a straight line, as predicted by the Big Bang theory. The least squares method provides a way to estimate the slope and intercept in this regression. Two questions are asked: is the intercept zero, as predicted by the theory, and what is the slope; which, according to the theory, is the age of the universe? The evidence in answer to the first question is expressed through a p-value for the test that β_0 is zero. The evidence in answer to the second is expressed through a confidence interval for β_1.

Meat Processing Analysis

Even though the question of interest calls for finding the value of time after slaughter at which pH is 6, it is necessary to specify pH to be the response and time (log hours after slaughter) to be the explanatory variable, since the latter was controlled by the researchers. The regression of pH on log time is well approximated by a straight line, at least for times up to 8 hours. The calibration-like question of interest requires the data analyst to find the value of X (log time) at which Y (pH) is predicted to be 6.

7.7 EXERCISES

Conceptual Exercises

1. Big Bang Data. Can the estimated regression equation be used to make inferences about (a) the mean distance of nebulae whose recession velocities are zero? (b) the mean distance of nebulae whose recession velocities are 1,000 km/sec? (c) the mean distance of nebulae whose velocities are 2,000 km/sec? (See Display 7.1.)

2. Big Bang Data. Explain why improved measurement of distance would lead to more precise estimates of the regression coefficients.

3. Meat Processing Data. By inspecting Display 7.4, describe the distribution of pH's for steer carcasses 1.65 hours after slaughter (where $X = \log(1.65) = 0.5$).

4. What is wrong with this formulation of the regression model: $Y = \beta_0 + \beta_1 X$?

5. Consider a regression of weight (kg) on height (cm) for a sample of adult males. What are the units of measurement of (a) the intercept? (b) the slope? (c) the SD? (d) the correlation?

6. Explain the differences between the following terms: regression, regression model, and simple linear regression model.

7. What assumptions are made about the distribution of the explanatory variable in the normal simple linear regression model?

8. A group of children were all given a dexterity test in their fifth grade physical education class. The teacher noted a small group who performed exceptionally well, and she informed the grade six teacher as to which children were in the group. When the same children were given a similar dexterity test the next year, they performed reasonably well, but not as well as the sixth grade teacher had expected. What might have caused this?

9. Consider the regression of weight on height for a sample of adult males. Suppose the intercept is 5 kg. (a) Does this imply that males of height 0 weigh 5 kg, on average? (b) Would this imply that the simple linear regression model is meaningless?

10. (a) At what value of X will there be the most precise estimate of the mean of Y? (b) At what value of X will there be the most precise prediction of a future Y?

11. What is the standard error of prediction as the sample size approaches infinity?

Computational Exercises

12. Suppose that the fit to the simple linear regression of Y on X from 6 observations produces the following residuals: -3.3, 2.1, -4.0, -1.5, 5.1, 1.6. (a) What is the estimate of σ^2? (b) What is the estimate of σ? (c) What are the degrees of freedom?

13. Big Bang Data. Using the results shown in Display 7.9, find a 95% confidence interval for the intercept in the regression of measured distance on recession velocity.

14. Big Bang Data. Using the results in Display 7.9, find a 95% confidence interval for the slope in the regression of measured distance on recession velocity.

15. Pollen Removal. Reconsider the data in Exercise 3.27. (a) Draw a scatterplot of proportion of pollen removed versus duration of visit, for the bumblebee queens. (b) Fit the simple linear regression of proportion of pollen removed on duration of visit. Draw the estimated regression line on the scatterplot. (Do this with the computer, if possible; otherwise draw the line with pencil and ruler.) Is there any indication of a problem with this fit? (This problem will be continued in the next chapter.)

16. Most computational procedures utilize the following identities:

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2$$

and

$$\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right).$$

The left-hand expressions are needed for calculating $\hat{\beta}_1$, but the right-hand equivalents require less effort. Using the meat processing data, calculate the left-hand versions. Then calculate the right-hand versions. Did you get the same answers? Which version requires fewer steps? Which version requires less storage in a computer?

17. Meat Processing. (a) Enter the data from Display 7.3 into a computer and find the least squares estimates for the simple linear regression of pH on log hours. (b) Noting that the average and sample standard deviation of X are provided in Display 7.12, calculate the estimated mean pH at 5 hours after slaughter, and the standard error of the estimated mean (using the formula in Section 7.4.2). (c) Find the standard error of the estimated mean pH at 5 hours after slaughter using the computer trick described in Section 7.4.2.

18. Meat Processing. (a) Find the standard error of prediction for the prediction of pH at 5 hours after slaughter. (b) Construct a 95% prediction interval at 5 hours after slaughter.

19. Meat Processing. Compute a 95% calibration interval (using the graphical approach) for the time at which pH of steer carcasses should be 7.

20. Meat Processing (Sample Size Determination). The standard error of the estimated slope based on the 10 data points is 0.0344. Using the formula for SE in Section 7.3.5, and supposing that the spread of the X's and the estimate of σ will be about the same in a future study, calculate how large the sample size would have to be in order for the SE of the estimated slope to be 0.01.

21. Planetary Distances and Order from the Sun. The first three columns in Display 7.14 show the distances from the sun (scaled so that earth = 1) and the order from the sun of the 8 planets in

DISPLAY 7.14 Orders and distances from the sun (in astronomical units, so that the distance from earth to the sun is 1) of planets in our solar system, without and with the asteroid belt

Name	Order	Distance	Name 2	Order 2	Distance 2
Mercury	1	0.387	Mercury	1	0.387
Venus	2	0.723	Venus	2	0.723
Earth	3	1	Earth	3	1
Mars	4	1.524	Mars	4	1.524
Jupiter	5	5.203	(asteroids)	5	2.9
Saturn	6	9.546	Jupiter	6	5.203
Uranus	7	19.2	Saturn	7	9.546
Neptune	8	30.09	Uranus	8	19.2
Pluto	9	39.5	Neptune	9	30.09
			Pluto	10	39.5

our solar system and the dwarf planet, Pluto. The last three columns are the same but also include the asteroid belt. Using the first three columns, (a) draw a scatterplot of log of distance versus order, (b) include the least squares estimated simple linear regression line on the plot, (c) find the estimate of σ from the least squares fit, and (d) draw a scatterplot of the residuals versus the fitted values from this fit. Using the last three columns, (e) draw a scatterplot of log of distance versus order, (f) include the least squares estimated simple linear regression line on the plot, (g) find the estimate of σ from the least squares fit, and (h) draw a scatter plot of the residuals versus the fitted values from this fit. (i) Does it appear that the simple linear (straight line) regression model fits better to the first set of 9 planets or the second set of 10 "planets"? Explain.

22. Crab Claw Size and Force. As part of a study of the effects of predatory intertidal crab species on snail populations, researchers measured the mean closing forces and the propodus heights of the claws on several crabs of three species. Their data (read from their Figure 3) appear in Display 7.15. (Data from S. B. Yamada and E. G. Boulding, "Claw Morphology, Prey Size Selection and Foraging Efficiency in Generalist and Specialist Shell-Breaking Crabs," *Journal of Experimental Marine Biology and Ecology*, 220 (1998): 191–211.)

(a) Estimate the slope in the simple linear regression of log force on log height, separately for each crab species. Obtain the standard errors of the estimated slopes.

(b) Use a t-test to compare the slopes for *C. productus* and *L. bellus*. Then compare the slopes for *C. productus* and *H. nudus*. The standard error for the difference in two slope estimates from independent samples is the following:

$$\text{SE}[\hat{\beta}_{1(1)} - \hat{\beta}_{1(2)}] = \sqrt{[\text{SE}(\hat{\beta}_{1(1)})]^2 + [\text{SE}(\hat{\beta}_{1(2)})]^2},$$

where $\hat{\beta}_{1(j)}$ represents the estimate of slope from sample j. Use t-tests with the sum of the degrees of freedom associated with the two standard errors. What do you conclude? (*Note*: A better way to perform this test, using multiple regression, is described in Chapter 9.)

23. For Those Literate in Calculus and Linear Algebra. The least squares problem is that of finding estimates of β_0 and β_1 that minimize the sum of squares,

$$\text{SS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

DISPLAY 7.15 Closing force (in newtons) and propodus heights (in mm) in three predatory crab species

Hemigrapsus nudus ($n=14$)		Lophopanopeus bellus ($n=12$)		Cancer productus ($n=12$)	
Force	Height	Force	Height	Force	Height
3.2	5.0	2.1	5.1	5.0	6.7
6.4	6.0	8.7	5.9	7.8	7.1
2.0	6.4	2.9	6.6	14.6	11.2
2.0	6.5	6.9	7.2	16.8	11.4
4.9	6.6	8.7	8.6	17.7	9.4
3.0	7.0	15.1	7.9	19.8	10.7
2.9	7.9	14.6	8.1	19.6	13.1
9.5	7.9	17.6	9.6	22.5	9.4
4.0	8.0	20.6	10.2	23.6	11.6
3.4	8.2	19.6	10.5	24.4	10.2
7.4	8.3	27.4	8.2	26.0	12.5
2.4	8.8	29.4	11.0	29.4	11.8
4.0	12.1				
5.2	12.2				

(a) Setting the partial derivatives of $\text{SS}(\beta_0, \beta_1)$ with respect to each parameter equal to zero, show that β_0 and β_1 must satisfy the *normal equations*:

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

$$\beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i{}^2 = \sum_{i=1}^{n} X_i Y_i.$$

(b) Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ given in Section 7.3.1 satisfy the normal equations. (c) Verify that the solutions provide a minimum to the sum of squares.

24. Decline in Male Births. Display 7.16 shows the proportion of male births in Denmark, The Netherlands, Canada, and the United States for a number of years. (Data read from graphs in Davis et al., "Reduced Ratio of Male to Female Births in Several Industrial Countries," *Journal of the American Medical Association*, 279 (1998): 1018–23.) Also shown are the results of least squares fitting to the simple linear regression of proportion of males on year, separately for each country, with standard errors of estimated coefficients in parentheses.

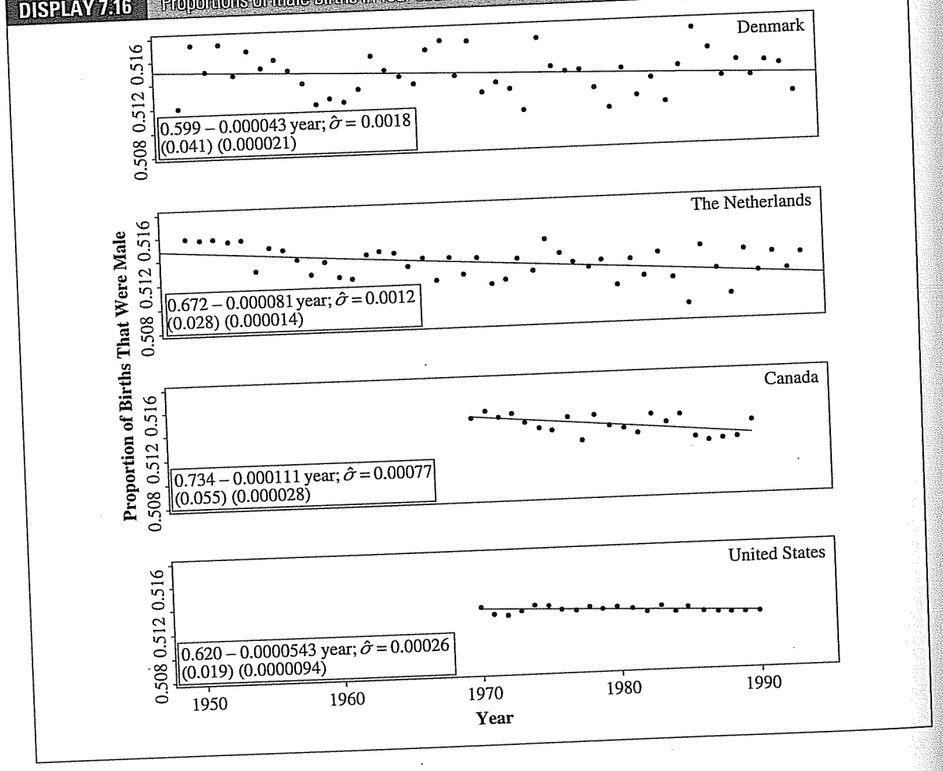
(a) With a statistical computer package and the data in the file ex0725, obtain the least squares fits to the four simple regressions, individually, to confirm the estimates and standard errors presented in Display 7.17.

(b) Obtain the t-statistic for the test that the slopes of the regressions are zero, for each of the four countries. Is there evidence that the proportion of male births is truly declining?

(c) Explain why the United States can have the largest of the four t-statistics (in absolute value) even though its slope is only the third largest (in absolute value).

(d) Explain why the standard error of the estimated slope is smaller for the United States than for Canada, even though the sample size is the same.

DISPLAY 7.16 Proportions of male births in four countries, and simple regression statistics



(e) Can you think of any reason why the standard deviations about the regression line might be different for the four countries? (*Hint*: The proportion of males is a kind of average, i.e., the average number of births that are male.)

Data Problems

25. Big Bang II. The data in Display 7.17 are measured distances and recession velocities for 10 clusters of nebulae, much farther from earth than the nebulae reported in Section 7.1.1. (Data from E. Hubble and M. Humason, "The Velocity–Distance Relation Among Extra-Galactic Nebulae," *Astrophysics Journal* 74 (1931): 43–50.) If Hubble's theory is correct, then the mean of the measured distance, as a function of velocity, should be β_1 Velocity, and β_1 is the age of the universe. Are the data consistent with the theory (that the intercept is zero)? What is the estimated age of the universe? (*Note*: The slope here is in units of megaparsecs-seconds per kilometer. Multiply by 979.8 to get an answer in billions of years. You should find out how to fit simple linear regression through the origin—that is, how to drop the intercept term—with your statistical computer package.) To what extent is the relationship shown by these far-away nebulae clusters similar to and different from the relationship indicated in Case Study 7.1.1? (Analyze the data and write a brief statistical report including a summary of statistical findings, a graphical display, and a details section describing the details of the particular methods used.)

DISPLAY 7.17 Measured distance (million parsecs) and recession velocity (km/sec) for 10 clusters of nebulae

Cluster	Distance	Velocity
Virgo	1.8	890
Pegasus	7.25	3,810
Pisces	7.00	4,638
Cancer	9.00	4,820
Perseus	11.00	5,230
Coma	13.80	7,500
Ursa Major	22.00	11,800
Leo	32.00	19,600
Isolated nebulae I	4.20	2350
Isolated nebulae II	2.15	630

DISPLAY 7.18 Galton's data on heights (in inches) of adult children and their parents; first 5 of 933 rows

Gender	Family	Height	Father	Mother
Male	1	73.2	78.5	67
Female	1	69.2	78.5	67
Female	1	69	78.5	67
Female	1	69	78.5	67
Male	2	73.5	78.5	66.5

26. Origin of the Term *Regression*. Motivated by the work of his cousin, Charles Darwin, the English scientist Francis Galton studied the degree to which human traits were passed from one generation to the next. In an 1885 study, he measured the heights of 933 adult children and their parents. Display 7.18 shows the first five rows of his data set. Galton multiplied all female heights by 1.08 to convert them to a male-equivalent height. He estimated the simple linear regression line of child's height on average parent's height and, upon finding that the slope was positive but less than 1, concluded that children of taller-than-average parents tended to also be taller than average but not as tall as their parents; and, similarly, children of shorter-than-average parents tended to be shorter than average, but not as short as their parents. He labeled this "regression towards mediocrity" because of the apparent regression (i.e., reversion) of children's height toward the average. Other scientists began to refer to the line as "Galton's regression line". Although the term was not intended to describe the model for the mean, the name stuck. Reproduce Galton's analysis by converting females' heights to their male-equivalents (multiply them by 1.08). Do the same for mothers' heights. Compute the parent height by taking the average of the father's height and the converted mother's height. Then fit the simple linear regression of child height on parent height. (For now, do as Galton did and ignore the fact that the heights of individuals from the same family are probably not independent.) Find the predicted height and a 95% prediction interval for the adult height of a boy whose average parent height is 65 inches. Repeat for a boy whose average parent height is 76 inches. (These are the raw

data used in the paper by F. Galton, "Regression towards Mediocrity in Hereditary Stature" in the *Journal of the Anthropological Institute* in 1886, as described in "'Transmuting' Women into Men: Galton's Family Data On Human Stature" by James Henley in *The American Statistician*, 58 (2004): 237–43.)

27. Male Displays. Black wheatears, *Oenanthe leucura*, are small birds of Spain and Morocco. Males of the species demonstrate an exaggerated sexual display by carrying many heavy stones to nesting cavities. This 35-gram bird transports, on average, 3.1 kg of stones per nesting season! Different males carry somewhat different sized stones, prompting a study of whether larger stones may be a signal of higher health status. M. Soler et al. ("Weight Lifting and Health Status in the Black Wheatear," *Behavioral Ecology* 10(3) (1999): 281–86) calculated the average stone mass (g) carried by each of 21 male black wheatears, along with T-cell response measurements reflecting their immune systems' strengths. The data in Display 7.19 were taken from their Figure 1. Analyze the data and write a statistical report summarizing the evidence supporting whether health—as measured by T-cell response—is associated with stone mass, and quantifying the association.

DISPLAY 7.19 Mass of stones carried and immune system strength for 21 wheatear birds, first 5 of 15 rows

Bird	Mean stone mass (g)	T-cell response (mm)
1	3.33	0.252
2	4.62	0.263
3	5.43	0.251
4	5.73	0.251
5	6.12	0.183

28. Brain Activity in Violin and String Players. Studies over the past two decades have shown that activity can effect the reorganization of the human central nervous system. For example, it is known that the part of the brain associated with activity of a finger or limb is taken over for other purposes in individuals whose limb or finger has been lost. In one study, psychologists used magnetic source imaging (MSI) to measure neuronal activity in the brains of nine string players (six violinists, two cellists, and one guitarist) and six controls who had never played a musical instrument, when the thumb and fifth finger of the left hand were exposed to mild stimulation. The researchers felt that stringed instrument players, who use the fingers of their left hand extensively, might show different behavior in the brain—as a result of this extensive physical activity—than individuals who did not play stringed instruments. Display 7.20 shows a neuron activity index from the MSI and the years that the individual had been playing a stringed instrument (zero for the controls). (Data based on a graph in Elbert et al., "Increased Cortical Representation of the Fingers of the Left Hand in String Players," *Science* 270 (13 October, 1995) 305–7.) Is the neuron activity different in the stringed musicians and the controls? Is the amount of activity associated with the number of years the individual has been playing the instrument?

29. Sampling Bias in Exit Polls. Exit pollsters predict election results before final counts are tallied by sampling voters leaving voting locations. The pollsters have no way of selecting a random sample, so they instruct their interviewers to select every third exiting voter, or fourth, or tenth, or some other specified number. Some voters refuse to participate or avoid the interviewer. If the refusers and avoiders have the same voting patterns as the rest of the population, then it shouldn't matter; the sample, although not random, wouldn't be biased. If, however, one candidate's voters are more likely to refuse or avoid interview, the sample would be biased and could lead to a misleading conclusion.

DISPLAY 7.20 Years that the individual has been playing a stringed instrument and neuronal activity index ("D5 dipole strength, in nA-m") for nine stringed musicians and six controls

Years playing	Neuron activity index
0	5
0	6
0	7.5
0	9
0	9.5
0	11
5	16
6	16.5
8	11.5
10	16
12	25
13	25.5
17	25.5
18	23
19	26.5

On November 4, 2004, exit pollsters incorrectly predicted that John Kerry would win the U.S. presidential election over George W. Bush. The exit polls overstated the Kerry advantage in 42 of 50 states. No one expects exit polls to be exact, but chance alone cannot reasonably explain this discrepancy. Although fraud is a possibility, the data are also consistent, with Bush supporters being more likely than Kerry supporters to refuse or avoid participation in the exit poll.

In a postelection evaluation, the exit polling agency investigated voter avoidance of interviewers. Display 7.21 shows the average Kerry exit poll overestimate (determined after the actual counts were available) for a large number of voting precincts, grouped according to the distance of the interviewer from the door. If Bush voters were more likely to avoid interviewers in general, one might also expect a greater avoidance with increasing distance to the interviewer (since there is more opportunity for escape). A positive relationship between distance of the interviewer from the door and amount of Kerry overestimate, therefore, would lend credibility to the theory that Bush voters were more likely to avoid exit poll interviews. How strong is the evidence that the mean Kerry overestimate increases with increasing distance of interviewer from

DISPLAY 7.21 Exit poll error in favor of Kerry and distance of exit poll interviewer from the voting precinct door, in the 2004 U.S. presidential election between George W. Bush and John Kerry

Overestimate	Distance
5.3	0
6.4	5
5.6	17
7.6	37
9.6	75
12.3	100

DISPLAY 7.22 Average exit poll interview refusal rates for precincts grouped according to the approximate age of the interviewer in the 2004 U.S. presidential election between George W. Bush and John Kerry

Age	Refusal
22	0.39
30	0.38
40	0.35
50	0.32
60	0.31
65	0.29

the door? (Data from Evaluation of Edison/Mitofsky Election System 2004 prepared by Edison Media Research and Mitofsky International for the National Election Pool (NEP), January 15, 2005. http://abcnews.go.com/images/Politics/EvaluationofEdisonMitofskyElectionSystem.pdf (accessed May 9, 2008).)

30. Sampling Bias in Exit Polls 2. This exercise is about differential interview *refusal* rates in the exit polls conducted in the 2004 U.S. presidential election. Display 7.22 shows the average proportion of voters who refused to be interviewed at precincts grouped according to the approximate age of the interviewer. What evidence do these data provide that the mean refusal rate decreased with increasing age of interviewer? An affirmative answer to this question doesn't provide any direct evidence of a difference between Kerry and Bush voters, but is consistent with an undercount of Bush votes in the exit polls if, as one might speculate based on the relative conservativeness of Bush supporters, Bush voters were more likely to avoid younger interviewers. (See Exercise 29.)

Answers to Conceptual Exercises

1. (a) Yes. (b) Yes. (c) Such an extrapolation would be risky.

2. The standard deviation σ about the regression reflects measurement error variation. Making this smaller will cause the standard deviations of the sampling distributions of the least squares estimates to be smaller (see Display 7.7).

3. The model says that the distribution is normal. The estimated mean pH is about 6.6. The prediction limits will be about 2 SDs up and down, so the SD is about 0.1 pH units.

4. This implies an exact relationship between Y and X. The model should be for the *mean* of Y as a function of X.

5. (a) kg; (b) kg/cm; (c) kg; (d) none.

6. Regression refers to the mean of a response variable as a function of an explanatory variable. A regression model is a function used to describe the regression. The simple linear regression model is a particular regression model in which the regression is a straight-line function of a single explanatory variable.

7. None.

8. This is the regression effect. It is exactly what you can expect to happen.

9. (a) No. Height = 0 is outside the range of observed values, so the model may not extend to that situation. (b) No. It may be useful for answering questions pertaining to the regression of weight on height for heights in a certain range.

10. (a) At the sample average of the X's used in the estimation. (b) Same as (a).

11. σ.

CHAPTER 8

A Closer Look at Assumptions for Simple Linear Regression

The inferential tools of the previous chapter are based on the normal simple linear regression model with constant variance. Since real data do not necessarily conform to this model, the data analyst must size up the situation and choose a course of action based on an understanding of the robustness of the tools to model violations.

This chapter presents some informal graphical tools and a formal test for assessing the lack of fit. As before, the graphical procedures are used to find suitable transformations to scales where the simple linear regression model seems appropriate. The lack-of-fit test, on the other hand, looks specifically at the issue of whether the straight line assumption is plausible.