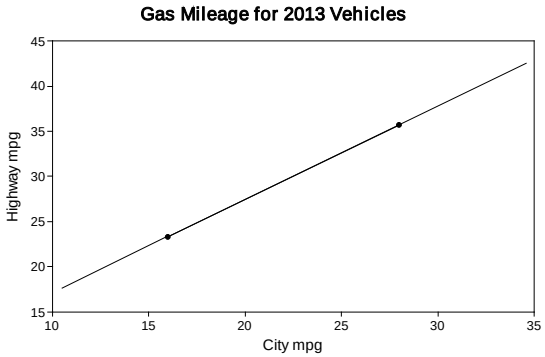
Chapter 5 – Regression

5.1 (a) The slope is 1.033. On average, highway mileage increases by 1.033 mpg for each additional 1 mpg change in city mileage. **(b)** The intercept is 6.785 mpg. This is the highway mileage for a nonexistent car that gets 0 mpg in the city. Although this interpretation is valid, such a prediction would be invalid, since 0 is outside the range of the data (this is extrapolation, which will be addressed later in the chapter). **(c)** For a car that gets 16 mpg in the city, we predict highway mileage to be 6.785 + (1.033)(16) = 23.31 mpg. For a car that gets 28 mpg in the city, we predict highway mileage to be 6.785 + (1.033)(28) = 35.71 mpg. **(d)** The regression line passes through all the points of prediction. The plot was created by drawing a line through the two points (16, 23.31) and (28, 35.71), corresponding to the city mileages and predicted highway mileages for the two cars described in (c).



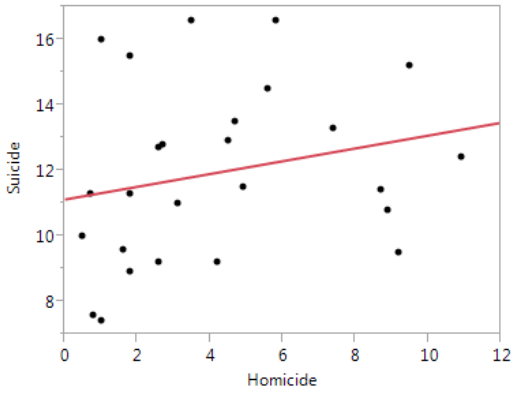
5.2 The equation is cigarettes = 48,000,000 – 0.178(new runners).

5.3 (a) The slope is 1021. This means that for each year since 2000, forest loss averages about 1021 km^2. **(b)** If we measured in square meters, the slope would be 1021×10^6 = 1,021,000,000; a loss of 1 billion square meters per year (on average). In thousands of km^2, the slope would be 1.021; a loss of a bit more than 1000 km^2 per year (on average).

Note: *The point of this exercise is that units matter a great deal in regression. All these slopes represent the same relationship.*

5.4 (a) $\bar{x}$ = 30.280, s_x = 0.4296, $\bar{y}$ = 2.4557, s_y = 0.1579, and r = –0.8914. $b = r\frac{s_y}{s_x}$ = $(-0.8914)\frac{0.1579}{0.4296}$ = –0.3276, and $a = \bar{y} - b\bar{x}$ = 2.4557 – (–0.3276)(30.280) = 12.3754. The equation is $\widehat{Coral\ growth} = 12.3754 - 0.3276$(Celsius Temperature). **(b)** Software agrees with these values to three decimal places, since we rounded to the fourth decimal place (where values are rounded will affect these results). **(c)** The slope is –0.3276. This means that every increase of one degree Celsius means about 0.3276 fewer mean millimeters of coral growth per year.

5.5 (a) The scatterplot (with the regression line) is shown. This relationship is certainly weak. **(b)** JMP output is shown. The regression equation is $\widehat{\text{Suicide}}$ = 11.125 + 0.195(Homicide). **(c)** The slope means that for every suicide (per 100,000 people), there are about 0.195 homicides (per 100,000 people) in these Ohio counties. **(d)** We would predict 11.125 + 0.195(8.0) = 12.685 suicides.



Linear Fit

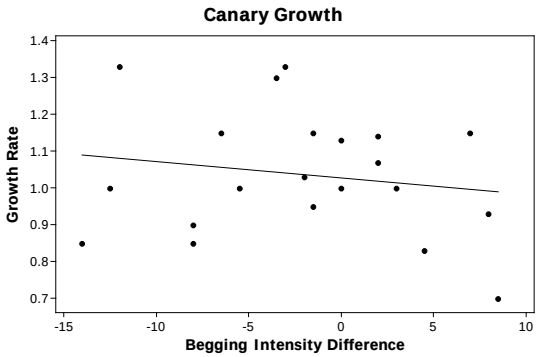
Suicide = 11.125 + 0.1953551*Homicide

Summary of Fit

RSquare	0.052692
RSquare Adj	0.013221
Root Mean Square Error	2.657216
Mean of Response	11.95
Observations (or Sum Wgts)	26

5.6 The farther r is from 0 (in either direction), the stronger the linear relationship is between two variables. In Exercise 4.30, the relationship between SRD and DMS is very strongly linear, and a regression line should enable relatively more accurate prediction than a regression line for the golfers' scores.

5.7 (a) The scatterplot is provided, with the regression line. Regression gives $\hat{y} = 1.0284 - 0.004498x$ (see Minitab output). The plot suggests a slightly curved pattern, not a strong linear pattern. A regression line is not useful for making predictions. **(b)** $r^2 = 0.031$. This confirms what we see in the graph: the regression line does a poor job summarizing the relationship between difference in begging intensity and growth rate. Only about 3% of the variation in growth rate is explained by the least-squares regression on difference in begging intensity.



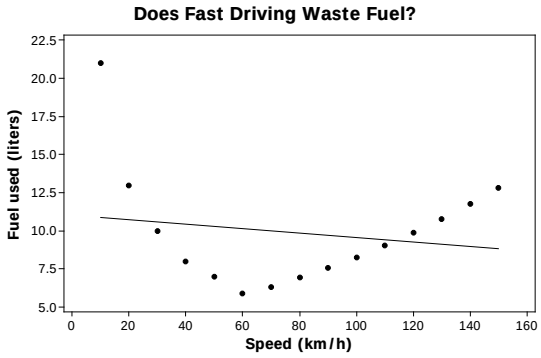
Minitab output

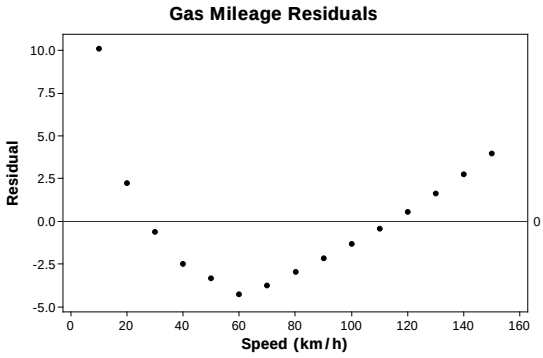
```
The regression equation is Growth = 1.028 – 0.0045 Difference

Predictor       Coef       Stdev     t-ratio    P
Constant     1.028409   0.039042     26.341  0.000
Difference  –0.004498   0.005808     –0.774  0.448

s = 0.1704  R-Sq = 3.1%
```

5.8 (a) The residuals are computed in the table using $\hat{y} = 12.3754 - 0.3276x$, as computed in Exercise 5.4. **(b)** They sum to zero, except for rounding error. **(c)** From software, the correlation between x and $y - \hat{y}$ is 0.000025, which is zero except for rounding.

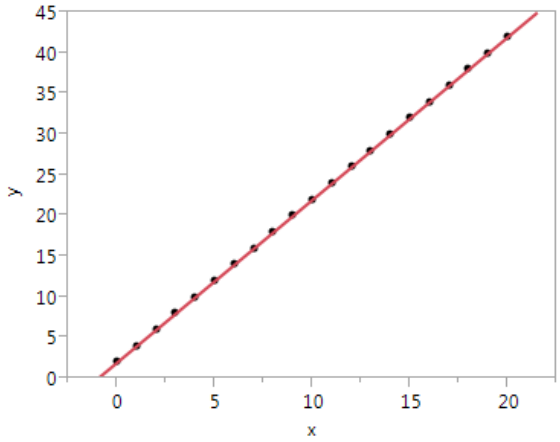
x	y	$\hat{y}$	$y - \hat{y}$
29.68	2.63	2.652	–0.022
29.87	2.58	2.590	–0.010
30.16	2.60	2.495	0.105
30.22	2.48	2.475	0.005
30.48	2.26	2.390	–0.130
30.65	2.38	2.335	0.045
30.90	2.26	2.253	0.007
		0	

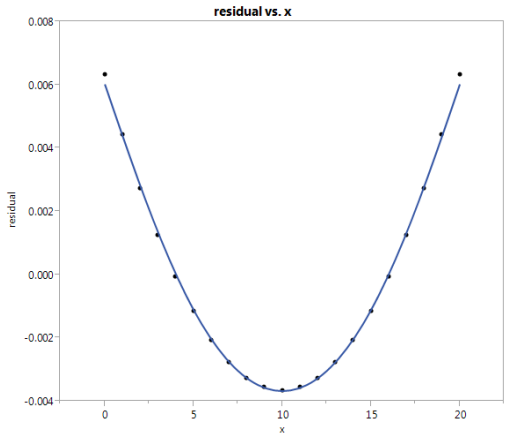
5.9 (a) Plot is provided following, left. **(b)** No; the pattern is curved, so linear regression is not appropriate for prediction. **(c)** For $x = 10$, we estimate $\hat{y} = 11.058 - 0.01466(10) = 10.91$, so the residual is 21.00 – 10.91 = 10.09. The sum of the residuals is –0.01. **(d)** The first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



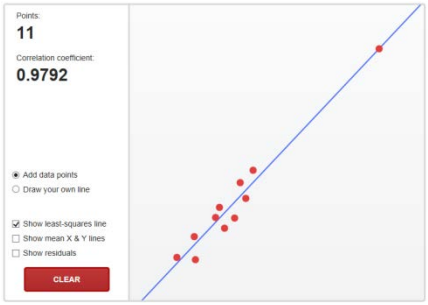


5.10 (a) The scatterplot with the line is shown below. **(b)** This looks like an excellent fit; all the data points are very close to the line, so predictions should be accurate. **(c)** The residuals plot is shown. This plot is clearly a curve that could not be seen in the original data. A linear model is not correct here.

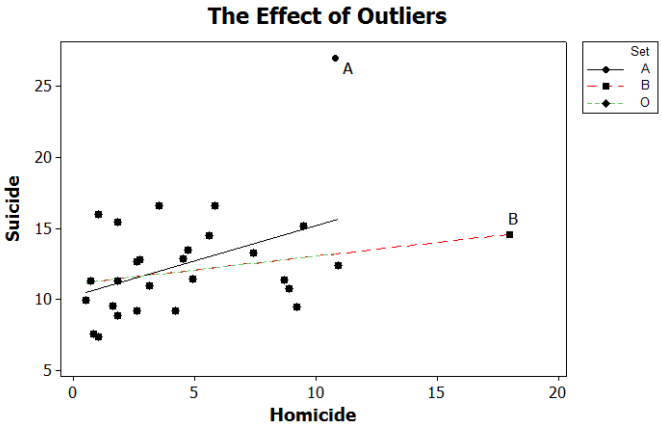




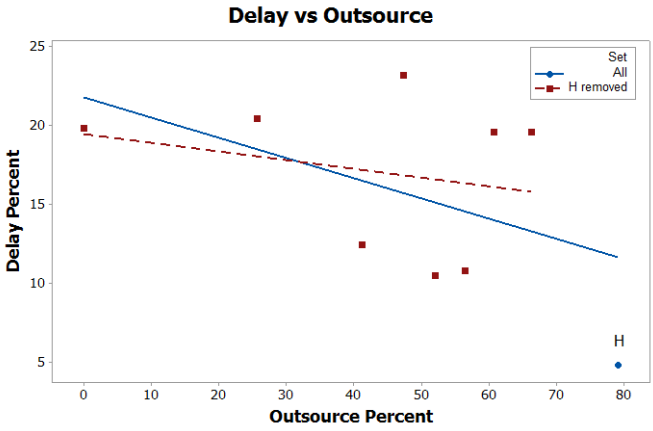
5.11 (a) Any point that falls exactly on the regression line will not increase the sum of squared vertical distances (which the regression line minimizes). Thus the regression line does not change. Possible output is shown, below left. Any other line (even if it passes through this new point) will necessarily have a higher total sum of squared prediction errors. The correlation changes (increases) because the new point reduces the relative scatter about the regression line. **(b)** Influential points are those whose *x*-coordinates are outliers. The regression line will "follow" an influential point if it is moved up or down in the *y* direction. An example is provided, below right.



5.12 (a) Point A lies above the other points; that is, the suicide rate is higher than we expect for the given homicide rate and is also an outlier in *x*. Point B lies to the right of the other points; it is an outlier in the *x* (homicide) direction, with a suicide rate more like the others. **(b)** In the plot, the regression lines for the original data and for the set including Point B are essentially identical. The solid line above that includes Point A; Point A is more influential than Point B.



5.13 (a) In the plot, Hawaiian Airlines is the point identified with "H." Since this point is an outlier and falls outside the *x* range of the other data points, it is influential, and will affect the regression line by "pulling" it. **(b)** With the outlier, $r = -0.488$. If the outlier is deleted from the data, $r = -0.241$. Notice that with the outlier, the correlation suggests a stronger linear relationship. **(c)** The two regression lines (one including the outlier, and the other without) are plotted. We see that the line based on the full data set (including the outlier) has been pulled down toward the outlier, indicating that the outlier is influential. Now, the regression line based on the complete (original) data set, including the outlier, is $\hat{y} = 21.815 - 0.12851x$. Using this, when $x = 79.1$, we predict 11.65% delays. The other regression line (fit without the outlier), is $\hat{y} = 19.460 - 0.05528x$, so our prediction would be 15.09% delays. The outlier impacts predictions because it impacts the regression line.



5.14 The correlation between *mean* SAT scores is an ecological correlation. There is far more variability among individuals than among the averages. Correlation would be much smaller (i.e., closer to zero) if we calculated it based on scores for individual students.

5.15 (a) The regression line is $\hat{y} = -44.831 + 0.1323\,x$ (or, $\widehat{\text{Kills}} = -44.831 + 0.1323$ Boats). **(b)** If 890,000 boats are registered, then by our scale, $x = 890$, and $\hat{y} = -44.831 + (0.1323)(890) = 72.92$ manatees killed. The prediction seems reasonable, as long as conditions remain the same, because "890" is within the space of observed values of *x* on which the regression line was based. That is, this is not extrapolation. **(c)** If $x = 0$ (corresponding to no registered boats), then we would "predict" –44.831 manatees to be killed by boats. This is absurd, because it is clearly impossible for fewer than 0 manatees to be killed. This illustrates the folly of extrapolation… $x = 0$ is well outside the range of observed values of *x* on which the regression line was based.

5.16 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.

5.17 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 toxins from the smoke. So, perhaps smoking mothers gradually grow less smart and are less able to nurture their children's cognitive development.

5.18 Socioeconomic 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 or their employers might pay for them to get further education.

5.19 One example would be that men who are married, widowed, or divorced may be more "invested" in their careers than men who are single. There is still a feeling of societal pressure for a man to "provide" for his family.

5.20 (b) 7.5. The regression line seems to pass through the point (110, 7.5).

5.21 (b) 0.2. Consider two points on the regression line—say (90,4) and (130,11). The slope of the line segment connecting these points is $\frac{11-4}{130-90} = 7/40 = 0.175$.

5.22 (c) −3

5.23 (a) $y = 1000 + 100x$

5.24 (b) will be less than 0. As the number of packs increases, average age at death decreases. Correlation is negative, and so is the slope of the regression line.

5.25 (c) 16 cubic feet

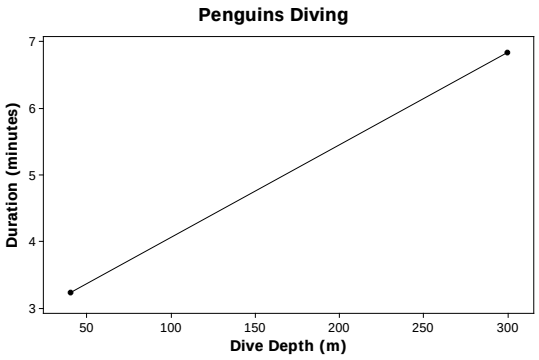
5.26 (a) 405 cubic feet

5.27 (a) The slope of the line is positive.

5.28 (c) prediction of gas used from degree-days will be quite accurate.

5.27 (a) $\hat{y} = 24.2 + 6.0x$

5.30 (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.

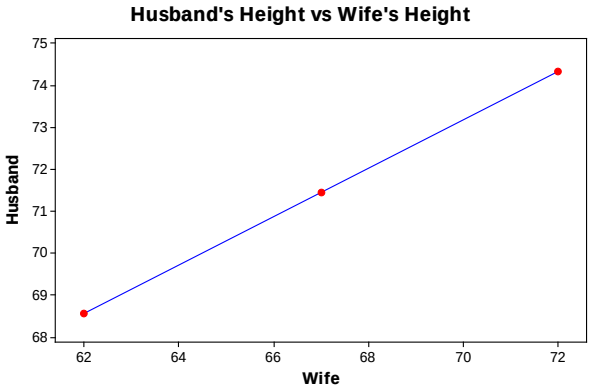


5.31 (a) Since the slope is 3721.02, the least-squares regression line says that increasing the size of a diamond by 1 carat increases its price by 3721.02 Singapore dollars, on average. **(b)** A diamond of size 0 carats would have a predicted price of 259.63 Singapore dollars. This is probably an extrapolation, since the data set on which the line was constructed almost certainly had no rings with diamonds of size 0 carats. However, if the number is meaningful (dubious), then it refers to the cost of the gold content and other materials in the ring.

5.32 (a) The regression equation is $\hat{y} = -0.126 + 0.0608x$. Each increase of one unit in social distress increases brain activity by about 0.0608 units. For $x = 2.0$, this formula gives $\hat{y} = -0.0044$. (A student who uses the more precise coefficient estimates 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 $r^2 = 0.771$, we find $r = \sqrt{r^2} = 0.878$; the sign is positive because it has the same sign as the slope coefficient.

5.33 (a) The regression equation is $\hat{y} = 0.919 + 2.0647x$. For every degree Celsius, the toucan will lose about 2.06% more heat through its beak. **(b)** $\hat{y} = 0.919 + 2.0647(25) = 52.5$. At a temperature of 25 degrees Celsius, we predict a toucan to lose 52.5% more heat through its beak, on average. **(c)** Since R-Sq = 83.6%, 83.6% of the total variation in beak heat loss is explained by the straight-line relationship with temperature. **(c)** $r = \sqrt{r^2} = \sqrt{0.836} = 0.914$. Correlation is positive here, since the least-squares regression line has a positive slope.

5.34 Since we wish to regress husbands' heights on wives' heights, the women's heights will be the x-values, and the men's heights will be the y-values. **(a)** $b = r\, s_y / s_x = (0.5)(3.1/2.7) = 0.574$, and $a = \bar{y} - b\bar{x} = 69.9 - (0.574)(64.3) = 32.99$ inches. The regression equation is $\hat{y} = 32.99 + 0.574x$. For every inch of a wife's height, her husband is about 0.574 inches taller. **(b)** If

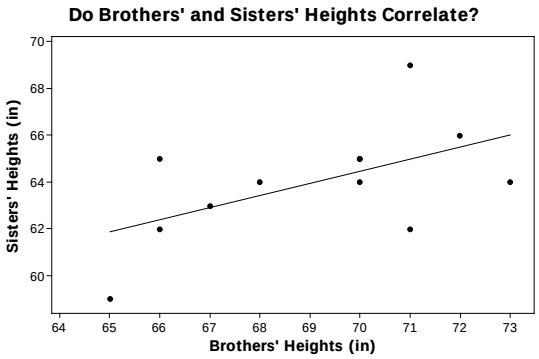


a wife is 67 inches tall, we predict her husband to have height $\hat{y}$ = 32.99 + (0.574)(67) = 71.448 inches. The plot, with this pair identified, is provided. **(c)** We don't expect this prediction to be very accurate because the heights of men having wives 67 inches tall varies a lot. Also, $r^2 = (0.5)^2 = 0.25$, so the linear regression explains only 25% of the variation in men's heights.

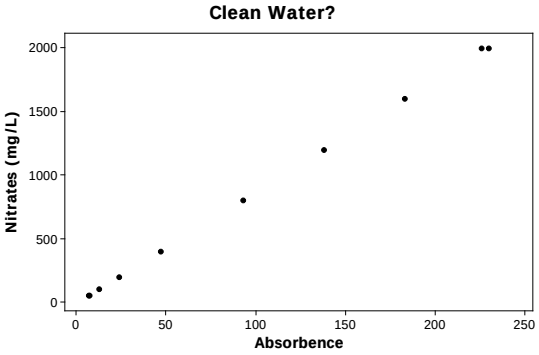
5.35 (a) $b = r\, s_y / s_x = (0.5)\left(\frac{8}{40}\right) = 0.1$, and $a = \bar{y} - b\bar{x} = 75 - (0.1)(280) = 47$. The regression equation is $\hat{y} = 47 + 0.1x$. Each point of pre-exam total score means an additional 0.1 points on the final exam, on average. **(b)** Julie's pre-final exam total was 300, so we would predict a final exam score of $\hat{y}$ = 47 + (0.1)(300) = 77. **(c)** Julie is right; with a correlation of r = 0.5, $r^2 = (0.5)^2 = 0.25$, so the regression line accounts for only 25% of the variability in student final exam scores. That is, the regression line doesn't predict final exam scores very well. Julie's score could, indeed, be much higher or lower than the predicted 77. Since she is making this argument, one might guess that her score was, in fact, higher. Julie should visit the Dean.

5.36 $r = \sqrt{0.16} = 0.40$ (high attendance goes with high grades, so the correlation must be positive).

5.37 (a) The regression equation is $\hat{y}$ = 28.037 + 0.521x. r = 0.555. **(b)** The plot is provided. Based on Damien's height of 70 inches, we predict his sister Tonya to have height $\hat{y}$ = 28.037 + (0.521)(70) = 64.5 inches (rounded). This prediction isn't expected to be very accurate because the correlation isn't very large; $r^2 = (0.555)^2 = 0.308$. The regression line explains only 30.8% of the variation in sister heights.

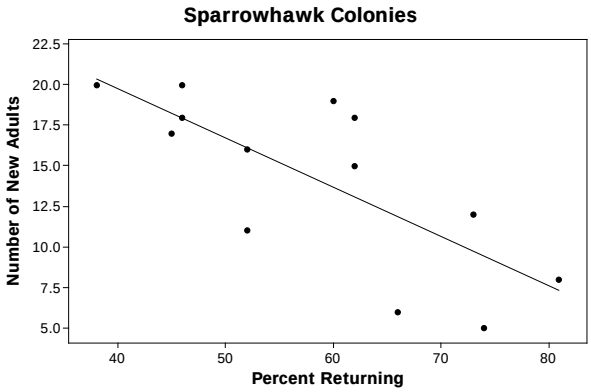


5.38 (a) The scatterplot suggests that the relationship between absorbence and Nitrates is extremely linear. From software, r = 0.99994 > 0.997, so the calibration does not need to be repeated. **(b)** From software, the equation of the least-squares regression line for predicting nitrates from absorbence is $\hat{y}$ = –14.522 + 8.825x. The slope of the line tells us that for each additional 1 unit increase in absorbence, nitrates are expected to increase by 8.825 mg/L, on average. If the water sample has absorbence of 40, we predict Nitrate concentration of –14.522 + (8.825)(40) = 338.478 mg/L. **(c)** We expect estimates of nitrate concentration from absorbence to be very accurate since the linear regression explains



virtually all of the variation in nitrate concentration. That is, $r^2 = (0.99994)^2 = 0.9999$, or 99.99% of the variation in nitrate concentration is explained by the regression on Absorbence.

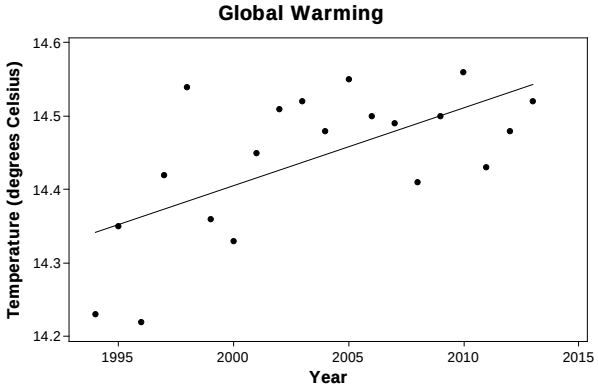
5.39 (a) The regression equation is $\hat{y} = 31.934 - 0.304x$. **(b)** The slope (–0.304) tells us that, on the average, for each additional 1% increase in returning birds, the number of new birds joining the colony decreases by 0.304. **(c)** When $x = 60$, we predict $\hat{y}$ =13.69 new birds will join the colony.



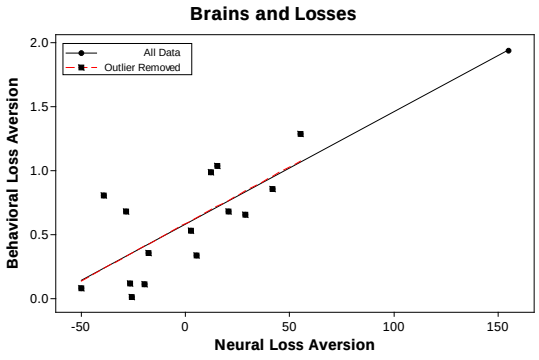
Minitab output

```
The regression equation is New = 31.93 – 0.3040PctRtn

Predictor   Coef    Stdev t-ratio    p
Constant  31.934    4.838   6.60  0.000
PctRtn  -0.30402  0.0812  -3.74   0.003
s= 3.667   R-sq=56.0%    R-sq(adj)=52.0%
```

5.40 (a) $\widehat{\text{Temp}} = -6.87 + 0.0106\text{Year}$. **(b)** The slope indicates that global temperatures are increasing about 0.01°C per year. **(c)** We estimate −6.87+0.0106(2050) = 14.86°C. Because this is extrapolation, we should not have much faith in the estimate.



5.41 (a) The outlier is in the upper-right corner. **(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 line) is greater with the outlier removed, and the outlier is located consistently with the linear pattern of the rest of the points. **(d)** The correlation changes from 0.8486 (with all points) to 0.7015 (without the outlier). With all points included, the regression line is $\hat{y} = 0.585 + 0.0879x$ (nearly indistinguishable from the other regression



line).

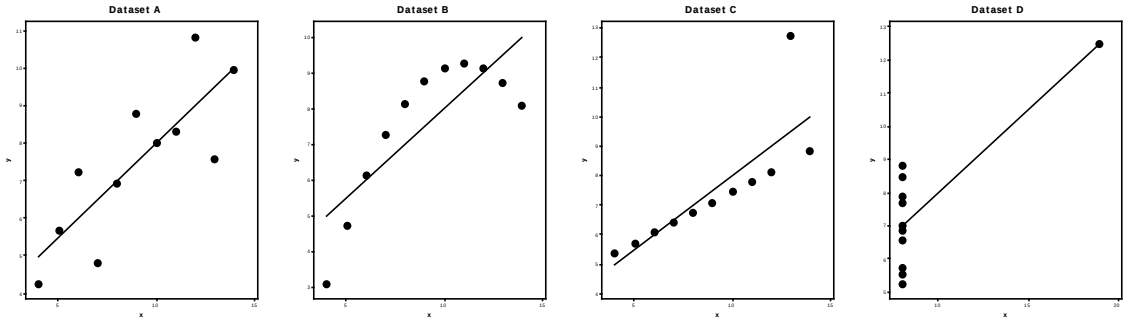
Minitab output – all points

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

Predictor   Coef  SE Coef   T   P
Constant  0.58496  0.07093 8.25 0.000
Neural   0.008794 0.001465 6.00 0.000
```

Minitab output – outlier removed

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

Predictor   Coef  SE Coef   T   P
Constant  0.58581  0.07506 7.80 0.000
Neural   0.008909 0.002510 3.55 0.004
```

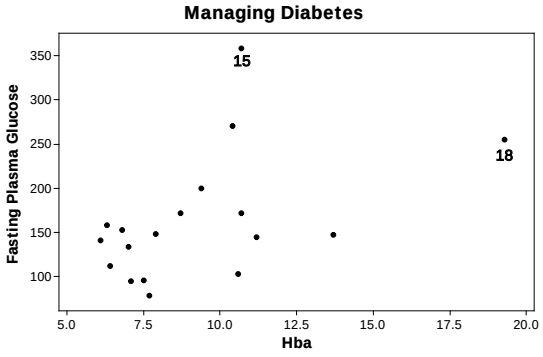
5.42 (a) To three decimal places, the correlations are all approximately 0.816 (for Set D, r actually rounds to 0.817), and the regression lines are all approximately $\hat{y} = 3.00 + 0.500x$. For all four sets, we predict $\hat{y} = 8$ when $x = 10$. **(b)** Plots below. **(c)** For set A, the use 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.)



5.43 (a) The two unusual observations are indicated on the scatterplot. **(b)** The correlations are

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

Both outliers change the correlation. Removing Subject 15 decreases r because its presence makes the scatterplot less linear. Removing



Subject 18 increases *r* because its presence decreases the relative scatter about the linear pattern.

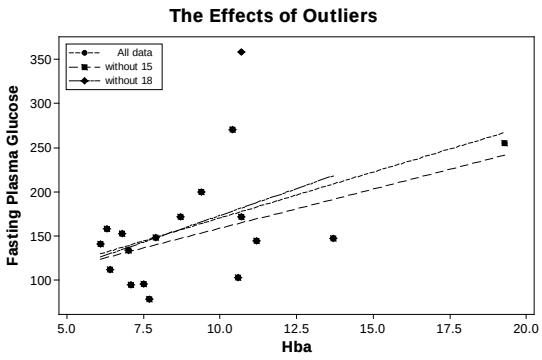
5.44 (a) $\hat{y}$ = 24.2 + 6.0*x*. **(b)** $\hat{y}$ = 24.2+ 0.6*x*. **(c)** When *x* = 0.7 mg, the first regression equation gives $\hat{y}$ = 28.4% body fat. Using the second equation, with *x* = 7 mg, $\hat{y}$ = 28.4% body fat. These are the same (as they should be).

5.45 The scatterplot with regression lines added is given. The equations are

$\hat{y}$ = 66.4 + 10.4*x* (all observations)

$\hat{y}$ = 69.5 + 8.92*x* (without #15)

$\hat{y}$ = 52.3 + 12.1*x* (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 it is not particularly influential because it is consistent with the linear pattern suggested by the other points.

5.46 The correlation would be much lower, because there is much greater variation in individuals than in the averages. The correlation in Exercise 4.25 was an ecological correlation, which obscures the variability in individuals.

5.47 The correlation would be smaller. Individual weight will vary much more than the average weight for a given height.

5.48 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.)

5.49 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).

5.50 (a) The regression equation is $\widehat{\text{MathSAT}} = 616.6 - 0.00148(\text{TeachSal})$. For each additional dollar of average teacher salary in a state, the average Math SAT score is expected to go down about 0.001 points. **(b)** This is unreasonable; the highest (and lowest) Math SAT averages correspond to states that pay teachers in the middle of the range (Illinois and Delaware). New York pays teachers the most and has a low average Math SAT score; it's very expensive to live in New York.

Regression Analysis: MathSAT versus Avg. teacher salaries 2013

```
The regression equation is
MathSAT = 617 - 0.00148 Avg. teacher salaries 2013

Predictor             Coef  SE Coef    T    P
Constant            616.60   40.11 15.37 0.000
Avg. teacher salaries 2013 -0.0014795 0.0007298 -2.03 0.048

S = 43.9615  R-Sq = 7.7%  R-Sq(adj) = 5.9%
```

5.51 (a) For states where more than 40% take the SAT, we have $\widehat{\text{MathSAT}} = 471.82 + 0.00048(\text{TeachSal})$. For these states, increasing the average teacher salary increases the mean Math SAT score by 0.00048 points, on average. **(b)** For states where less than 40% take the SAT, we have $\widehat{\text{MathSAT}} = 472.8 + 0.0020(\text{TeachSal})$. For these states, increasing the average teacher salary increases the mean Math SAT score by 0.0020 points, on average. **(c)** The slopes here have opposite signs from that found in Exercise 5.50. This is an example of Simpson's paradox with continuous variables (although none of the relationships are particularly strong). Consideration of a third (lurking) variable changed the relationship.

```
The regression equation is
MathSAT = 472 + 0.000480 Avg.tchrsal2013

Predictor       Coef  SE Coef
Constant      471.82   26.15
Avg.tchrsal2013 0.0004800 0.0004432

S = 20.2536  R-Sq = 4.7%
```

```
The regression equation is
MathSAT = 473 + 0.00202 Avg.tchrsal2013

Predictor      Coef  SE Coef
Constant      472.80   55.37
Avg.tchrsal2013 0.002022 0.001099

S = 28.7808  R-Sq = 12.8%
```

5.52 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.

5.53 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.

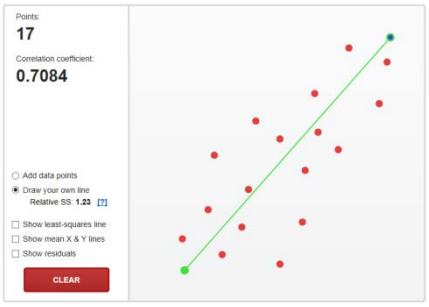
5.54 We have slope $b = r\, s_y / s_x$, and intercept $a = \bar{y} - b\bar{x}$, and $\hat{y} = a + bx$. When $x = \bar{x}$, $\hat{y} = a + b\,\bar{x} = (\bar{y} - b\,\bar{x}) + b\,\bar{x} = \bar{y}$.

5.55 For a player who shot 80 in the first round, we predict a second-round score of $\hat{y} = 56.47 + (0.243)(80) = 75.91$. For a player who shot 70 in the first round, we predict a second-round score of $\hat{y} = 56.47 + (0.243)(70) = 73.48$. Notice that the player who shot 80 the first round (worse than average) is predicted to have a worse-than-average score the second round, but better than the first round. Similarly, the player who shot 70 the first round (better than average) is predicted to do better than average in the second round, but not as well (relatively) as in the first round. Both players are predicted to "regress" to the mean.

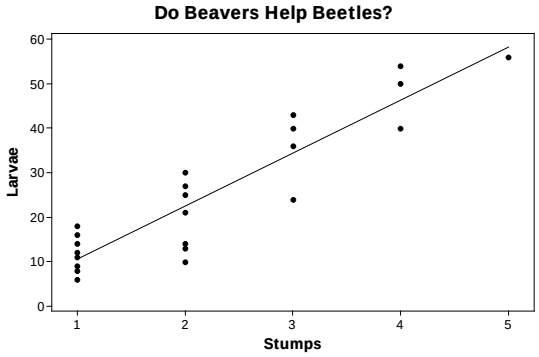
5.56 Note that $\bar{y} = 46.6 + 0.41\,\bar{x}$. We predict that Octavio will score 4.1 points above the mean on the final exam: $\hat{y} = 46.6 + 0.41(\bar{x} + 10) = 46.6 + 0.41\,\bar{x} + 4.1 = \bar{y} + 4.1$. (Alternatively, because the slope is 0.41, we can observe that an increase of 10 points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

5.57 See Exercise 4.41 for the 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.

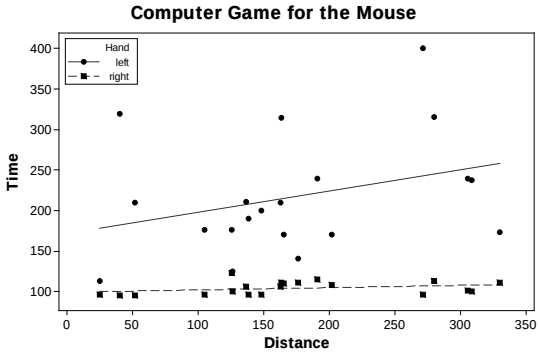
5.58 (a) Drawing the "best line" by eye is a very inaccurate process; few people choose the best line. **(b)** Most people tend to overestimate the slope for a scatterplot with $r = 0.7$; that is, most students will find that the least-squares line is less steep than the one they draw.



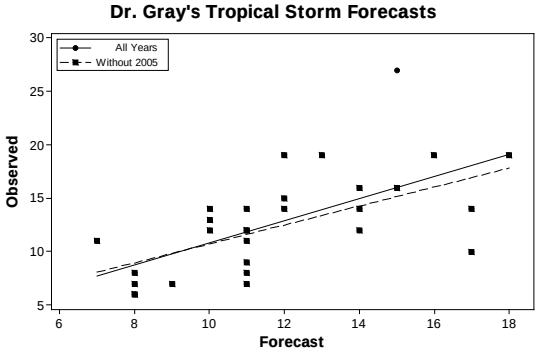
5.59 PLAN: We construct a scatterplot (with beaver stumps as the explanatory variable), and if appropriate, find the regression line and correlation. SOLVE: The scatterplot shows a positive linear association. Regression seems to be an appropriate way to summarize the relationship; the regression line is $\hat{y} = -1.286+11.89x$. The straight-line relationship explains r^2 =83.9% of the variation in beetle larvae. CONCLUDE: The strong positive association supports the idea that beavers benefit beetles.



5.60 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 squares and left-hand points are circles. In general, the right-hand points lie below the left-hand points, meaning the right-hand times are shorter, so the subject is likely 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, $\hat{y} = 99.36 + 0.0283x$ ($r = 0.305$, $r^2 = 9.3\%$), and for the left hand, $\hat{y} = 171.5 + 0.2619x$ ($r = 0.318$, $r^2 = 10.1\%$). CONCLUDE: Neither regression is particularly useful for prediction; distance accounts for only 9.3% (right) and 10.1% (left) of the variation in time.

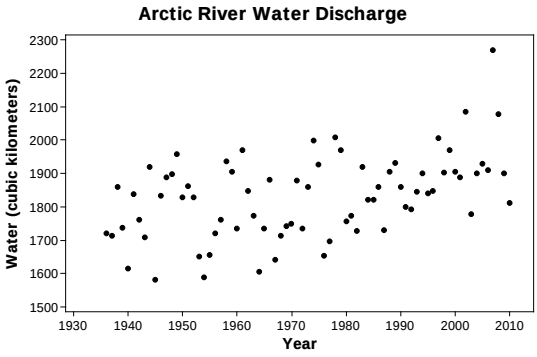


5.61 PLAN: We construct a scatterplot, with Forecast as the explanatory variable, and Actual as the response variable. If appropriate, we find the least-squares regression line. We consider the impact of the potential outlier (2005 season). SOLVE: The scatterplot shows a reasonable, but not very strong linear relationship between Forecast and Actual named storms. In recent years, it seems that there is no relationship between Forecast and Actual (a flat line). In the plot, the 2005 season is a noticeable outlier at the upper right. It is influential, pulling the regression line somewhat. We might consider deleting this point and fitting the line again. Deleting the point, we obtain the solid regression line, $\hat{y} = 2.753 +$



0.7964 Forecast when the original equation was $\hat{y}$ = 1.668 + 0.920 Forecast. If the forecasts were perfect, the intercept of this line would be 0, and the slope would be 1, for reference. Deleting the 2005 season, r = 0.628, and r^2 = 39.4%. Even after deleting the outlier, the regression line explains only 39.4% of variation in number of hurricanes. CONCLUDE: Predictions using the regression line are not very accurate. However, there is a positive association… so a forecast of many hurricanes may reasonably be expected to forebode a heavy season for hurricanes.

5.62 PLAN: We plot the data, producing a time-series plot. If appropriate, we consider fitting a regression line. SOLVE: The plot follows. We see that during the recent 10–15 years, the volume of discharge has become highly variable, but before then, the rate increased slowly, if at all. CONCLUDE: If there is a relationship between Year and Discharge, it isn't strongly linear, and use of a regression line would not be useful to predict Discharge from Year.

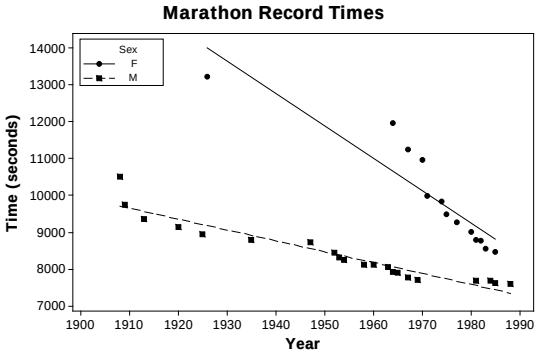


5.63 PLAN: We plot marathon times by year for each sex, using different symbols. If appropriate, we fit least-squares regression lines for predicting time from year for each gender. We then use these lines to guess when the times will concur. SOLVE: The scatterplot is provided below, with regression lines plotted. The regression lines are:

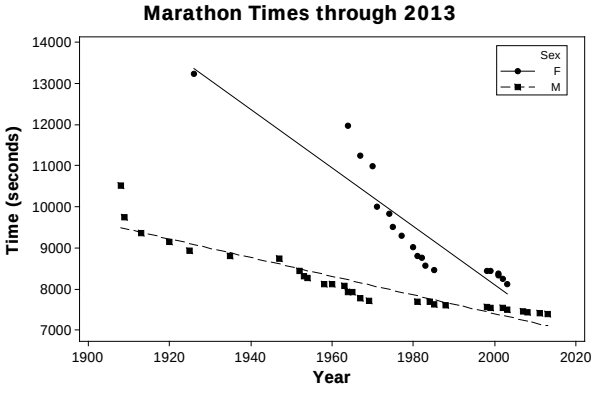
For men: $\hat{y}$ = 66,072 – 29.535x

For women: $\hat{y}$ = 182,976.15 – 87.73x

Although the lines appear to fit the data reasonably well (and the regression line for women would fit better if we omitted the outlier associated with year 1926), this analysis is inviting you to extrapolate, which is never advisable. CONCLUDE: Using the regression lines plotted, we might expect women to "outrun" men by the year 2009. Omitting the outlier, the line for women would decrease more steeply, and the intersection would occur sooner, by 1995. We'll note that as of 2014, this prediction has not happened.



5.64 For the men, we have $\widehat{\text{Time}}$ = 52,852 – 22.725(Year); for the women, $\widehat{\text{Time}}$ = 150,053 - 70.974(Year). When 52,852 – 22.725(Year) = 150,053 - 70.974(Year), the women will pass the men. Solving for Year, we have 48.249(Year) = 97,201, and Year =



2014.57. Given that at this writing, we are in August 2014, this estimate is not reliable, but it is better than the estimate obtained in Exercise 5.63.

5.65 – **5.67** are Web-based exercises.