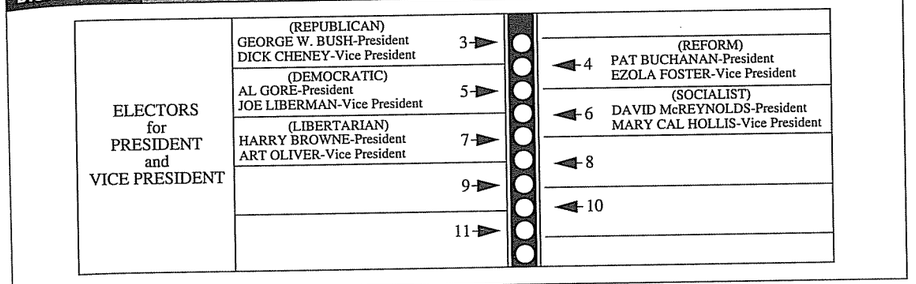
| Child | 1 | 2 | 3 | 4 | 5 | 6 | … | 618 |
|---|---|---|---|---|---|---|---|---|
| Age: | 0.1 | 0.2 | 0.3 | 0.3 | 0.3 | 0.4 | … | 36.0 |
| Rate: | 53 | 38 | 58 | 52 | 42 | 62 | … | 31 |

**25. The Dramatic U.S. Presidential Election of 2000.** The U.S. presidential election of November 7, 2000, was one of the closest in history. As returns were counted on election night it became clear that the outcome in the state of Florida would determine the next president. At one point in the evening, television networks projected that the state was carried by the Democratic nominee, Al Gore, but a retraction of the projection followed a few hours later. Then, early in the morning of November 8, the networks projected that the Republican nominee, George W. Bush, had carried Florida and won the presidency. Gore called Bush to concede. While en route to his concession speech, though, the Florida count changed rapidly in his favor. The networks once again reversed their projection, and Gore called Bush to retract his concession. When the roughly 6 million Florida votes had been counted, Bush was shown to be leading by only 1,738, and the narrow margin triggered an automatic recount. The recount, completed in the evening of November 9, showed Bush's lead to be less than 400.

Meanwhile, angry Democratic voters in Palm Beach County complained that a confusing "butterfly" lay-out ballot caused them to accidentally vote for the Reform Party candidate Pat Buchanan instead of Gore. The ballot, as illustrated in Display 8.22, listed presidential candidates on both a left-hand and a right-hand page. Voters were to register their vote by punching the circle



**DISPLAY 8.22** Confusing ballot in Palm Beach County, Florida




corresponding to their choice, from the column of circles between the pages. It was suspected that since Bush's name was listed first on the left-hand page, Bush voters likely selected the first circle. Since Gore's name was listed second on the left-hand side, many voters—who already knew who they wished to vote for—did not bother examining the right-hand side and consequently selected the second circle in the column; the one actually corresponding to Buchanan. Two pieces of evidence supported this claim: Buchanan had an unusually high percentage of the vote in that county, and an unusually large number of ballots (19,000) were discarded because voters had marked two circles (possibly by inadvertently voting for Buchanan and then trying to correct the mistake by then voting for Gore).

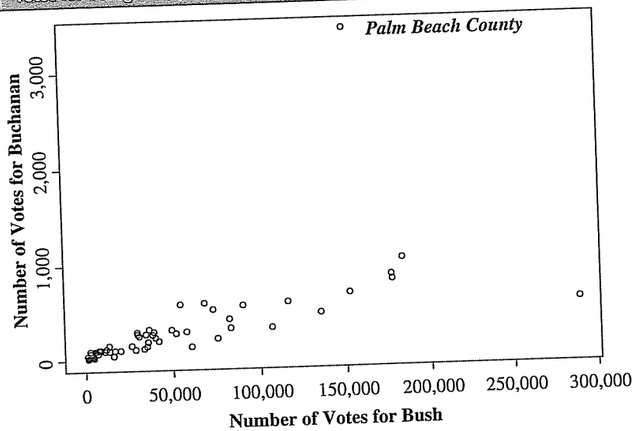
Display 8.23 shows the first few rows of a data set containing the numbers of votes for Buchanan and Bush in all 67 counties in Florida. What evidence is there in the scatterplot of Display 8.24 that Buchanan received more votes than expected in Palm Beach County? Analyze the data without Palm Beach County results to obtain an equation for predicting Buchanan votes from Bush votes. Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach from this result—assuming the relationship is the same in this county as in the others. If it is assumed that Buchanan's actual count contains a number of votes intended for Gore, what can be said about the likely size of this number from the prediction interval? (Consider transformation.)

**DISPLAY 8.23** Votes for Bush and Buchanan in all Florida counties (first 5 of 67 rows)

| County | Bush votes | Buchanan votes |
|---|---|---|
| Alachua | 34,062 | 262 |
| Baker | 5,610 | 73 |
| Bay | 38,637 | 248 |
| Bradford | 5,413 | 65 |
| Brevard | 115,185 | 570 |

**26. Kleiber's Law.** Display 8.25 shows the first five rows of a data set with average mass, metabolic rate, and average lifespan for 95 species of mammals. (From A. T. Atanasov, "The Linear Allometric Relationship Between Total Metabolic Energy per Life Span and Body Mass of Mammals," *Biosystems* 90 (2007): 224–33.) Kleiber's law states that the metabolic rate of an animal species, on

**DISPLAY 8.24** Votes for George W. Bush and Pat Buchanan in all Florida counties




**DISPLAY 8.25** Average mass (kg), average basal metabolic rate (kJ per day), and lifespan (years) for 95 mammal species; first 5 of 95 rows

| CommonName | Species | Mass | Metab | Life |
|---|---|---|---|---|
| Echidna | *Tachiglossus aculeatus* | 2.50E+00 | 3.02E+02 | 14 |
| Long-beaked echidna | *Zaglossus bruijni* | 1.03E+01 | 5.94E+02 | 20 |
| Platypus | *Ornithorhynchus anatinus* | 1.30E+00 | 2.29E+02 | 9 |
| Opossum | *Lutreolina crassicaudata* | 8.12E−01 | 1.96E+02 | 5 |
| South American opossum | *Didelphis marsupialis* | 1.33E+00 | 2.99E+02 | 6 |

average, is proportional to its mass raised to the power of 3/4. Judge the adequacy of this theory with these data.

**27. Metabolic Rate and Lifespan.** It has been suggested that metabolic rate is one of the best single predictors of species lifespan. Analyze the data in Exercise 26 to describe an equation for predicting mammal lifespan from metabolic rate. Also provide a measure of the amount of variation in the distribution of mammal lifespans that can be explained by metabolic rate.

**28. IQ, Education, and Future Income.** Display 8.26 is a partial listing of a data set with IQ scores in 1981, years of education completed by 2006, and annual income in 2005 for 2,584 Americans who were selected for the National Longitudinal Study of Youth in 1979 (NLSY79), who were re-interviewed in 2006, and who had paying jobs in 2005. (See Exercises 2.22 and 3.30 for a more detailed description of the survey.) (a) Describe the distribution of 2005 income as a function of IQ test score. What percentage of variation in the distribution is explained by the regression? (b) Describe the distribution of 2005 income as a function of years of education. What percentage of variation in the distribution is explained by the regression?



**DISPLAY 8.26** IQ test score from 1981 (AFQT—armed forces qualifying test score), number of years of education, and annual income in 2005 (dollars) for 2,584 Americans in the NLSY79 survey; first 5 of 2,584 rows

| Subject | AFQT | Educ | Income2005 |
|---|---|---|---|
| 2 | 6.841 | 12 | 5,500 |
| 6 | 99.393 | 16 | 65,000 |
| 7 | 47.412 | 12 | 19,000 |
| 8 | 44.022 | 14 | 36,000 |
| 9 | 59.683 | 14 | 65,000 |

**DISPLAY 8.27** Autism prevalence per 10,000 ten-year olds in each of five years

| Year | Prevalence |
|---|---|
| 1992 | 3.5 |
| 1994 | 5.3 |
| 1996 | 7.8 |
| 1998 | 11.8 |
| 2000 | 18.3 |

**29. Autism Rates.** Display 8.27 shows the prevalence of autism per 10,000 ten-year old children in the United States in each of five years. Analyze the data to describe the change in the distribution of autism prevalence per year in this time period. (Data from C. J. Newschaffer, M. D. Falb, and J. G. Gurney, "National Autism Prevalence Trends From United States Special Education Data," *Pediatrics*, 115 (2005): e277–e282.)

## Answers to Conceptual Exercises

**1.** Median {species|area} $= \exp(1.94)\text{area}^{0.250}$. So Median{species | 0.5 area}/Median{species | area} $= 0.5^{.250} = 0.84$. Thus Median{species | 0.5 area} $= 0.84$ Median{species | area} and, finally, [Median{species | area} − Median{species | 0.5 area}]/Median{species | area} $= 1 - 0.84 = 0.16$.

**2.** The ANOVA model states that there are seven means, one for each voltage level tested, but does not describe any relation between mean and voltage. The regression model establishes a pattern between mean log breakdown time and voltage, for all voltages in the range of 26 to 38 kV.

**3.** Yes. (*Note*: The multiple observations occurred because of rounding, so they do not represent repeated draws from the same distribution. Nevertheless, they are near replicates, at least, and can be used as such for the lack-of-fit test.)

**4.** $R^2$ = square of correlation coefficient = $(-0.648)^2 = 0.420$.

**5.** Although a high $R^2$ reflects a strong degree of linear association, this linear association may well be accompanied by curvature (and by nonconstant variance).

**6.** In the simple linear regression model, the nine mean stress levels lie on a straight line against volume. In the one-way classification (the separate-means model) the mean stress levels may or may not lie on the straight line—their values are not restricted.

**7.** (a) If the data do not fit the model, then the parameters of the model are not adequate descriptive summaries. No inference should be drawn, at least until a better model is found. (b) (i) 8; (ii) 2.

**8.** Even in laboratory circumstances, confounding variables are possible. If, for example, batches are spooned from a large container that has density stratification, then assigning consecutive batches to the lowest voltage, then the next lowest, and so on, will confound fluid density with voltage. Randomization never hurts, and usually helps.

**9.** Yes, very much so. The least squares method is not resistant to the effects of outliers.

**10.** (a) linearity, constant variance, normality, independence. (b) normality.

**11.** It may appear that this is a case where the spread of $Y$ increases as the mean of $Y$ does, so a simple transformation may be in order. This will not produce good results. Notice that the average $Y$ at a large $X$ lies in a no-man's land with few observations. Looking at the distributions in strips, you will see two separate groups in each distribution. Look for some important characteristic that separates the data into two groups, then build a separate regression model for each group.

**12.** In fitting the separate-means model, all residuals are zero. The denominator of the $F$-statistic is zero, so the $F$-statistic is not defined.

**13.** Put two classes at 25 students and two at 185 students. This makes the standard deviation in the sampling distribution of the slope parameter as small as possible, given the constraints of the situation. (But it gives no information on lack of fit; check the degrees of freedom.)

**14.** The regression estimate. It is more precise (see Display 8.11).

CHAPTER 9

# Multiple Regression

Multiple regression analysis is one of the most widely used statistical tools, and for good reason: It is remarkably effective for answering questions involving many variables. Although more difficult to visualize than simple regression, multiple regression is a straightforward extension. It models the mean of a response variable as a function of *several* explanatory variables.

Many issues, tools, and strategies are associated with multiple regression analysis, as discussed in this and the next three chapters. This chapter focuses on the meaning of the regression model and strategies for analysis. The details of estimation and inferential tools are deliberately postponed until the next chapter so that the student may concentrate first on understanding regression coefficients and the types of data structures that may be analyzed with multiple regression analysis.