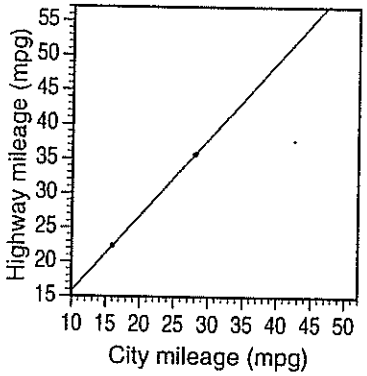
Chapter 5 Solutions

5.1. (a) The slope is 1.109. On average, highway mileage increases by 1.109 mpg for each 1 mpg change in city mileage. (b) The intercept is 4.62 mpg; this is the highway mileage for a nonexistent car that gets 0 mpg in the city. (c) With city mileage equal to 16 mpg, predicted highway mileage is $4.62 + 1.109 \times 16 \doteq 22.36$ mpg. With city mileage equal to 28 mpg, predicted highway mileage is $4.62 + 1.109 \times 28 \doteq 35.67$ mpg. (d) The graph is shown on the right. It can be drawn by drawing a line between any two points on the line; the two marked points are the two predictions computed in part (c). Because both variables are in units of mpg, the vertical and horizontal scales on this graph are the same. This is not a crucial detail, but it has the benefit of making the slope look "right"; that is, the line is slightly steeper than a line with a slope of 1.



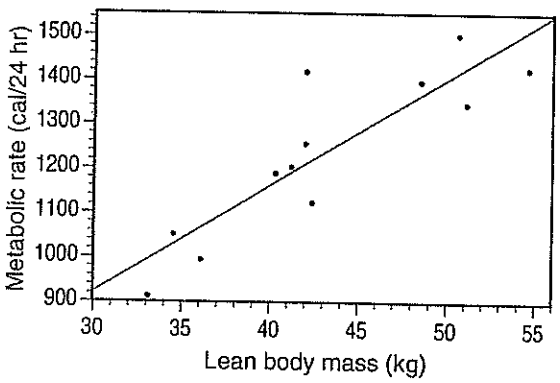
5.2. The equation is weight $= 80 - 6 \times$ days; the intercept is 80 g (the initial weight), and the slope is -6 grams/day.

5.3. Note that the means, standard deviations, and correlation were previously computed in the solution to Exercise 4.10. (a) The means and standard deviations are $\bar{x} = 3.5$ and $s_x \doteq 1.3784$ ranges, and $\bar{y} = 31.\overline{3}$ and $s_y \doteq 16.1328$ days. The correlation is $r \doteq 0.9623$. Therefore, the slope and intercept of the regression line are (respectively)

$$b = r\frac{s_y}{s_x} \doteq 11.26 \quad \text{and} \quad a = \bar{y} - b\bar{x} \doteq -8.088,$$

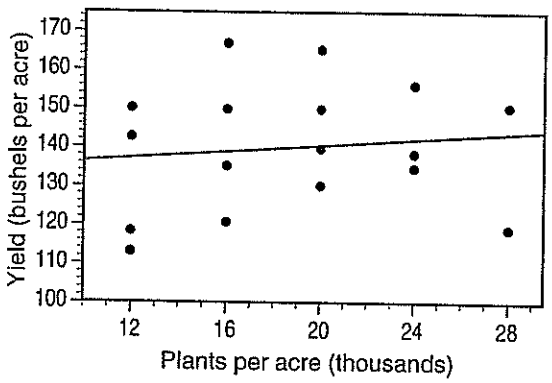
so the regression equation is $\hat{y} \doteq -8.088 + 11.26x$. (b) Obviously, the software result should be the same.

5.4. See also the solutions to Exercises 4.4 and 4.12. (a) The scatterplot is shown on the right. (b) The regression equation is $\hat{y} = 201.2 + 24.026x$. (c) The slope tells us that on the average, metabolic rate increases by about 24 cal/day for each additional kilogram of body mass. (d) For $x = 45$ kg, the predicted metabolic rate is $\hat{y} \doteq 1282.3$ cal/day.



5.5. A correlation close to 1 (or -1) means a strong linear relationship, so the points in the DMS/SRD scatterplot fall close to the regression line, so predictions based on the line are accurate. With a smaller correlation, the points are more widely spread around the line, so a prediction based on the line is less accurate.

5.6. **(a)** Scatterplot at right. Regression gives $\hat{y} = 132.45 + 0.402x$ (Minitab output below). The plot suggests a curved pattern, so a linear formula is not appropriate for making predictions. **(b)** $r^2 = 0.0182$. This confirms what we see in the graph: this line does a poor job of summarizing the relationship.



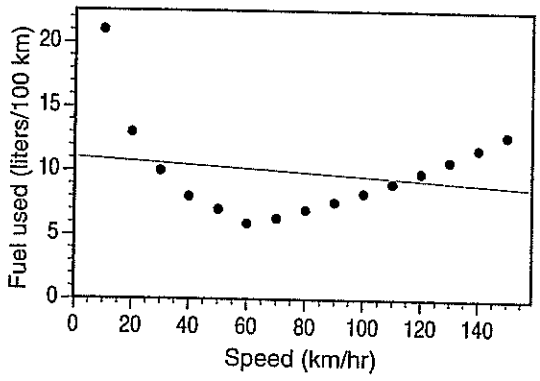
Minitab output

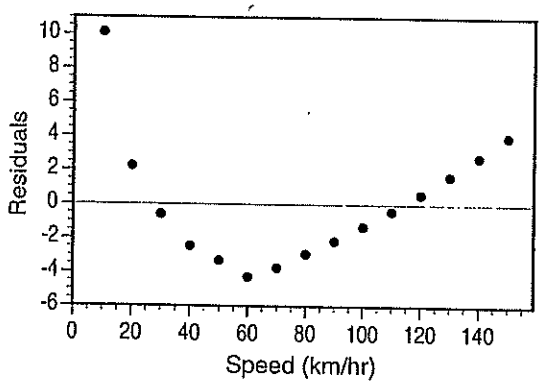
```
The regression equation is Yield = 132 + 0.402 Plants

Predictor       Coef      Stdev    t-ratio      p
Constant      132.45      14.91       8.89  0.000
Plants        0.4020     0.7625       0.53  0.606

s = 16.57      R-sq = 1.8%     R-sq(adj) = 0.0%
```

5.7. **(a)** Using the regression equation $\hat{y} \doteq -8.088 + 11.26x$, the predicted values and residuals are given in the table on the right. **(b)** Depending on the amount of rounding, the sum is either 0 or very close to 0. **(c)** The correlation between x and the residuals is no more than 0.0017 regardless of the amount of rounding.

Ranges (x)	Days (y)	$\hat{y}$	Residual $y - \hat{y}$
1	4	3.1754	0.8246
3	21	25.7018	−4.7018
4	33	36.9649	−3.9649
4	41	36.9649	4.0351
4	43	36.9649	6.0351
5	46	48.2281	−2.2281

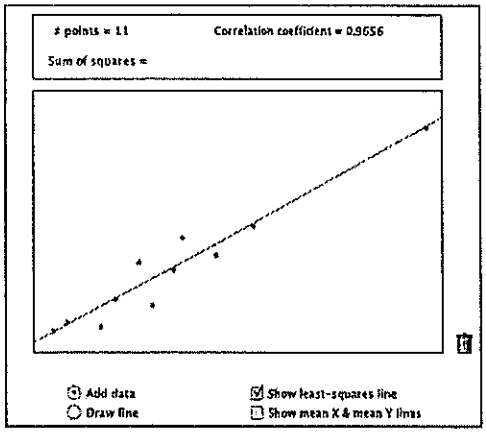
5.8. **(a)** Below, left. **(b)** No; the pattern is curved, so a linear formula is not the appropriate choice for prediction. **(c)** For $x = 10$, we estimate $\hat{y} = 11.058 - 0.01466(10) \doteq 10.91$, so the residual is $21.00 - 10.91 = 10.09$. The sum of the residuals is -0.01. **(d)** The first two and last four residuals are positive, and those in the middle are negative. Plot below, right.

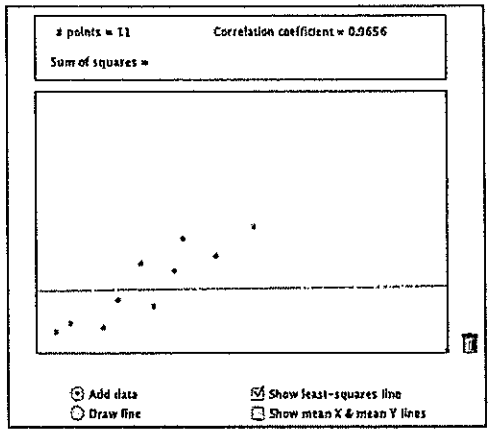




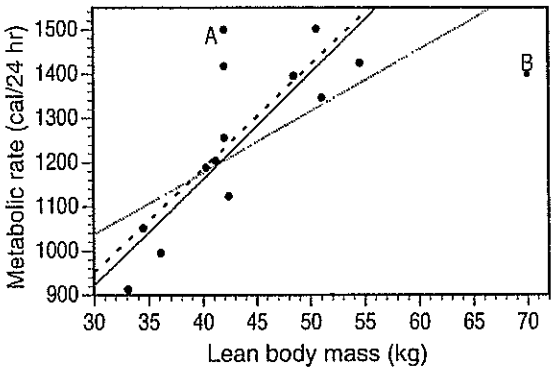
5.9. **(a)** Any point that falls exactly on the regression line will not increase the sum of squared vertical distances (which the regression line minimizes). Any other line—even if it passes through this new point—will necessarily have a higher total sum of squares. Thus the regression line does not change. Possible output is shown on the following page, left. The correlation changes (increases) because the new point reduces the *relative* scatter about the regression line. (That is, the distance of the points above and below the line remains

the same, but the spread of the x values increases.) **(b)** Influential points are those whose x coordinates are outliers; this point is on the right side, while all others are on the left. Possible output is shown below, right.



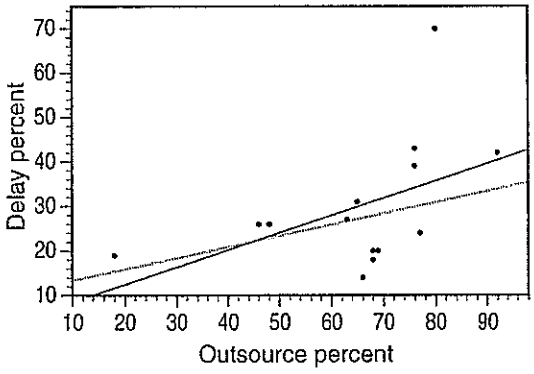


5.10. See also the solution to Exercise 5.4.



(a) Point A lies above the other points; that is, the metabolic rate is higher than we expect for the given body mass. Point B lies to the right of the other points; that is, it is an outlier in the x (mass) direction, and the metabolic rate is lower than we would expect. **(b)** In the plot, the solid line is the regression line for the original data. The dashed line slightly above that includes Point A; it has a very similar slope to the original line, but a slightly higher intercept, because Point A pulls the line up. The third line includes Point B, the more influential point; because Point B is an outlier in the x direction, it "pulls" the line down so that it is less steep.

5.11. See also the solution to Exercise 4.5.



(a) The scatterplot (with regression lines) is shown on the right. **(b)** The correlation is $r \doteq 0.4765$ with all points. It rises slightly to 0.4838 with the outlier removed; this is too small a change to consider the outlier influential for correlation. **(c)** With all points, $\hat{y} \doteq 4.73 + 0.3868x$ (the solid line), and the prediction for $x = 76$ is 34.13%. With Hawaiian Airlines removed, $\hat{y} \doteq 10.88 + 0.2495x$ (the dotted line), and the prediction is 29.84%. This difference in prediction—and the visible difference in the two lines—indicates that the outlier is influential for regression.

5.12. **(a)** The regression equation is $\hat{y} = -43.81 + 0.1302x$. **(b)** With $x = 1027$ thousand boats, we predict about 90 manatee deaths ($\hat{y} \doteq 89.87$). Assuming conditions in 2007 were similar to the previous 30 years, this is a fairly reliable prediction because of the strong linear association visible in the scatterplot. **(c)** With $x = 0$ boats, our prediction is the intercept: $\hat{y} \doteq -43.81$ manatee deaths. A negative number of deaths makes no sense, unless we are making a horror film called "Attack of the Zombie Manatees."

Note: *The fact that we trust our prediction in (b) does not guarantee that it is exactly right. In fact, the actual number of manatee deaths in 2007 was 73—quite a bit lower than our prediction (90 deaths). However, the point (1027, 73) fits reasonably well with the other points in the scatterplot; it just happens to be on the "edge" of the scatterplot, rather than in the center (next to the regression line).*

Minitab output

```
The regression equation is Kills = - 43.8 + 0.130 Boats

Predictor       Coef      Stdev    t-ratio       p
Constant     -43.812      5.717      -7.66   0.000
Boats       0.130164   0.007822      16.64   0.000

s = 7.445      R-sq = 90.8%     R-sq(adj) = 90.5%
```

5.13. A student's intelligence may be a lurking variable: stronger students (who are more likely to succeed when they get to college) are more likely to choose to take these math courses, while weaker students may avoid them. Other possible answers might be variations on this idea; for example, if we believe that success in college depends on a student's self-confidence, and perhaps confident students are more likely to choose math courses.

5.14. Possible lurking variables include the IQ and socioeconomic status of the mother, as well as the mother's other habits (drinking, diet, etc.). These variables are associated with smoking in various ways, and are also predictive of a child's IQ.

Note: *There may be an indirect cause-and-effect relationship at work here: some studies have found evidence that over time, smokers lose IQ points, perhaps due to brain damage caused by free radicals from the smoke. So perhaps smoking mothers gradually grow less smart, and are less able to nurture their children's cognitive development.*

5.15. Social status is a possible lurking variable: children from upper-class families can more easily afford higher education, and they would typically have had better preparation for college as well. They may also have some advantages when seeking employment, and have more money should they want to start their own businesses.

This could be compounded by racial distinctions: some minority groups receive worse educations than other groups, and prejudicial hiring practices may keep minorities out of higher-paying positions.

It could also be that some causation goes the other way: people who are doing well in their jobs might be encouraged to pursue further education.

5.16. Age is probably the most important lurking variable: married men would generally be older than single men, so they would have been in the workforce longer, and therefore had more time to advance in their careers.

5.17. (b) The line passes through (or near) the point (110, 60).

5.18. (c) The line is clearly positively sloped.

5.19. (c) The slope is the coefficient of x.

5.20. (a) The slope is \$100/yr, and the intercept is \$500 (his beginning balance).

5.21. (b) Age at death and packs per day are negatively associated. In other words, the more one smokes, the shorter one's life.

5.22. (a) This is what the slope of the regression line tells us.

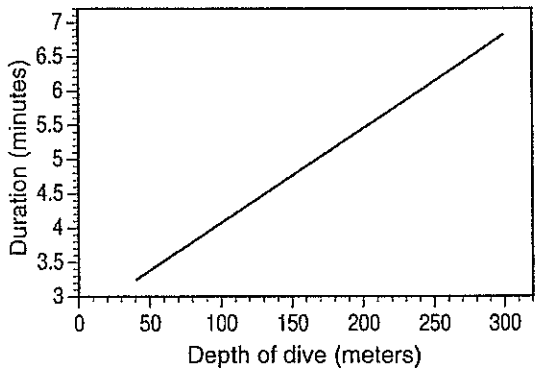
5.23. (b) $\hat{y} = 6.4 + 0.93(100) = 6.4 + 93 = 99.4$ cm.

5.24. (a) The slope and the correlation always have the same sign.

5.25. (c) The regression line explains 95% of the variation in height.

5.26. (b) One can also guess this by considering the slope between the first two points: y changes by about -40 when x changes by about -10. The only slope that is even close to that is 2.4. Alternatively, note that when $x = 50$ cm, the data suggests that y should be about 160 cm, and only the second equation gives a result close to that.

5.27. **(a)** The slope is 0.0138 minutes per meter. On the average, if the depth of the dive is increased by one meter, it adds 0.0138 minutes (about 0.83 seconds) to the time spent underwater. **(b)** When D = 200, the regression formula estimates DD to be 5.45 minutes. **(c)** To plot the line, compute DD = 3.242 minutes when D = 40 meters, and DD = 6.83 minutes when D = 300 meters.



5.28. **(a)** The slope (1.507) says that, on the average, BOD rises (falls) by 1.507 mg/l for every 1 mg/l increase (decrease) in TOC. **(b)** When TOC = 0 mg/l, the predicted BOD level is -55.43 mg/l. This must arise from extrapolation; the data used to find this regression formula must not have included values of TOC near 0.

5.29. See also the solution to Exercise 4.45. **(a)** The regression equation is $\hat{y} = -0.126 + 0.0608x$. For $x = 2.0$, this formula gives $\hat{y} = -0.0044$. (A student who uses the numbers listed under "Coef" in the Minitab output might report the predicted brain activity as -0.0045.) **(b)** This is given in the Minitab output as "R-sq": 77.1%. The linear relationship explains 77.1% of the variation in brain activity. **(c)** Knowing that $r^2 \doteq 0.771$,

we find $r = \sqrt{r^2} \doteq 0.88$; the sign is positive because it has the same sign as the slope coefficient.

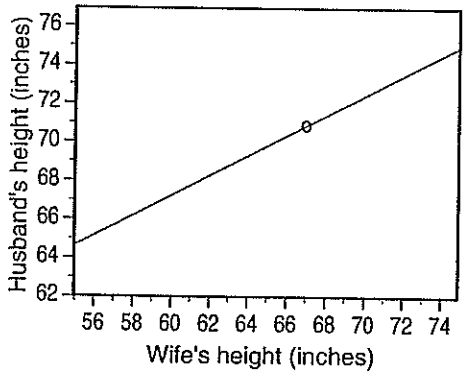
5.30. See also the solution to Exercise 4.44. **(a)** The regression line is $\hat{y} = 158 - 2.99x$. Following a season with 30 breeding pairs, we find $\hat{y} \doteq 68.3\%$, so we predict that about 68% of males will return. (A student who uses the numbers listed under "Coef" in the Minitab output might report the prediction as $\hat{y} = 67.875\%$.) **(b)** This is given in the Minitab output as "R-sq": 63.1%. The linear relationship explains 63.1% of the variation in the percent of returning males. **(c)** Knowing that $r^2 \doteq 0.631$, we find $r = -\sqrt{r^2} \doteq -0.79$; the sign is negative because it has the same sign as the slope coefficient.

5.31. Women's heights are the x values; men's are the y values. **(a)** The slope and intercept are

$$b = r \cdot s_y/s_x = 0.5 \cdot 2.8/2.7 \doteq 0.5185$$

$$a = \bar{y} - b\bar{x} = 69.3 - (0.5185)(64) \doteq 36.115.$$

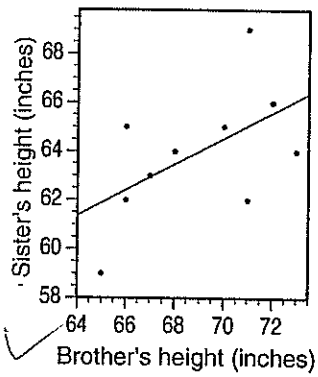
(b) The regression equation is $\hat{y} = 36.115 + 0.5185x$. Ideally, the scales should be the same on both axes. For a 67-inch-tall wife, we predict the husband's height will be about 70.85 inches. **(c)** The regression line only explains $r^2 = 25\%$ of the variation in the height of the husband.



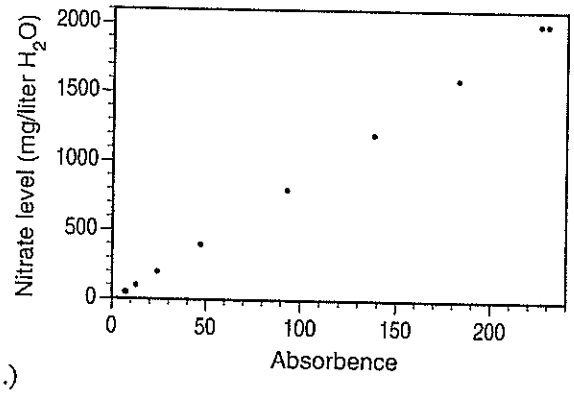
5.32. **(a)** The slope is $b = r \cdot s_y/s_x = (0.6)(8)/(30) = 0.16$, and the intercept is $a = \bar{y} - b\bar{x} = 30.2$. **(b)** Julie's predicted score is $\hat{y} = 78.2$. **(c)** $r^2 = 0.36$; only 36% of the variability in y is accounted for by the regression, so the estimate $\hat{y} = 78.2$ could be quite different from the real score.

5.33. $r = \sqrt{0.16} = 0.40$ (high attendance goes with high grades, so the correlation must be positive).

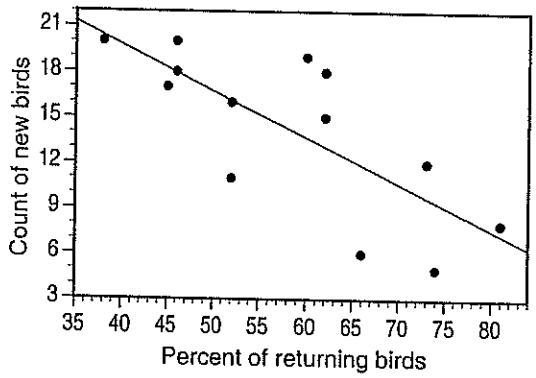
5.34. **(a)** The correlation is $r \doteq 0.558$, and the regression equation is $\hat{y} = 27.64 + 0.527x$. **(b)** When $x = 70$ inches, we predict Tonya's height to be $\hat{y} = 64.5$ inches. Because of the relatively low correlation ($r^2 \doteq 0.311$) and the variation about the line in the scatterplot, we should not place too much confidence in this prediction.



5.35. **(a)** Plot at right; based on the discussion in part (b), absorbence is the explanatory variable, so it has been placed on the horizontal axis. The correlation is $r \doteq 0.9999$, so recalibration is not necessary. **(b)** The regression line is $\hat{y} = 8.825x - 14.52$; when $x = 40$, we predict $\hat{y} \doteq 338.5$ mg/l. **(c)** This prediction should be very accurate because the relationship is so strong. (It explains $r^2 = \doteq 99.99\%$ of the variation in nitrate level.)



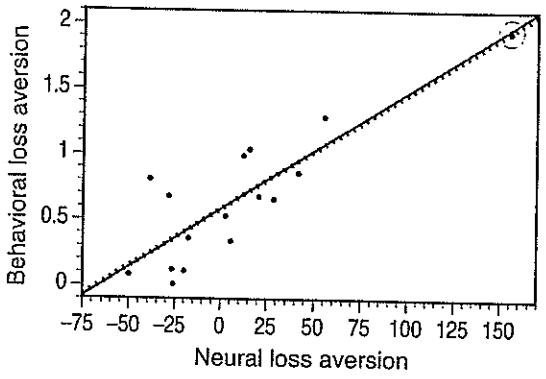
5.36. See also the solution to Exercise 4.28. **(a)** The regression equation is $\hat{y} = 31.9 - 0.304x$. **(b)** The slope (-0.304) tells us that, on the average, for every 1% increase in returning birds, the number of new birds joining the colony decreases by 0.304. **(c)** When $x = 60$, we predict $\hat{y} \doteq 13.69$ new birds will join the colony.



Minitab output

```
The regression equation is New = 31.9 - 0.304 PctRtn

Predictor       Coef      Stdev    t-ratio       p
Constant      31.934      4.838       6.60   0.000
PctRtn      -0.30402    0.08122      -3.74   0.003

s = 3.667       R-sq = 56.0%     R-sq(adj) = 52.0%
```

5.37. See also the solution to Exercise 4.29. **(a)** The outlier (in the upper right corner) is circled, because it is hard to see behind the two regression lines. **(b)** With the outlier omitted, the regression line is $\hat{y} = 0.586 + 0.00891x$. (This is the solid line in the plot.) **(c)** The line does not change much because the outlier fits the pattern of the other points; r changes because the scatter (relative to the length of the line) is greater with the outlier removed. **(d)** The correlation changes from 0.8486 (with all points) to 0.7015 (without the outlier). With all points included, the regression line is $\hat{y} = 0.585 + 0.00879x$ (the dotted line in the plot—nearly indistinguishable from the other regression line).



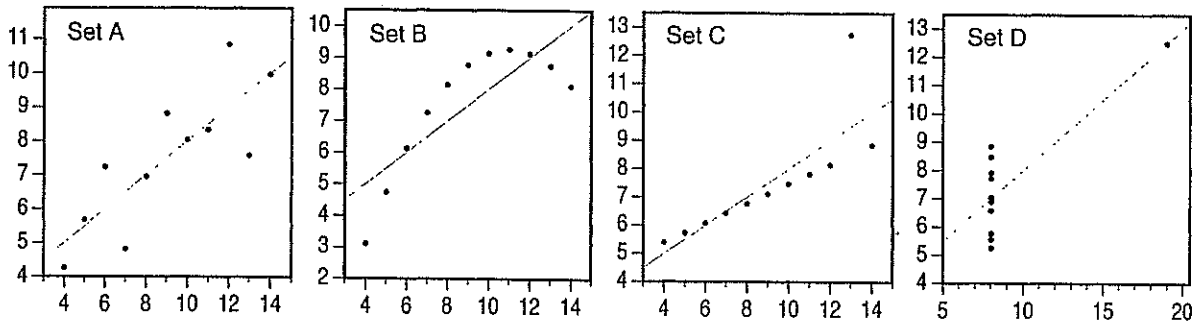
Minitab output: All points

```
The regression equation is Behave = 0.585 + 0.00879 Neural

Predictor       Coef      Stdev    t-ratio       p
Constant     0.58496    0.07093       8.25   0.000
Neural      0.008794   0.001465       6.00   0.000
```

With outlier removed

```
The regression equation is Behave = 0.586 + 0.00891 Neural

Predictor       Coef      Stdev    t-ratio       p
Constant     0.58581    0.07506       7.80   0.000
Neural      0.008909   0.002510       3.55   0.004
```

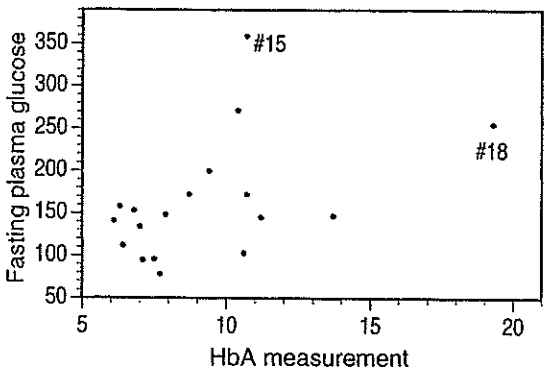
5.38. **(a)** To three decimal places, the correlations are all approximately 0.816 (for Set D, r actually rounds to 0.817), and the regression lines are all approximately $\hat{y} = 3.000 + 0.500x$. For all four sets, we predict $\hat{y} \doteq 8$ when $x = 10$. **(b)** Plots below. **(c)** For Set A, the use of the regression line seems to be reasonable—the data seem to have a moderate linear association (albeit with a fair amount of scatter). For Set B, there is an obvious *non*linear relationship; we should fit a parabola or other curve. For Set C, the point (13, 12.74) deviates from the (highly linear) pattern of the other points; if we can exclude it, the (new) regression formula would be very useful for prediction. For Set D, the data point with $x = 19$ is a very influential point—the other points alone give no indication of slope for the line. Seeing how widely scattered the y-coordinates of the other points are, we cannot place too much faith in the y-coordinate of the influential point; thus we cannot depend on the slope of the line, and so we cannot depend on the estimate when $x = 10$. (We also have no evidence as to whether or not a line is an appropriate model for this relationship.)



5.39. **(a)** The two unusual observations are marked on the scatterplot. **(b)** The correlations are

$r_1 \doteq 0.4819$ (all observations)
$r_2 \doteq 0.5684$ (without Subject 15)
$r_3 \doteq 0.3837$ (without Subject 18)

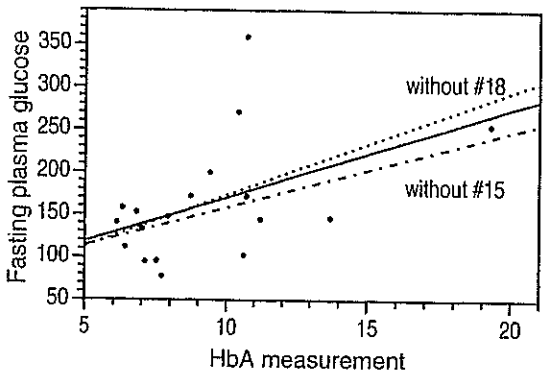
Both outliers change the correlation. Removing Subject 15 increases r, because its presence makes the scatterplot less linear, while removing Subject 18 decreases r, because its presence decreases the relative scatter about the linear pattern.



5.40. (a) The regression equation is $\hat{y} = 44.13 + 2.4254x$. (b) With the altered data, the equation is $\hat{y} = 0.4413 + 0.0024254x$. (c) With $x = 50$ cm, the first equation predicts $\hat{y} \doteq 165.4$ cm. With $x = 500$ mm, the second equation predicts $\hat{y} \doteq 1.654$ m.

5.41. The scatterplot from Exercise 5.39 is reproduced here with the regression lines added. The equations are

$\hat{y} \doteq 66.4 + 10.4x$ (all observations)
$\hat{y} \doteq 69.5 + 8.92x$ (without #15)
$\hat{y} \doteq 52.3 + 12.1x$ (without #18)



While the equation changes in response to removing either subject, one could argue that neither one is particularly influential, because the line moves very little over the range of x (HbA) values. Subject 15 is an outlier in terms of its y value; such points are typically not influential. Subject 18 is an outlier in terms of its x value, but is not particularly influential because it is consistent with the linear pattern suggested by the other points.

5.42. In this case, there may be a causative effect, but in the direction opposite to the one suggested: People who are overweight are more likely to be on diets, and so choose artificial sweeteners over sugar. (Also, heavier people are at a higher risk to develop Type 2 diabetes; if they do, they are likely to switch to artificial sweeteners.)

5.43. Responses will vary. For example, students who choose the online course might have more self-motivation, or have better computer skills (which might be helpful in doing well in the class; e.g., such students might do better at researching course topics on the Internet).

5.44. For example, a student who in the past might have received a grade of B (and a lower SAT score) now receives an A (but has a lower SAT score than an A student in the past). While this is a bit of an oversimplification, this means that today's A students are yesterday's A and B students, today's B students are yesterday's C students, and so on. Because of the grade inflation, we are not comparing students with equal abilities in the past and today.

5.45. Here is a (relatively) simple example to show how this can happen: suppose that most workers are currently 30 to 50 years old; of course, some are older or younger than that, but this age group dominates. Suppose further that each worker's current salary is his/her age (in thousands of dollars); for example, a 30-year-old worker is currently making $30,000.

Over the next 10 years, all workers age, and their salaries increase. Suppose every worker's salary increases by between $4000 and $8000. Then every worker will be making *more* money than he/she did 10 years before, but *less* money than a worker of that same age 10 years before.

During that time, a few workers will retire, and others will enter the workforce, but that large cluster that had been between the ages of 30 and 50 (now between 40 and 60) will bring up the overall median salary despite the changes in older and younger workers.

5.46. We have slope $b = r\,s_y/s_x$ and intercept $a = \bar{y} - b\bar{x}$, and $\hat{y} = a + bx$, so when $x = \bar{x}$,

$$\hat{y} = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}.$$

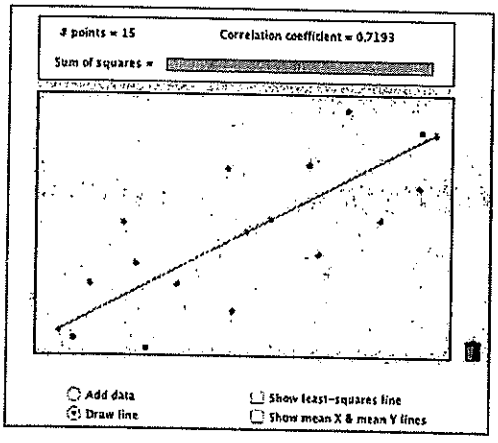
(Note that the value of the slope does not actually matter.)

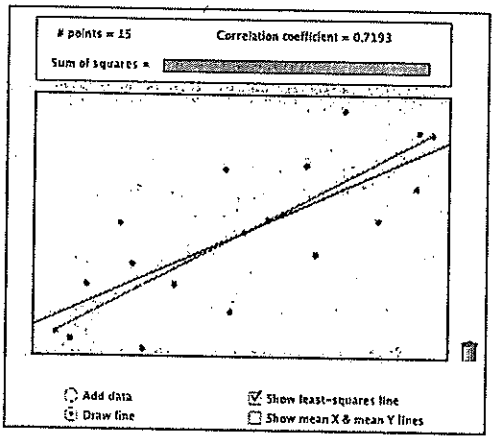
5.47. With the regression equation, $\hat{y} = 61.93 + 0.180x$, a first-round score of $x = 80$ leads to a predicted second-round score of $\hat{y} \doteq 76.33$, while a first-round score of $x = 70$ leads to a predicted second-round score of $\hat{y} \doteq 74.53$. As the text notes, an above-average first-round score predicts a slightly-less-above-average score in the second round—and likewise for below-average scores.

5.48. Note that $\bar{y} = 46.6 + 0.41\bar{x}$. We predict that Octavio will score 4.1 points above the mean on the final exam: $\hat{y} = 46.6 + 0.41(\bar{x} + 10) = 46.6 + 0.41\bar{x} + 4.1 = \bar{y} + 4.1$. (Alternatively, because the slope is 0.41, we can observe that an increase of 10 points on the midterm yields an increase of 4.1 on the predicted final exam score.)

5.49. See the solution to Exercise 4.41 for three sample scatterplots. A regression line is appropriate only for the scatterplot of part (b). For the graph in (c), the point not in the vertical stack is very influential—the stacked points alone give no indication of slope for the line (if indeed a line is an appropriate model). If the stacked points are scattered, we cannot place too much faith in the y-coordinate of the influential point; thus we cannot depend on the slope of the line, and so we cannot depend on predictions made with the regression line. The curved relationship exhibited by the scatterplot in (d) clearly indicates that predictions based on a straight line are not appropriate.

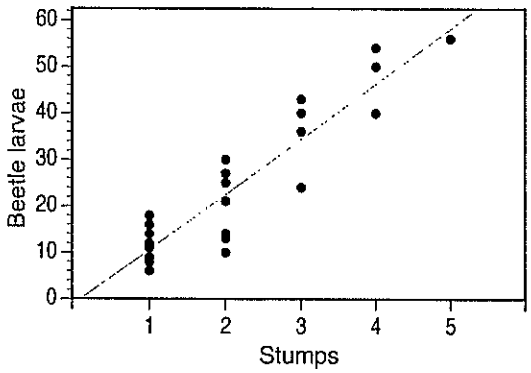
5.50. **(a)** Drawing the "best line" by eye is a very inaccurate process; few people choose the best line (although you can get better at it with practice). **(b)** Most people tend to overestimate the slope for a scatterplot with $r \doteq 0.7$; that is, most students will find that the least-squares line is less steep than the one they draw.





5.51. PLAN: We construct a scatterplot (with beaver stumps as the explanatory variable), and if appropriate, find the regression line and correlation.

SOLVE: The scatterplot shows a positive linear association. Regression seems to be an appropriate way to summarize the relationship; the regression line is $\hat{y} = -1.286 + 11.89x$. The straight-line relationship explains $r^2 \doteq 83.9\%$ of the variation in beetle larvae.



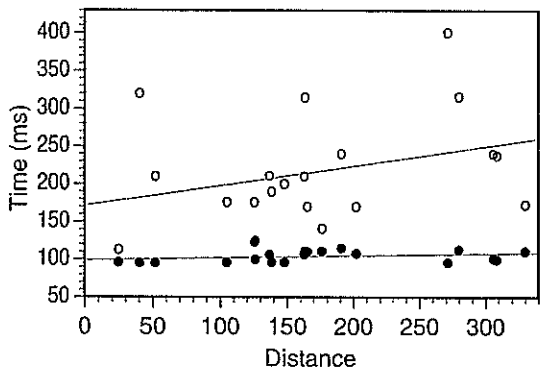
CONCLUDE: The strong positive association supports the idea that beavers benefit beetles.

Minitab output

```
The regression equation is larvae = - 1.29 + 11.9 stumps

Predictor       Coef      Stdev    t-ratio        p
Constant      -1.286      2.853      -0.45    0.657
stumps        11.894      1.136      10.47    0.000

s = 6.419       R-sq = 83.9%     R-sq(adj) = 83.1%
```

5.52. PLAN: We construct a scatterplot, with distance as the explanatory variable, using different symbols for the left and right hands, and (if appropriate) find separate regression lines for each hand.

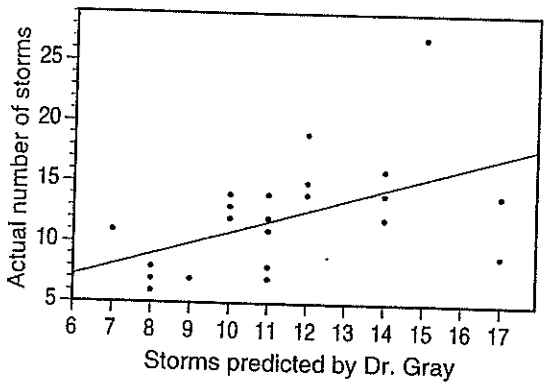
SOLVE: In the scatterplot, right-hand points are filled circles and left-hand points are open circles. In general, the right-hand points lie below the left-hand points, meaning the right-hand times are shorter, so the subject is right-handed. There is no striking pattern for the left-hand points; the pattern for right-hand points is obscured because they are squeezed at the bottom of the plot. While neither plot looks particularly linear, we might nonetheless find the two regression lines: for the right hand, $\hat{y} = 99.4 + 0.0283x$ ($r = 0.305$, $r^2 = 9.3\%$), and for the left hand, $\hat{y} = 172 + 0.262x$ ($r = 0.318$, $r^2 = 10.1\%$).



CONCLUDE: Neither regression is particularly useful for prediction; distance accounts for only 9.3% (right) and 10.1% (left) of the variation in time.

5.53. PLAN: We construct a scatterplot with Dr. Gray's forecast as the explanatory variable, and if appropriate, find the regression equation.

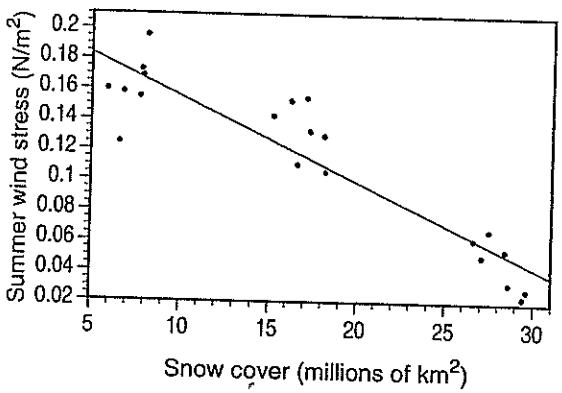
SOLVE: The scatterplot shows a moderate positive association; the regression line is $\hat{y} = 1.803 + 0.9031x$, with $r^2 \doteq 28\%$. The relationship is strengthened by the large number of storms in the 2005 season, but it is weakened by the last two years of data, when Gray's forecasts were the highest, but the actual numbers of storms were unremarkable. As an indication of the influence of the 2005 season, we might find the regression line without that point; it is $\hat{y} = 4.421 + 0.6224x$, with $r^2 \doteq 22.6\%$.



CONCLUDE: If Dr. Gray forecasts $x = 16$ tropical storms, we expect 16.25 storms in that year. However, we do not have very much confidence in this estimate, because the regression line explains only 28% of the variation in tropical storms. (If we exclude 2005, the prediction is 14.4 storms, but this estimate is less reliable than the first.)

5.54. PLAN: We examine a scatterplot of wind stress against snow cover—viewing the latter as explanatory—and (if appropriate) compute correlation and regression lines.

SOLVE: The scatterplot suggests a negative linear association, with correlation $r \doteq -0.9179$. The regression line is $\hat{y} = 0.212 - 0.00561x$; the linear relationship explains $r^2 \doteq 84.3\%$ of the variation in wind stress.



CONCLUDE: We have good evidence that decreasing snow cover is strongly associated with increasing wind stress.

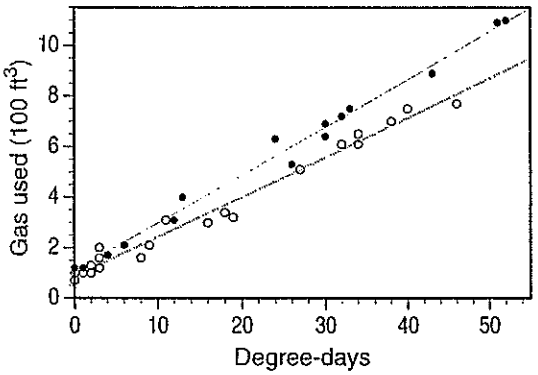
Minitab output

```
The regression equation is wind = 0.212 - 0.00561 snow

Predictor        Coef        Stdev    t-ratio        p
Constant      0.21172      0.01083      19.56    0.000
snow       -0.0056096    0.0005562     -10.09    0.000

s = 0.02191     R-sq = 84.3%     R-sq(adj) = 83.4%
```

5.55. See also the solution to Exercise 4.43.

PLAN: We construct a scatterplot of gas use against outside temperature (the explanatory variable), using separate symbols for before and after solar panels were installed. We also find before and after regression lines, and estimate gas usage when $x = 45$.

SOLVE: Both sets of points show a strong positive linear association between degree-days and gas usage. The new points (open circles) are generally slightly lower than the pre-solar-panel points. The regression lines are $\hat{y} = 1.089 + 0.1890x$ (before) and $\hat{y} = 0.8532 + 0.1569x$ (after). Both lines give very reliable predictions ($r^2 \doteq 99.1\%$ and $r^2 \doteq 98.2\%$, respectively).



CONCLUDE: With $x = 45$, the predictions (before and after, respectively) are 9.59 and 7.91 hundred cubic feet. This gives an estimated savings of about 168 cubic feet.

5.56. PLAN: We construct scatterplots of female life expectancy and infant mortality against health care spending (the explanatory variable), and compute regression lines if appropriate.

SOLVE: The two scatterplots (below) show a positive association between spending and life expectancy, and a negative association with infant mortality; these associations are what we might expect. In both cases, the United States and South Africa stand out as outliers.

One could choose from many possible regression lines. The scatterplots show only the lines based on all points, but here is a more complete list of possibilities:

	Life expectancy		Infant mortality	
	Regression line	r^2	Regression line	r^2
All points	$\hat{y} = 74.73 + 0.001971x$	30.4%	$\hat{y} = 12.22 - 0.002613x$	12.0%
Without U.S.	$\hat{y} = 73.43 + 0.002753x$	41.9%	$\hat{y} = 14.03 - 0.003700x$	17.0%
Without S.A.	$\hat{y} = 76.14 + 0.001494x$	40.4%	$\hat{y} = 8.398 - 0.001319x$	17.1%
Without both	$\hat{y} = 75.01 + 0.002154x$	58.8%	$\hat{y} = 9.614 - 0.002033x$	28.5%

For both life expectancy and infant mortality, the best predictions come from the lines which exclude both outliers—but for infant mortality, even those predictions are not very good.

CONCLUDE: Health care spending allows some prediction of infant mortality and life expectancy, but those predictions are not too reliable unless the outliers are excluded.