# Chapter 3

## Section 3.1

***Check Your Understanding, page 144:***

1. The explanatory variable is the number of cans of beer. The response variable is the blood alcohol level.
2. There are two explanatory variables: amount of debt and income. The response variable is stress caused by college debt.

***Check Your Understanding, page 149:***

1. The relationship is positive. The longer the duration of the eruption, the longer the wait between eruptions is. One reason for this may be that if the geyser erupted for longer, it expended more energy and it will take longer to build up the energy needed to erupt again.
2. The form is roughly linear with two clusters. The clusters indicate that in general there are two types of eruptions, one shorter, the other somewhat longer.
3. The relationship is fairly strong. Two points define a line and in this case we could think of each cluster as a point, so the two clusters seem to define a line.
4. There are a few outliers around the clusters, but not many and not very distant from the main grouping of points.
5. The Starnes family needs to know how long the last eruption was in order to predict how long it will be until the next one.

***Check Your Understanding, page 154:***

1. (a) The correlation is about 0.9. This indicates that there is a strong, positive linear relationship between the number of boats registered in Florida and the number of manatees killed. (b) The correlation is about 0.5. This indicates that there is a moderate, positive linear relationship between the number of named storms predicted and the actual number of named storms. (c) The correlation is about 0.3. This indicates that there is a weak, positive linear relationship between the healing rate of the two front limbs of the newts. (d) The correlation is about -0.1. This indicates that there is a weak, negative linear relationship between last year's percent return and this year's percent return in the stock market.
2. If we remove the outlier in this scatterplot, the correlation would decrease. This point has the effect of strengthening the observed linear relationship we see.

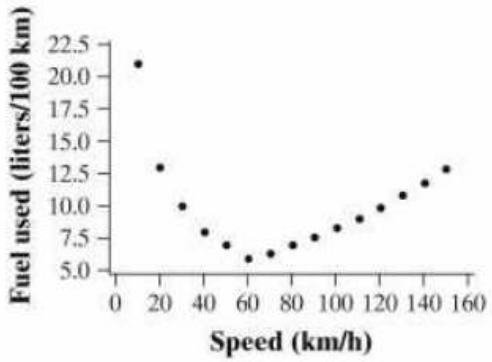
***Exercises, page 158:***

3.1 Water temperature is the explanatory variable, and weight change (growth) is the response variable. Both are quantitative.

3.2 The explanatory variable is the type of treatment—removal of the breast or removal of only the tumor and nearby lymph nodes, followed by radiation, and survival time is the response variable. Type of treatment is a categorical variable, and survival time is a quantitative variable.

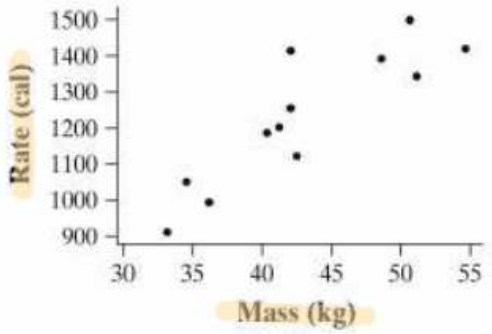
3.3 (a) A positive association between IQ and GPA means that students with higher IQs tend to have higher GPAs, and those with lower IQs generally have lower GPAs. The plot does show a positive association. (b) The form of the relationship is roughly linear, because a line through the scatterplot of points would provide a good summary. The positive association is moderately strong (with a few exceptions) because most of the points would be close to the line. (c) The lowest point on the plot is for a student with an IQ of about 103 and a GPA of about 0.4.

3.9 (a) A scatterplot with speed as the explanatory variable is shown below.




(b) The relationship is curved or quadratic. High amounts of fuel were used for low and high values of speed and low amounts of fuel were used for moderate speeds. This makes sense because the best fuel efficiency is obtained by driving at moderate speeds. (Note: 60 km/hr is about 37 mph) (c) Both are present. The first part of the graph (low speeds) would be described as negative and the second part (higher speeds) would be positive. (d) The relationship is very strong, with little deviation for a curve that can be drawn through the points.

3.10 (a) A scatterplot with mass as the explanatory variable is shown below.




r = .88

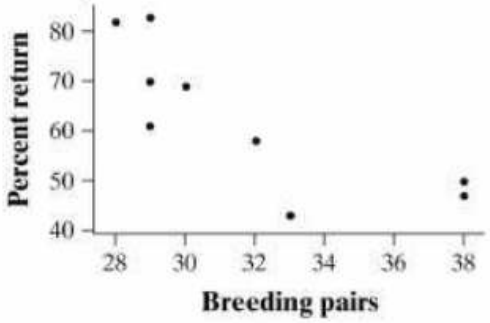
(b) The association is positive, and the relationship is linear and moderately strong.

3.11 (a) Most of the southern states blend in with the rest of the country. Several southern states do lie at the lower edges of their clusters. This means that, in general, the students in the southern states do not do as well as their counterparts in other portions of the country. (b) West Virginia is an outlier because it has a much lower mean SAT Math score than the other states which have a similar percent of students taking the exam.

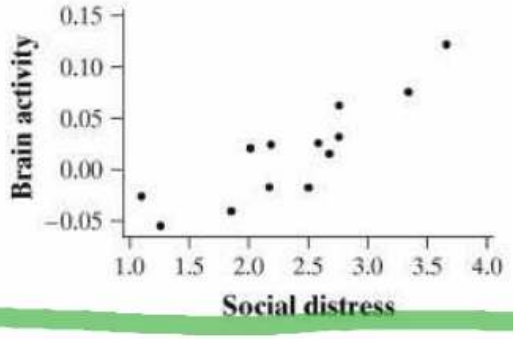
3.12 The scatterplot shows that the pattern of the relationship does hold both for men and women. However, the relationship between mass and rate is not as strong for men as it is for women. The group of men has higher lean body masses and metabolic rates than the group of women.

3.13 *State:* Is the relationship between the number of breeding pairs of merlins and the percent of males who return the next season negative? *Plan:* We will begin with a scatterplot, and compute the correlation if appropriate. *Do:* A scatterplot of the percent returning against the number of breeding pairs (shown below) shows the expected negative association. Though slightly curved, it is reasonable to compute

$r = -0.7943$ as a measure of the strength of the linear association. *Conclude:* This supports the theory: a smaller percent of birds survive following a successful breeding season.




3.14 *State:* Does social rejection cause activity in areas of the brain that are known to be activated by physical pain? *Plan:* We will begin with a scatterplot, and compute the correlation if appropriate. *Do:* A scatterplot (shown below) shows a fairly strong positive linear association. There are no particular outliers; each variable has low and high values, but those points do not deviate from the pattern of the rest. The relationship seems to be reasonably linear, so we compute $r = 0.8782$. *Conclude:* Social exclusion does appear to trigger a pain response: higher social distress measurements are associated with increased activity in the pain-sensing area of the brain.




3.15 (a) $r = 0.9$ (b) $r = 0$ (c) $r = 0.7$ (d) $r = -0.3$ (e) $r = -0.9$

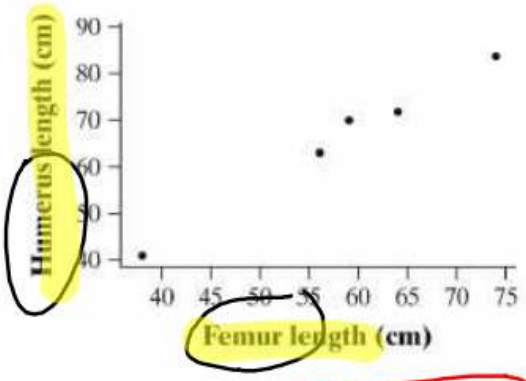
3.16 Answers may vary. We would expect the height of women at age 4 and their height as women at age 18 to be the highest correlation since it is reasonable to expect taller children to become taller adults and shorter children to become shorter adults. The next highest would be the correlation between the heights of male parents and their adult children. Tall fathers tend to have tall sons, but typically not as tall, and likewise for shorter fathers. The lowest correlation would be between husbands and their wives. Husbands may be taller than their wives in general, but there is no reason to expect anything more than a weak positive correlation.

3.17 (a) Gender is a categorical variable and the correlation coefficient $r$ measures the strength of linear association for two quantitative variables. (b) The largest possible value of the correlation coefficient $r$ is 1. (c) The correlation coefficient $r$ has no units.

3.18 The paper's report is wrong because the correlation ($r = 0.0$) is interpreted incorrectly. The author incorrectly suggests that a correlation of zero indicates a negative association between research productivity and teaching rating. The psychologist meant that there is no linear association between

research productivity and teaching rating. In other words, knowledge of a professor's research productivity will not help you predict her teaching rating.

3.19 (a) The scatterplot below shows a strong, positive, linear relationship between the two measurements. Thus, all five specimens appear to be from the same species.




r = .99

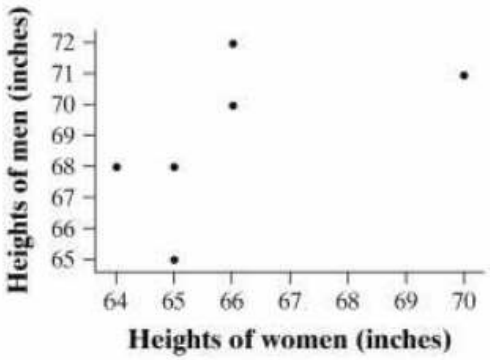
$$Z = \frac{X - M}{\sigma}$$

| femur | Humerus | zfemur | zhumerus | product |
|---|---|---|---|---|
| 38 | 41 | -1.53048 | -1.57329 | 2.40789 |
| 56 | 63 | -0.16669 | -0.18880 | 0.03147 |
| 59 | 70 | 0.06061 | 0.25173 | 0.01526 |
| 64 | 72 | 0.43944 | 0.37759 | 0.16593 |
| 74 | 84 | 1.19711 | 1.13277 | 1.35605 |

(b) The femur measurements have mean of 58.2 and a standard deviation of 13.2. The humerus measurements have a mean of 66 and a standard deviation of 15.89. The table below shows the standardized measurements (labeled zfemur and zhumerus) obtained by subtracting the mean and dividing by the standard deviation. The column labeled "product" contains the product (zfemur×zhumerus) of the standardized measurements. The sum of the products is 3.97659, so the correlation coefficient is

$r = \frac{1}{4}(3.97659) = 0.994$.

3.20 (a) The scatterplot shows a moderate positive association, so $r$ should be positive, but not close to 1.




(b) For the women, the mean is 66 and the standard deviation is 2.098. For the men, the mean is 69 and the standard deviation is 2.53. The table below shows the standardized measurements (labeled zfemale and zmale) obtained by subtracting the mean and dividing by the standard deviation. The column labeled "product" contains the product (zfemale×zmale) of the standardized measurements. The sum of the products is 2.82667, so the correlation coefficient is $r = \frac{1}{5}(2.82667) = 0.5653$.

| female | male | zfemale | zmale | product |
|---|---|---|---|---|
| 66 | 72 | 0 | 1.18585 | 0 |
| 64 | 68 | -0.95346 | -0.39528 | 0.37689 |
| 66 | 70 | 0 | 0.39528 | 0 |
| 65 | 68 | -0.47673 | -0.39528 | 0.18844 |
| 70 | 71 | 1.90693 | 0.79057 | 1.50756 |
| 65 | 65 | -0.47673 | -1.58114 | 0.75378 |

There is some evidence that taller women tend to date taller men (and shorter women date shorter men), but it is hardly overwhelming--and the small sample size makes any conclusion suspect.

3.21 (a) There is a strong positive linear association. High-calorie hot dogs tend to be high in salt, and low-calorie hot dogs tend to have lower sodium. (b) It would tend to decrease the correlation. An outlier generally in line with the bulk of the data will tend to increase the correlation.

3.22 (a) There is a strong positive linear association between body weight and brain weight of mammals. (b) It would tend to decrease the correlation. An outlier generally in line with the bulk of the data will tend to increase the correlation.

3.23 (a) The correlation would not change. It does not have units associated with it, so a change in units for either variable (or both) will not change the correlation. Multiplying both the $x$ and $y$ values by 10 will also multiply their standard deviations by 10, so the $z$-scores will not change. (b) The correlation would not change. The correlation measures the strength of the linear relationship between two quantitative variables. It does not distinguish between the explanatory and response variables.

3.24 (a) If all the men were 6 inches shorter, the correlation would not change. The correlation tells us that there is a weak to moderate association between women's heights and men's heights (that is, that taller women tend to date taller men), but it does not tell us whether or not they tend to date men taller than themselves. Subtracting 6 from each $y$ would also subtract 6 from the mean so the $z$-scores would

***Check Your Understanding, page 179:***

1. There is a moderate, positive linear relationship with one outlier in the bottom right corner of the plot.
2. The average error (residual) in predicting the backpack weight is 2.27 using the least-squares regression line.

***Check Your Understanding, page 181:***

1. c.
2. d.

***Exercises, page 191:***

3.35 The equation is $\hat{y} = 80 - 6x$ where $\hat{y}$ = the estimated weight of the soap and $x$ = the number of days since the bar was new.

3.36 The equation is $\hat{y} = 50 + x$ where $\hat{y}$ = the predicted reading test score and $x$ = the number of points above 100 for a child's IQ (this would be negative for a child whose IQ is less than 100).

3.37 (a) The slope is 1.109. We predict highway mileage will increase by 1.109 mpg for each 1 mpg increase in city mileage. (b) The intercept is 4.62 mpg. This is not statistically meaningful because this would represent the highway mileage for a car that gets 0 mpg in the city. (c) With city mpg of 16, the predicted highway mpg is $4.62 + 1.109(16) = 22.36$ mpg. With city mpg of 28, the predicted highway mpg is $4.62 + 1.109(28) = 35.67$ mpg.

3.38 (a) The slope is 0.882; this means that we predict reading scores will increase by 0.882 for each one-point increase in IQ. (b) The y-intercept is -33.4. This would only be statistically meaningful if a child could have an IQ score of 0. (c) The predicted scores for $x = 90$ and $x = 130$ are $-33.4 + 0.882(90) = 45.98$ and $-33.4 + 0.882(130) = 81.26$.

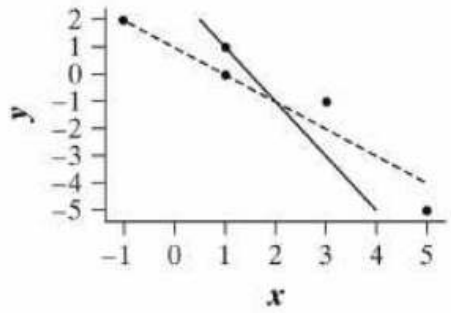
3.39 (a) The slope is −0.0053; this means that on the average for each additional week in the study the pH decreased by 0.0053 units. Thus, the acidity of the precipitation increased over time. (b) The $y$ intercept is 5.43 and it provides an estimate for the pH level at the beginning of the study. (c) At the end of the study pH is predicted to be $5.43 - 0.0053(150) = 4.635$.

3.40 (a) The slope is -19.87. We predict the amount of gas consumed in Joan's home to decrease by 19.87 cubic feet for every degree the average monthly temperature increases. (b) The $y$-intercept is 1425. When the average monthly temperature is 0°F, the predicted gas consumption for Joan's home is 1425 cubic feet. This is an extrapolation since the data only included points for months with an average temperature of more than 20°F. (c) $\widehat{\text{gas}} = 14.25 - 19.87(30) = 828.9$ cubic feet. We predict that the amount of natural gas Joan will use in a month with an average temperature of 30°F is 828.9 cubic feet.

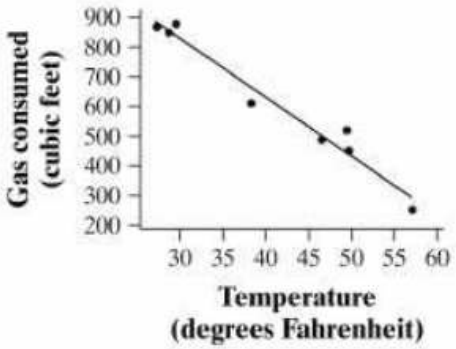
3.41 No. The data was collected weekly for 150 weeks. 1000 months corresponds to roughly 4000 weeks which is well outside the observed time period. We do not know that the linear relationship continues after 150 weeks. This constitutes extrapolation.

3.42 No. The average temperatures for the months where data was collected were between about 27°F and 57°F. 65°F is outside of this range so using the line to make a prediction here would be considered extrapolation. We do not know that the linear relationship continues after 57°F.

3.43 The dotted (red) line is the line $\hat{y} = 1 - x$ and the solid (black) line is the line $\hat{y} = 3 - 2x$. The dotted line comes closer to all of the data points. Thus, the line $\hat{y} = 1 - x$ fits the data best.




3.44




This line minimizes the square of the vertical distance between the points and the line.

3.45 The predicted value for this point is $\hat{y} = 5.43 - 0.0053(50) = 5.165$. So the residual is $5.08 - 5.185 = -0.085$. This means that the line predicted a pH value for that week that was 0.085 too large.

3.46 The predicted value for this point is $\hat{y} = 1425 - 19.87(46.4) = 503.032$. So the residual is $490 - 503.32 = -13.032$. This means that the line predicted that Joan would use 13.032 cubic feet of gas per day more than she actually did.

3.47 (a) The slope is $b = 0.5\left(\frac{2.7}{2.5}\right) = 0.54$. The $y$ intercept is $a = 68.5 - 0.54(64.5) = 33.67$. So the equation for predicting $y$ = husband's height from $x$ = wife's height is $\hat{y} = 33.67 + 0.54x$. (b) The predicted height is $\hat{y} = 33.67 + 0.54(67) = 69.85$ inches. 67 inches is one standard deviation above the mean for women. So the predicted value for husband's height would be $\overline{y} + rs_y = 68.5 + 0.5(2.7) = 69.85$.

3.48 (a) The slope is $b = 0.596\left(\frac{15.35}{5.36}\right) = 1.707$. The $y$ intercept is $a = 9.07 - 1.707(1.75) = 6.083$. So the regression equation is $\hat{y} = 6.083 + 1.707x$. ~~(b) The predicted change is~~

Annual Change = 6.083 + 1.707 (Jan Change)

$\hat{y} = 6.083 + 1.707(1.75) = 9.0703\%$. We could have given the answer without doing calculations because the regression line must pass through $(\bar{x}, \bar{y}) = (1.75, 9.07)$.

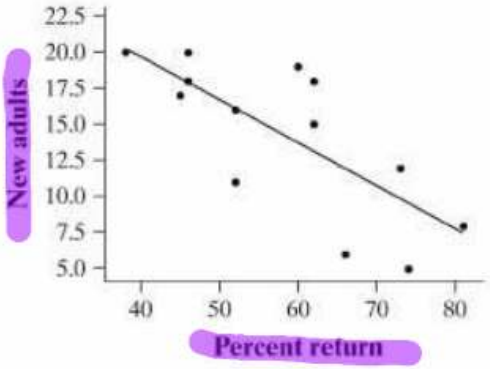
3.49 (a) $r^2 = (0.5)^2 = 0.25$. Thus, the straight-line relationship explains 25% of the variation in husbands heights. (b) The average error (residual) when using the line for prediction is 1.2 inches.

3.50 (a) $r^2 = (0.596)^2 = 0.3552$. Thus, the straight-line relationship explains 35.52% of the variation in yearly changes. (b) The average error (residual) when using the line for prediction is 8.3%.

3.51 (a) The least-squares line for predicting $y$ = GPA from $x$ = IQ has slope $b = 0.6337\left(\frac{2.1}{13.17}\right) = 0.101$ and intercept $a = 7.447 - 0.101(108.9) = -3.5519$. Thus, the regression line is [illegible]. (b) $r^2 = (0.6337)^2 = 0.4016$. Thus, 40.16% of the variation in GPA is accounted for by the linear relationship with IQ. (c) The predicted GPA for this student is $\hat{y} = -3.5519 + 0.101(103) = 6.8511$ and the residual is $0.53 - 6.8511 = -6.3211$. This means that the student had a GPA that was 6.3211 points worse than expected for someone with an IQ of 103.

3.52 Since the least-squares regression line must pass through the point of averages, we know that $\bar{y} = 46.6 + 0.41\bar{x}$. Octavio's predicted final exam score is $\hat{y} = 46.6 + 0.41(\bar{x} + 10) = (46.6 + 0.41\bar{x}) + 0.41(10) = \bar{y} + 4.1$. Thus, we predict that he will score 4.1 points above the mean on the final exam.

3.53 (a) The scatterplot is




(b) The least squares regression line is $\hat{y} = 31.9 - 0.304x$. Minitab output is shown below. See the scatterplot above as well.

**Minitab output**

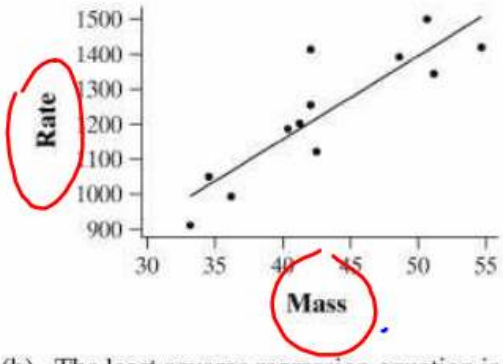
The regression equation is
newadults = 31.9 - 0.304 %returning

| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | 31.934 | 4.838 | 6.60 | 0.000 |
| %returning | -0.30402 | 0.08122 | -3.74 | 0.003 |

S = 3.66689 R-Sq = 56.0% R-Sq(adj) = 52.0%

(c) The slope tells us that as the percent of returning birds increases by one, we predict the number of new birds will decrease by −0.304. The $y$ intercept provides a prediction that we will see 31.9 new adults in a new colony when the percent of returning birds is zero. This value is clearly outside the range of values studied for the 13 colonies of sparrowhawks and has no practical meaning in this situation. (d) The predicted value for the number of new adults is $31.9 - 0.304(60) = 13.66$ or about 14.

3.54 (a) The scatterplot is shown below.




(b) The least squares regression equation is $\hat{y} = 201.2 + 24.026x$. Minitab output is shown below. See the scatterplot above as well.

Metabolic Rate = 201.2 + 24.026 (Mass)
Calories/day
kg

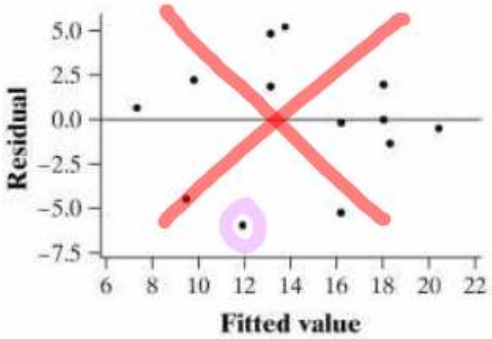
**Minitab output**
The regression equation is
Rate = 201 + 24.0 Mass

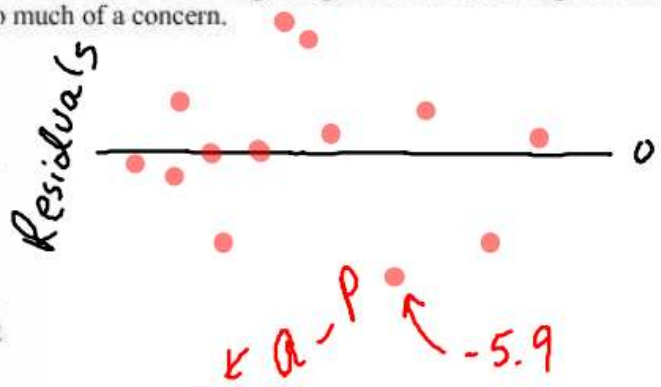
| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | 201.2 | 181.7 | 1.11 | 0.294 |
| Mass | 24.026 | 4.174 | 5.76 | 0.000 |

S = 95.0808 R-Sq = 76.8% R-Sq(adj) = 74.5%

(c) The slope tells us that we would predict an increase in the metabolic rate of about 24 cal/day for each additional kilogram of body mass. (d) For $x = 45$ kg, the predicted metabolic rate is $\hat{y} = 1282.3$ cal/day.

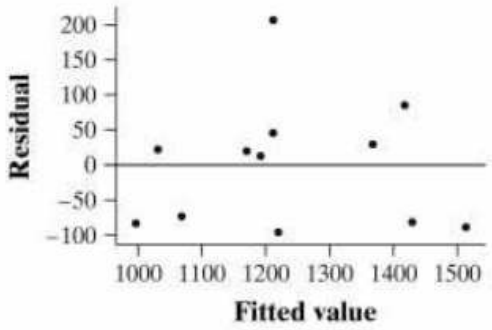
3.55 (a) The residual plot (shown below) suggests that the line is a decent fit. The points are all scattered around a residual value of 0. It is noteworthy that there are three larger negative residuals, but given the size of this data set, these are probably not too much of a concern.




(b) The point with the largest residual has a residual of about -6. This means that the line over-predicted the number of new adults by 6.

3.56 (a) The residual plot (shown below) shows that the linear fit is good. There is one large, positive, outlier, but since it is near the mean of the mass values, it does not influence the line very much.




(b) The point that has the largest residual has a residual of about 200. This means that the line greatly under-predicted the metabolic rate for this particular person.

3.57 56% of the variation in the number of new adult birds is explained by the straight-line relationship. The average error (residual) when using the line for prediction is 3.67%.

3.58 76.8% of the variation in the metabolic rate is explained by the straight-line relationship. The average error (residual) when using the line for prediction is 95.08 calories burned per 24 hours.

3.59 (a) There is a positive, linear association between the two variables. There is more variation in the field measurements for larger laboratory measurements. Also, the values are scattered above and below the line $y = x$ for small and moderate depths, indicating strong agreement, but the field measurements tend to be smaller than the laboratory measurements for large depths. (b) The points for the larger depths fall systematically below the line $y = x$ showing that the field measurements are too small compared to the laboratory measurements. (c) In order to minimize the sum of the squared distances from the points to the regression line, the top right part of the blue line in the scatterplot would need to be pulled down to go through the "middle" of the group of points that are currently below the blue line. Thus, the slope would decrease and the intercept would increase.

3.60 The residual plot clearly shows that the prediction errors increase for larger laboratory measurements. In other words, the variability in the field measurements increases as the laboratory measurements increase. The least squares line does not provide a great fit, especially for larger depths.

3.61 Clearly, this line does not fit the data very well; the data show a clearly curved pattern. The residual plot shows a clear curved pattern, with the first two and the last four residuals being negative and those between 3 and 8 months being positive.

3.62 We would certainly not use the regression line to predict fuel consumption. The scatterplot shows a nonlinear relationship.

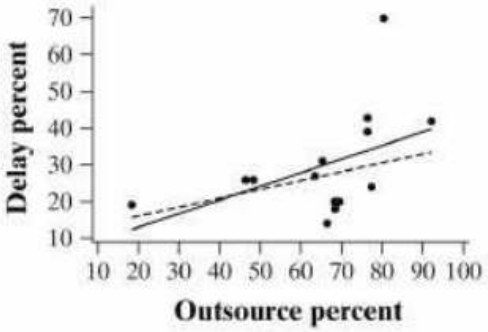
3.63 (a) The regression line is $\hat{y} = 157.68 - 2.99x$. Following a season with 30 breeding pairs, we find $\hat{y} = 157.68 - 2.99(30) = 67.98$ so we predict that about 68% of males will return. (b) This is given in the Minitab output as R-sq = 63.1%. The linear relationship explains 63.1% of the variation in the percent of returning males.

Percent Males Returning = 157.68 - 2.9935(# Breeding Pairs)

(c) Knowing that $r^2 = 0.631$, we find $r = -\sqrt{r^2} = -0.79$; the sign is negative because it has the same sign as the slope coefficient. (d) Since $s = 9.46$, the typical error when using the line to predict the return rate of males is about 9.46%.

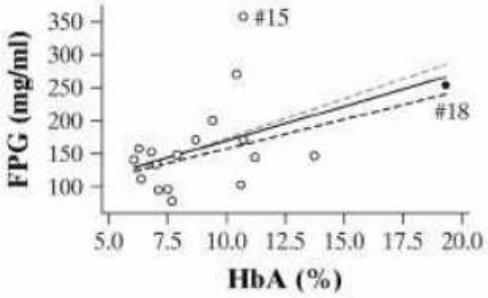
3.64 (a) The regression equation is $\hat{y} = -0.126 + 0.0608x$. For $x = 2.0$, this formula gives $\hat{y} = -0.126 + 0.0608(2) = -0.0044$. (b) This is given in the Minitab output as R-sq = 77.1%. The linear relationship explains 77.1% of the variation in brain activity. (c) Knowing that $r^2 = 0.771$, we find $r = +\sqrt{r^2} = 0.88$; the sign is positive because it has the same sign as the slope coefficient. (d) Since $s = 0.0251$, the typical error when using the line to predict the activity in the brain is about 0.025.

3.65 (a) The scatterplot (with regression lines) is shown below.




(b) The correlation is $r = 0.4765$ 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, $\hat{y} = 4.73 + 0.3868x$ (the solid, blue, line), and the prediction for $x = 76$ is 34.13%. With Hawaiian Airlines removed, $\hat{y} = 10.878 + 0.2495x$ (the dotted, black, 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.

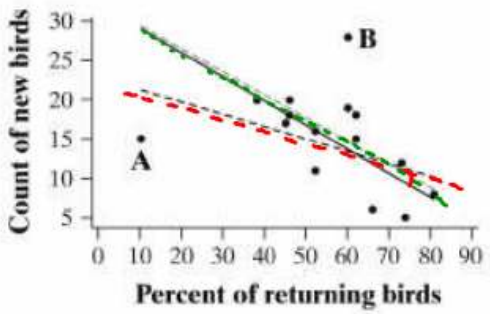
3.66 (a) A scatterplot, with the two unusual observations marked and the three separate regression lines added, is shown below.




(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 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. (c) The three regression lines shown in the

scatterplot above are: $\hat{y} = 66.4 + 10.4x$ (all observations); $\hat{y} = 69.5 + 8.92x$ (without #15); $\hat{y} = 52.3 + 12.1x$ (without #18). While the equation changes in response to removing either subject, one could argue that neither one is particularly influential, as 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.

3.67 (a) A scatterplot with the two new points is shown below. Point A is a horizontal outlier; that is, it has a much smaller $x$-value than the others. Point B is a vertical outlier; it has a higher $y$-value than the others.



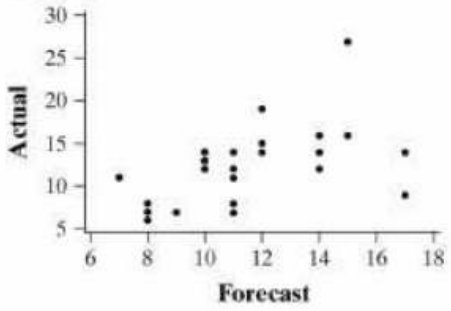

A and B are both outliers; only A is influential

(b) The three regression formulas are: $\hat{y} = 31.9 - 0.304x$ (the original data, solid black line); $\hat{y} = 22.8 - 0.156x$ (with Point A, dashed red line); $\hat{y} = 32.3 - 0.293x$ (with Point B, dotted green line). Adding Point B has little impact. Point A is influential; it pulls the line down, and changes how the line looks relative to the original 13 data points.

3.68 Answers may vary. For example: Weight, gender, other food eaten by the students, type of beer (light, imported, …).

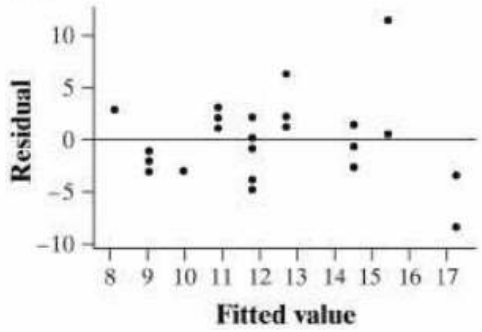
3.69 *State:* How accurate are Dr. Gray's forecasts? *Plan:* We will construct a scatterplot with Dr. Gray's forecast as the explanatory variable, and if appropriate, find the regression equation. Then we should make a residual plot and calculate $r^2$ and $s$. *Do:* The scatterplot shows a moderate positive association; the regression line is $\hat{y} = 1.688 + 0.9154x$, with $r^2 = 0.30$ and $s = 4.0$. The relationship is strengthened by the large number of storms in the 2005 season, but it is weakened by the 2006 and 2007, 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 $\hat{y} = 3.977 + 0.6699x$, with $r^2 = 0.265$ and $s = 3.14$.

(a)



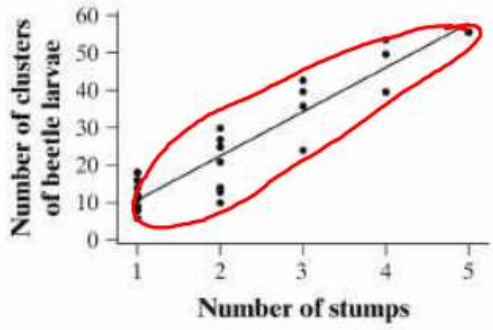

Finally, the residual plot does not indicate any problems with fitting the linear equation.

(b)




*Conclude:* If Dr. Gray forecasts $x = 16$ tropical storms, we expect 16.33 storms in that year. However, we do not have very much confidence in this estimate, because the regression line explains only 30% of the variation in tropical storms and the typical error we should expect when using this line for prediction is 4 storms. (If we exclude 2005, the prediction is 14.7 storms, but this estimate is less reliable than the first.)

3.70 *State:* Do more stumps result in more beetles? How accurate will our predictions be? *Plan:* We will construct a scatterplot with the number of stumps as the explanatory variable, and if appropriate, find the regression equation. Then we should make a residual plot and compute $r^2$ and $s$. *Do:* A scatterplot, with the least-squares regression line, is shown below. The plot shows a strong, positive linear association between the number of beaver-caused stumps and the number of beetle larvae clusters.




1) Scatterplot

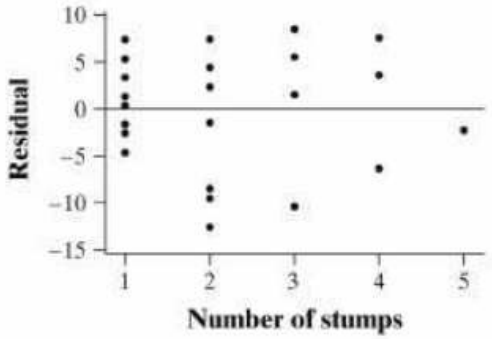
2) $r = .9159$

$r^2 = .839$

~~$s = 6.42$~~

$\widehat{\text{BeetleLarvae}} = -1.29 + 11.89(\#\text{ Stumps})$

The least-squares regression line is [illegible] and the residual plot is shown below. The linear model appears to provide a very good fit.




3) Residual Plot

4) Linear Model ?

Finally, $r^2 = 0.839$ and $s = 6.42$. In other words, about 84% of the variation in the number of beetle larvae clusters is accounted for by the linear relationship with the number of stumps and our average error in prediction is about 6.4 larvae. *Conclude:* More stumps (i.e. more beavers) does appear to lead to more beetle larvae. However we should be cautious because we have few observations with 4 or 5 stumps.

3.71 b

3.72 c

3.73 b

3.74 a

3.75 b

3.76 a

3.77 d

3.78 a

3.79 About 92.92%. For the N(18.7,4.3) distribution, $x < 25$ corresponds to $z < \dfrac{25 - 18.7}{4.3} = 1.47$, for which Table A gives $0.9292 = 92.92\%$.

3.80 At least 24.2 mpg. Search Table A for the proportion closest to 0.90; this is $z = 1.28$, the 90th percentile for the N(0,1) distribution. The top 10% of all vehicles are those with gas mileage at least 1.28 standard deviations above the mean:
$18.7 + 1.28(4.3) = 24.2$ mpg or more.

3.81 (a) There is evidence of an association between accident rate and marijuana use. Those people who use marijuana more are more likely to have caused accidents.

$\hat{Final\ Exam} = 30.2 + .16(Pre\ Exam\ Scores)$

R3.6 (a) The slope of the regression line for predicting final-exam score from pre-exam totals is $b = 0.6\left(\frac{8}{30}\right) = 0.16$; for every extra point earned on the midterm, we predict that the score on the final exam will increase by about 0.16. The intercept of the regression line is $a = 75 - 0.16(280) = 30.2$. (b) Julie's predicted final exam score is $\hat{y} = 30.2 + 0.16(300) = 78.2$. (c) $r^2 = 0.36$, so only 36% of the variability in the final exam scores is accounted for by the linear relationship with pre-exam totals. About 64% of the individual variation is not accounted for by the least squares regression line, so Julie has a good reason to think this is not a good estimate.

R3.7 (a) If we left out Hawaii, the correlation would decrease because Hawaii is above average for both the maximum 24 hour precipitation and the maximum annual precipitation. (b) The blue line is the line calculated with all 50 states. Hawaii's point is influential and pulls the line up toward it. The other line is the one with all states except Hawaii. (c) If we change the measurement on both $x$ and $y$ from inches to feet: the correlation will not change since it does not have units, $s$ will decrease since it would now be measured in feet as well, the slope of the regression line would not change since both x and y are measured in the same units leading to a slope without units, but the y-intercept would decrease since it changes units from inches to feet. If we switch the explanatory and response variables, the correlation will not change, but the standard error and the least squares line will.

## AP Statistics Practice Test (page 200)

T3.1 d. A correlation of near zero indicates no (or little) linear relationship, either positive or negative. Answers a, b and e indicate some form of linear relationship. Answer c implies no relationship whatsoever. It is possible to have a correlation where there is a strong relationship, just not a linear one.

T3.2 e. This point is influential because it is well above the mean for the amount spent on tobacco and well below the mean for the amount spent on alcohol. The observation (4.5, 6.0) is not an outlier because it does not have the greatest value in either dimension, nor does it fall outside the main pattern of the data set.

T3.3 c. This is the definition of $r^2$.

T3.4 a. The slope for the least-squares line depends on which variable is the explanatory variable and which is the response. Also, the slope $b = r\frac{s_y}{s_x}$ so $\frac{b}{r} = \frac{0.865}{0.79} = 1.09 = \frac{s_y}{s_x}$ which implies that $s_y > s_x$.

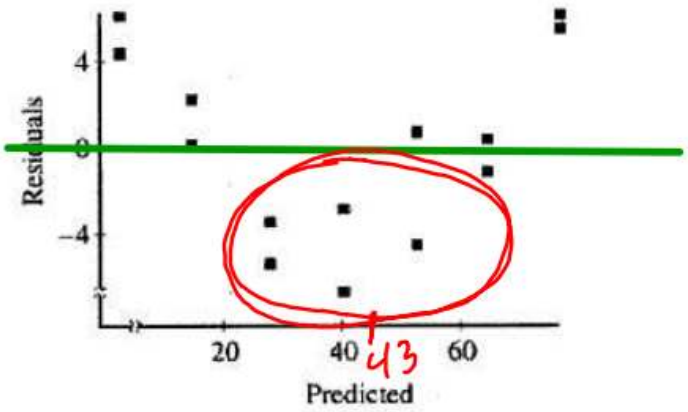
T3.5 a. The line predicts that a fish would have activity level $\hat{y} = 148.617 - 3.21667(20.4) = 83.0$. Looking at the residual plot, the fish with activity level 83 has a residual of about 3. Since residual $= y - \hat{y}$, we find that $y = \hat{y} + \text{residual} = 83 + 3 = 86$.

T3.6 c. This is another way of saying the average error between the actual values and the predicted values using the linear model.

4. In a study of the application of a certain type of weed killer, 14 fields containing large numbers of weeds were treated. The weed killer was prepared at seven different strengths by adding 1, 1.5, 2, 2.5, 3, 3.5, or 4 teaspoons to a gallon of water. Two randomly selected fields were treated with each strength of weed killer. After a few days, the percentage of weeds killed on each filed was measured. The computer output obtained from fitting a least squares regression line to the data is shown below. A plot of the residuals is provided as well.

Dependent variable is: percent killed
R squared = 97.2% R squared (adjusted) = 96.9%
$s$ = 4.505 with 14 − 2 = 12 degrees of freedom

| Source | Sum of Squares | df | Mean Square | F-ratio |
|---|---|---|---|---|
| Regression | 8330.16 | 1 | 8330.16 | 410 |
| Residual | 243.589 | 12 | 20.2990 | |
| **Variable** | **Coefficient** | **s.e. of Coeff** | **t-ratio** | **Prob** |
| Constant | -20.5893 | 3.242 | -6.35 | ≤ 0.0001 |
| No. Teaspoons | 24.3929 | 1.204 | 20.3 | ≤ 0.0001 |




| X | Y |
|---|---|
| 2.6 | 42.8% |

(a) What is the equation of the least squares regression line given by this analysis? Define any variables used in this equation.

Weeds Killed = -20.5893 + 24.3929 (# Teaspoons)

(b) If someone uses this equation to predict the percentage of weeds killed when 2.6 teaspoons of weed killer are used, which of the following would you expect?

- ○ The prediction will be too large.
- ○ The prediction will be too small.
- ○ A prediction cannot be made based on the information given on the computer output.

Explain your reasoning.

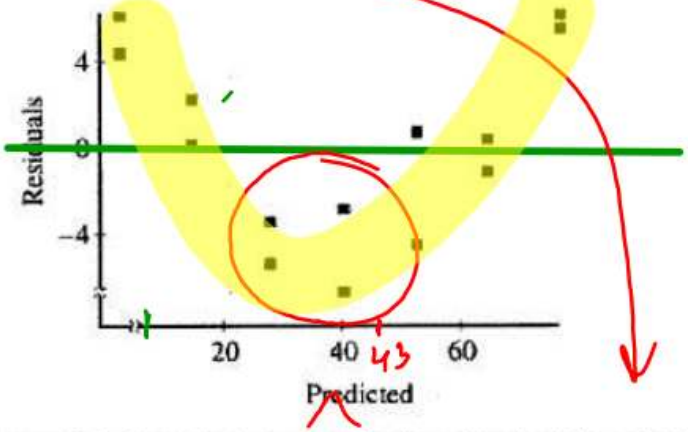
NEXT PAGE

4. In a study of the application of a certain type of weed killer, 14 fields containing large numbers of weeds were treated. The weed killer was prepared at seven different strengths by adding 1, 1.5, 2, 2.5, 3, 3.5, or 4 teaspoons to a gallon of water. Two randomly selected fields were treated with each strength of weed killer. After a few days, the percentage of weeds killed on each filed was measured. The computer output obtained from fitting a least squares regression line to the data is shown below. A plot of the residuals is provided as well.

Dependent variable is: percent killed
R squared = 97.2% R squared (adjusted) = 96.9%
$s = 4.505$ with $14 - 2 = 12$ degrees of freedom

| Source | Sum of Squares | df | Mean Square | F-ratio |
|---|---|---|---|---|
| Regression | 8330.16 | 1 | 8330.16 | 410 |
| Residual | 243.589 | 12 | 20.2990 | |

| Variable | Coefficient | s.e. of Coeff | t-ratio | Prob |
|---|---|---|---|---|
| Constant | -20.5893 | 3.242 | -6.35 | $\leq 0.0001$ |
| No. Teaspoons | 24.3929 | 1.204 | 20.3 | $\leq 0.0001$ |




X | Y
2.6 | 43%

(a) What is the equation of the least squares regression line given by this analysis? Define any variables used in this equation.

% Killed = -20.5893 + 24.3929 (# Tea)

(b) If someone uses this equation to predict the percentage of weeds killed when 2.6 teaspoons of weed killer are used, which of the following would you expect?

- The prediction will be too large.
- The prediction will be too small.
- A prediction cannot be made based on the information given on the computer output.

Explain your reasoning.

NEXT PAGE

STATISTICS

SECTION II

Part A

Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II grade—75

Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy of your results and explanation.

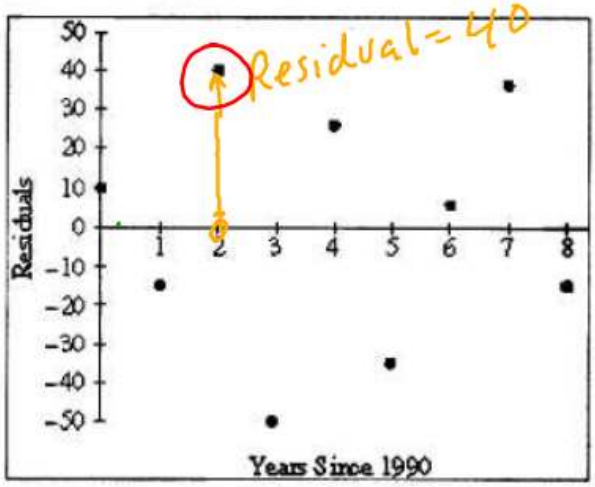
1. Lydia and Bob were searching the Internet to find information on air travel in the United States. They found data on the number of commercial aircraft flying in the United States during the years 1990-1998. The dates were recorded as years since 1990. Thus, the year 1990 was recorded as year 0. They fit a least squares regression line to the data. The graph of the residuals and part of the computer output for their regression are given below.

$r = a - p$

$40 = a - 3407$

$a = 3447$

Residual = 40




Random?

| Predictor | Coef | Stdev | t-ratio | p |
|---|---|---|---|---|
| Constant | 2939.93 | 20.55 | 143.09 | 0.000 |
| Years | 233.517 | 4.316 | 54.11 | 0.000 |

s = 33.43

slope

a. Is a line an appropriate model to use for these data? What information tells you this? yes - random

b. What is the value of the slope of the least squares regression line? Interpret the slope in the context of this situation. 233.517 → 234 planes inc every year

c. What is the value of the intercept of the least squares regression line? Interpret the intercept in the context of this situation. 2939.93 → In 1990, 2940 planes

d. What is the predicted number of commercial aircraft flying in 1992 ?

e. What was the actual number of commercial aircraft flying in 1992 ?

d) $\hat{\#\text{Airplanes}} = 2939.93 + 233.517(\text{year})$

| x | $\hat{y}$ |
|---|---|
| 2 | 3406.964 → 3407 ✓ |

GO ON TO THE NEXT PAGE