Section 12.1: Modeling How Two Variables Are Related

12.1 Car mileage and weight

a) The response variable is mileage, and the explanatory variable is weight.

b) $\hat{y} = 45.6 - 0.0052x$; The y-intercept is 45.6 and the slope is –0.0052.

c) For each 1000 pound increase in the vehicle, the predicted mileage will decrease by 5.2 miles per gallon.

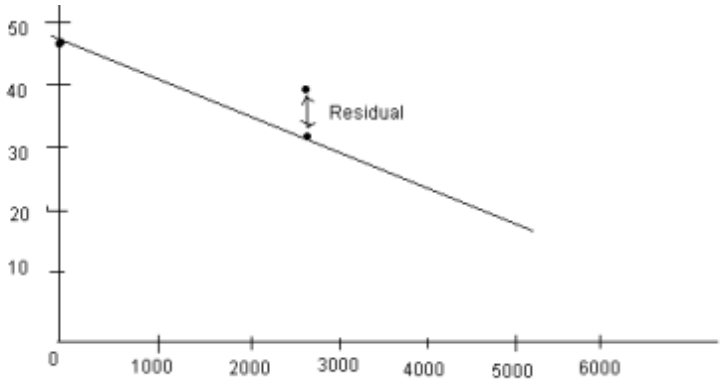
d) The y-intercept is the predicted miles per gallon for a car that weighs 0 pounds. This is far outside the range of the car weights in this database and, therefore, does not have contextual meaning for these data.

12.2 Predicting car mileage

a) $\hat{y} = 45.6 - 0.0052(2590) = 32.1$

b) $y - \hat{y} = 38 - 32.1 = 5.9$

c)



12.3 Predicting maximum bench strength in males

a) $\hat{y} = 117.5 + 5.86(35) = 322.6$

b) $\hat{y} = 117.5 + 5.86(0) = 117.5$

c) The y-intercept indicates that for male athletes who cannot perform any repetitions for a fatigue bench press, the predicted maximum bench press is 117.5 kg. As repetitions to fatigue bench press (repBP) increases from 0 to 35, the predicted maximum bench press (maxBP) increases from 117.5 kg to 322.6 kg.

12.4 Higher income with experience

a) The mean y value for an experience of 5 years is $\mu_y = -10,000 + 9,500(5) = 37,500$. The variability, based on $\sigma = 6,500$ (within three standard deviations of the mean, but with a floor of zero), would likely include values from 18,000 to 57,000.

b) The mean y value for an experience of 10 years is $\mu_y = -10,000 + 9,500(10) = 85,000$. The variability, based on $\sigma = 6,500$ (within three standard deviations of the mean, but with a floor of zero), would likely include values from 65,500 to 104,500.

12.5 Ensuring linear relationship

To ensure if the relationship between dependent variable and independent variable is linear or not, draw a scatter plot between dependent variable and independent variable to see if it exhibits a linear relationship, i.e. a line approximates the pattern between dependent variable and independent variable.

12.6 Fast food and indigestion

The mean would be the mean number of times fast food was eaten in the past month for individuals who had indigestion for a specific number of times (e.g., 3) in the past month. Variability would be the number that the actual individuals who had indigestion for a specific number of times varied in terms of number of times they ate fast food.

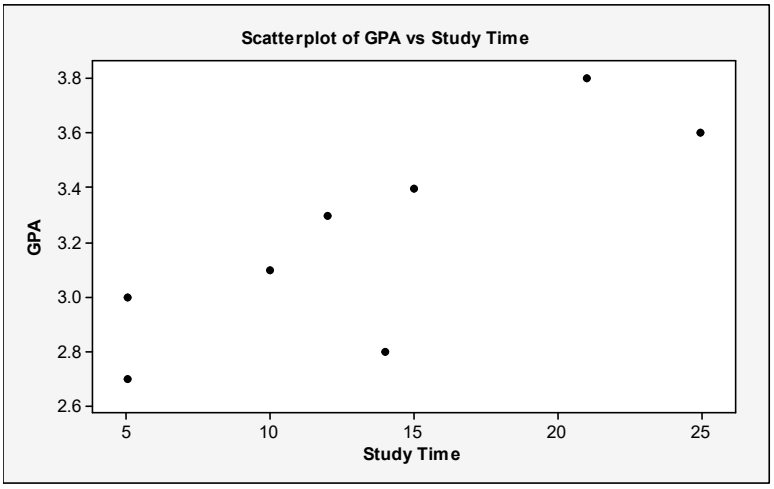
12.6 (continued)

a) It is more sensible to use a straight line model for the means of the conditional distributions rather than individual observations because individual observations are not likely to fall in a line, even if their means do.

b) The model needs to allow variations around the mean to account for the fact that people who had indigestion for a certain number of times–say, three times–might have eaten fast food for a different number of times.

12.7 Study time and college GPA

a)



Based on the scatterplot, there appears to be a positive association between GPA and study time.

b) From MINITAB: Predicted GPA = 2.63 + 0.0439Studytime. For every 1 hour increase in study time per week, GPA is predicted to increase by about 0.04 points.

c) Predicted GPA = 2.63 + 0.0439(25) = 3.73

d) $y - \hat{y} = 3.6 - 3.73 = -0.13$; The observed GPA for Student 2, who studies an average of 25 hours per week, is 3.6, which is 0.13 points below the predicted GPA of 3.73.

12.8 GPA and skipping class

a)

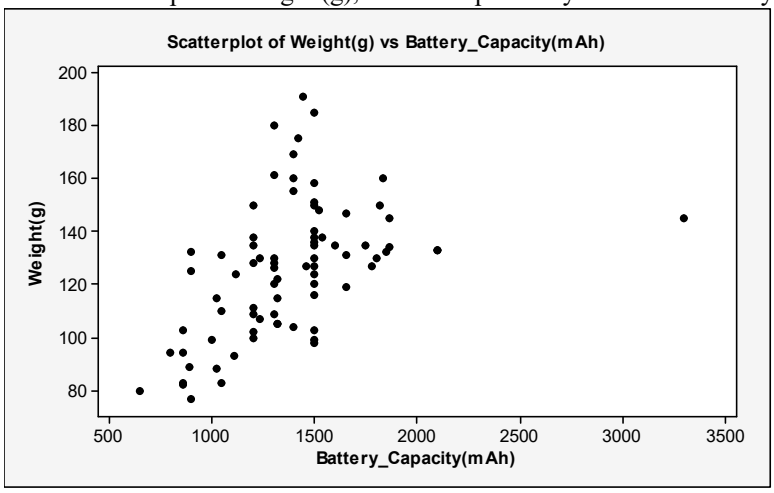
The association appears to be negative.

b) From MINITAB: Predicted GPA = 3.56 – 0.0820Skipped. For a student who does not skip any classes, the predicted GPA is 3.56 (the *y*-intercept). For every class a student skips, we predict their GPA to drop by 0.08 points.

c) The predicted GPA is 3.56 – 0.082(9) = 2.82. The residual is $y - \hat{y} = 2.8 - 2.82 = -0.02$.

12.9 Cell phone specs

a) The response variable is cell phone weight (g), and the explanatory variable is battery size (mAh).



A clear trend is visible in the scatterplot, showing that phones with larger capacity tend to weigh more. One phone (number 70 with battery capacity of 3300) has a much larger battery capacity than all the others, yet its weight is about average, not following this trend. This phone is a clear outlier.

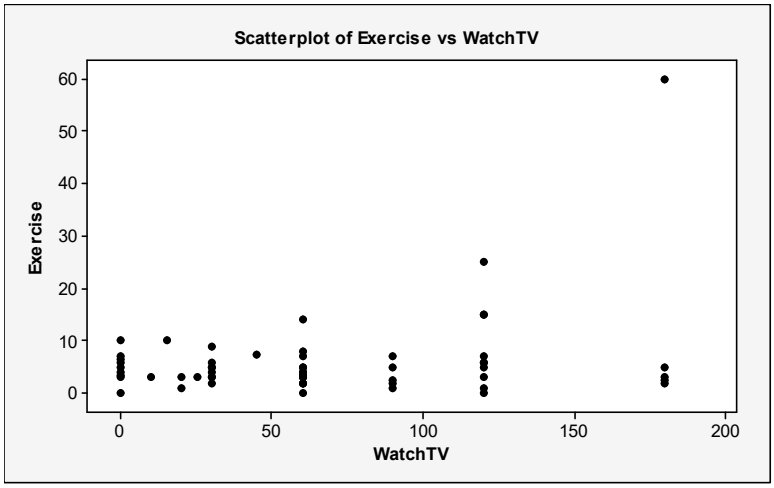
b) The outlier will pull the regression line toward it. Its residual will be negative and very large in absolute value.

c) From MINITAB: Predicted Weight = 66.7 + 0.0436Battery
   (i) Predicted Weight = 66.7 + 0.0436(1000) = 110g
   (ii) Predicted Weight = 66.7 + 0.0436(1500) = 132g

d) For every 100 mAh increase in the capacity of a cell phone's battery, the predicted weight increases by 4.3g.

12.10 Exercise and watching TV

a)



The point with a score of 60 on exercise is an outlier and could make the slope more positive.

b) With the exercise score of 60: Exercise = 2.90 + 0.0462WatchTV
Without the exercise score of 60: Exercise = 4.54 + 0.0075WatchTV
The observation decreased the intercept and increased the slope.

Section 12.2: Inference About Model Parameters and the Association

12.11 *t*-score?

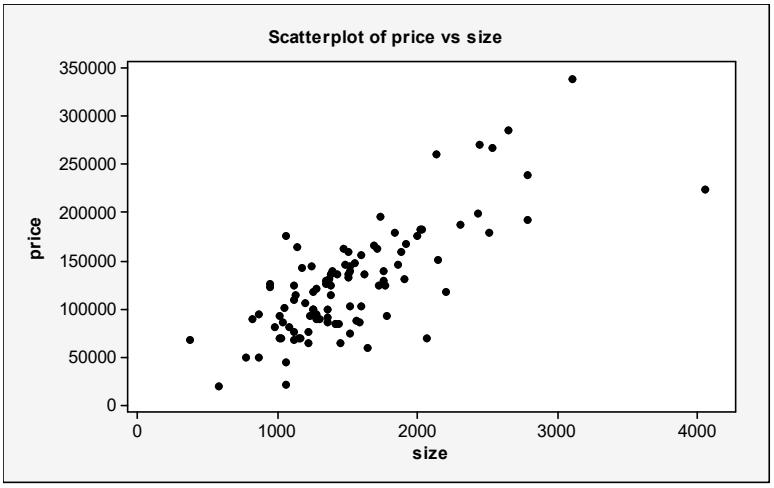
a) $df = n - 2 = 32 - 2 = 30$

b) –1.70 and 1.70

c) We'd use 1.70.

12.12 Predicting house prices

a) (i) Assumptions: Assume randomization, a linear relationship between mean selling price and size of a house in the population, and a normal conditional distribution of price for given sizes, with the same standard deviation.



(ii) Hypotheses: The null hypothesis that the variables are independent is H_0: $\beta = 0$. The two-sided alternative hypothesis of dependence is H_a: $\beta \neq 0$.

(iii) Test statistic: From technology, the test statistic is $t = 11.62$. We also could calculate the test statistic as follows: $t = b/se = 77.008/6.626 = 11.6$.

(iv) P-value: From technology, the P-value is 0.000.

(v) Conclusion: If H_0 were true that the population slope $\beta = 0$, it would be extremely unusual (the probability would be almost 0) to get a sample slope at least as far from 0 as $b = 77.008$. The P-value gives very strong evidence that an association exists between the size and price of houses; this is extremely unlikely to be due to random variation.

b) The 95% confidence interval is $b \pm t_{.025}(se) = 77.008 \pm 1.985(6.626)$, or (64, 90).

c) An increase of \$100 is outside the confidence interval and so is a very implausible value for the population slope.

12.13 Confidence interval for slope

a) On average, the selling price of a house increases between \$64 and \$90 for every one square foot increase in size.

b) On average, the selling price of a house increases between \$6400 and \$9000 for every 100 square feet increase in size.

12.14 House prices in bad part of town

a) The null hypothesis posits that the slope is 0; that is, it hypothesizes that there is no association between selling price and size of house. A data analyst might choose a one-sided alternative hypothesis for this test because the previous analysis showed a positive association for these variables.

b) (i) The test statistic would have to be 1.714 to get a P-value equal to 0.05.

(ii) The test statistic would have to be 2.500 to get a P-value equal to 0.01.

12.15 Strength through leg press

a) (i) Assumptions: Assume randomization; a linear relationship between mean maximum leg press and number of 200-pound leg presses; and a normal conditional distribution of maximum leg press for a given number of 200-pound leg presses with constant standard deviation. These data are not a random sample, so conclusions are highly tentative.

(ii) Hypotheses: The null hypothesis that the variables are independent is H_0: $\beta = 0$. The two-sided alternative hypothesis of dependence is H_a: $\beta \neq 0$.

(iii) Test statistic: From the table, the test statistic is 9.64. We also could calculate the test statistic as follows: $t = b/se = 5.271/0.547 = 9.64$.

12.15 (continued)

(iv) P-value: From software, the P-value is 0.000.

(v) Conclusion: If H_0 were true that the population slope $\beta = 0$, it would be extremely unusual (the probability would be close to 0) to get a sample slope at least as far from 0 as $b = 5.271$. The P-value gives very strong evidence that an association exists between maximum leg press and number of 200-pound leg presses.

b) The 95% confidence interval is $b \pm t_{.025}(se) = 5.2710 \pm 2.004(0.5469)$, or (4.2, 6.4). On average, maximum leg press increases between 4.2 pounds and 6.4 pounds for every additional 200-pound leg press that the athlete can do. The interval gives us a range of plausible values for the increase. The test only tells us there is strong evidence that the increase (slope) is significantly different from 0.

12.16 More boys are bad?

a) The negative slope indicates a negative association between life length and number of sons. Having more sons is bad.

b) (i) Assumptions: Assume randomization, linear trend with normal conditional distribution for y and the same standard deviation at different values of x.

(ii) Hypotheses: The null hypothesis that the variables are independent is H_0: $\beta = 0$. The two-sided alternative hypothesis of dependence is H_a: $\beta \neq 0$.

(iii) Test statistic: $t = b/se = -0.65/0.29 = -2.24$.

(iv) P-value: The P-value is 0.026.

(v) Conclusion: If H_0 were true that the population slope $\beta = 0$, it would be unusual to get a sample slope at least as far from 0 as $b = -0.65$. In fact, the probability would be 0.026. The P-value gives very strong evidence that an association exists between number of sons and life length.

c) The 95% confidence interval is $b \pm t_{.025}(se) = -0.65 \pm 1.966(0.29)$, or (–1.2, –0.1). The plausible values for the true population slope range from –1.2 to –0.1. It is not plausible that the true slope is 0.

12.17 More girls are good?

a) The positive slope indicates a positive association between life length and number of daughters. Having more daughters is good.

b) (i) Assumptions: Assume there was roughly a linear relationship between variables, and that the data were gathered using randomization and that the population y values at each x value follow a normal distribution, with roughly the same standard deviation at each x value.

(ii) Hypotheses: The null hypothesis that the variables are independent is H_0: $\beta = 0$. The two-sided alternative hypothesis of dependence is H_a: $\beta \neq 0$.

(iii) Test statistic: $t = b/se = 0.44/0.29 = 1.52$

(iv) P-value: From technology, the P-value is 0.13.

(v) Conclusion: If H_0 were true that the population slope $\beta = 0$, it would not be very unusual to get a sample slope at least as far from 0 as $b = 0.44$. The probability would be 0.13. It is plausible that there is no association between number of daughters and life length.

c) The 95% confidence interval is $b \pm t_{.025}(se) = 0.44 \pm 1.966(0.29)$, or (–0.1, 1.0). The plausible values for the true population slope range from –0.1 to 1.0. Zero is a plausible value for this slope.

12.18 CI and two-sided tests correspond

We would reject the null hypothesis regarding the association between number of sons and life length, but would not reject the null hypothesis regarding the association between number of daughters and life length. For boys, zero does not fall in the confidence interval, and we reject the null hypothesis. It does seem as if number of sons and life length are related. For girls, zero does fall in the confidence interval, and we do not reject the null hypothesis. It is plausible that number of daughters and life length are not related.

12.19 Investment and rate of interest

a) The mean for rate of interest is 6, and for investment is 7,000. The standard deviation for rate of interest is 2.16, and for investment is 2160.25.

b) $b = r\left(s_y / s_x\right) = 0.857\left(2.16/2.16\right) = 0.857$

$a = \bar{y} - \bar{x}\beta = 7000 - (0.857)(6) = 1857.14$

$\hat{y} = 1857.14 + 0.857x$

c) (i) Assumptions: Assume randomization, linear trend with normal conditional distribution for y and the same standard deviation at different values of x.

(ii) Hypotheses: The null hypothesis that the variables are independent is H_0: $\beta = 0$. The two-sided alternative hypothesis of dependence is H_a: $\beta \neq 0$.

(iii) Test statistic: $t = b/se$ = 857.14/364.21 = 2.35

(iv) P-value: From technology, the P-value is 0.14.

(v) Conclusion: If H_0 were true that the population slope $\beta = 0$, it would not be very unusual to get a sample slope at least as far from 0 as b = 857.14. The probability would be 0.14. The P-value is not below the significance level of 0.05, and, therefore, we cannot reject the null hypothesis. It is plausible that there is no association between rate of interest and investment.

12.20 GPA and study time–revisited

(i) Assumptions: Assume randomization, linear trend with normal conditional distribution for y and the same standard deviation at different values of x.

(ii) Hypotheses: The null hypothesis that the variables are independent is H_0: $\beta = 0$. The one-sided alternative hypothesis of dependence is H_a: $\beta > 0$.

(iii) Test statistic: From technology, t = 3.38

(iv) P-value: From technology, the P-value is 0.0075.

(v) Conclusion: If H_0 were true, the population slope $\beta = 0$, the probability of obtaining a sample slope at least as large as b = 0.044 would be 0.0075. There is strong evidence of an association between GPA and study time.

12.21 GPA and skipping class–revisited

The 90% confidence interval is $b \pm t_{.05}(se) = -0.082 \pm 1.94(0.016)$, or (–0.11, –0.05). We are 90% confident that the population slope β falls between –0.11 and –0.05. On average, GPA decreases by between 0.11 to 0.05 points for every additional class that is skipped.

12.22 Battery capacity

a) (i) Assumptions: Assume randomization; a linear relationship between mean weight of cell phone and the capacity of its battery with a normal conditional distribution of weight of phone for a given battery capacity with constant standard deviation.

(ii) Hypotheses: The null hypothesis that the variables are independent is H_0: $\beta = 0$. The two-sided alternative hypothesis of dependence is H_a: $\beta \neq 0$.

(iii) Test statistic: $t = b/se$ = 0.0436/0.00806 = 5.41

(iv) P-value: From technology, the P-value is 0.000.

(v) Conclusion: If H_0 were true that the population slope $\beta = 0$, it would be very unusual (the probability would be almost 0) to get a sample slope at least as far from 0 as b = 0.0436. The P-value is beyond the significance level of 0.05, and we can reject the null hypothesis. We have very strong evidence that an association exists between a phone's battery capacity and its weight.

b) The 95% confidence interval is $b \pm t_{.025}(se) = 0.0436 \pm 1.9917(0.00806)$, or (0.028, 0.060). With 95% confidence, we predict that an increase of 1000 mAh in the capacity of a phone's battery increases the weight of the phone by between 28 and 60 grams. Because 0 is not contained in this interval, the slope is significantly different from 0; i.e., an association exists. This is consistent with the result of the hypothesis test.

Section 12.3: Describing the Strength of Association

12.23 Euros and thousands of euros

The slope when investment is in euros is 1.63/1000 = 0.00163.

12.24 When can you compare slopes?

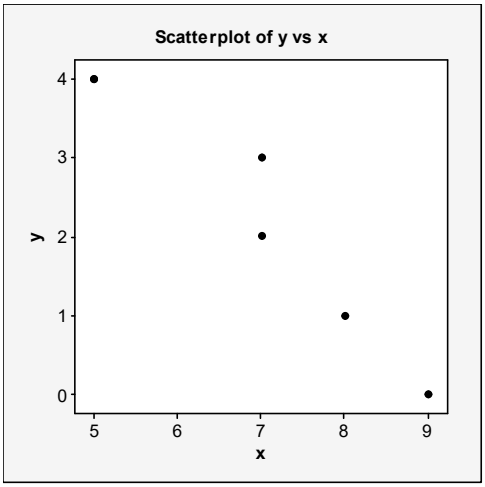
a) For a £1000 increase in per capita GDP, the predicted consumption expenditure increases by 0.52 × 1000 = £ 520, and the predicted investment expenditure increases by 0.27 × 1000 = £ 270.

b) Because the slope for GDP predicting consumption expenditure is larger than the slope for GDP predicting investment expenditure, an increase in GDP would have a slightly greater impact on consumption expenditure than on investment expenditure.

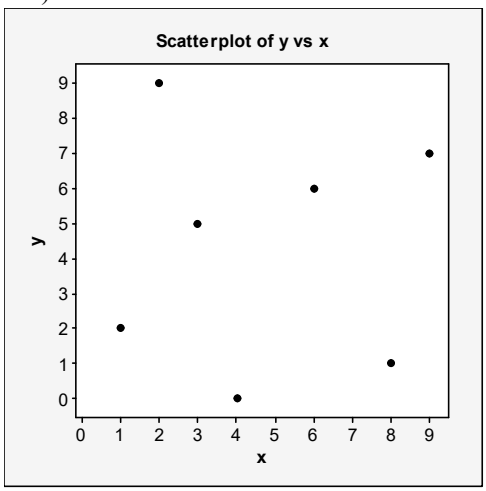
12.25 Sketch scatterplot

There are a number of possible scatterplots that would fit these scenarios. Here are possibilities.

a)



b)



12.26 Sit-ups and the 40-yard dash

a) (i) $\hat{y} = 6.707 - 0.02435(10) = 6.46$

(ii) $\hat{y} = 6.7065 - 0.024346(40) = 5.73$

For an increase of 30 in the number of sit-ups, the predicted change in the time for the 40-yard dash is (–0.0243)(30) = –0.73, That is, an athlete who can do 30 more sit-ups is predicted to be 6.46 – 5.73 = 0.73 seconds faster.

b) The time of the 40-yard dash will decrease by 0.46 standard deviations for a standard deviation increase in number of sit-ups.

c) Because the slope is negative, the correlation also will be negative.

$r = b\left(s_x / s_y\right) = -0.024346\left(6.887/0.365\right) = -0.46$

d) The predicted time difference is 0.46 standard deviations of the 40-yard dash time, which is 0.46(0.365) = 0.17 seconds.

12.27 Body fat

a) There is a strong, positive linear association between weight and percent body fat.

b) $r^2 = 0.78$; Using the regression equation with weight to predict percent body fat instead of predicting it with the sample mean results in a 78% reduction in the overall prediction error (the sum of the squared errors).

c) No, the correlation r does not depend on the units.

12.28 Verbal and math GRE scores

a) $\hat{y} = 30 + 0.80(150)$; Generally, at the x-value equal to its mean, the predicted value of y is equal to *its* mean.

12.28 (continued)

b) We can find the correlation as follows: $r = b\left(s_x / s_y\right) = 0.80\left(6.5/6.5\right) = 0.80.$

When the x and y variables have the same variability, the correlation equals slope.

c) $r^2 = (0.8)(0.8) = 0.64$; The sum of squared errors is 64% less when we use the regression equation to predict y instead of the mean of y.

12.29 GRE score regression toward mean

a) $\hat{y} = 30 + 0.80(170) = 166$

b) The predicted y value will be 0.8 standard deviations above the mean, for every one standard deviation above the mean that x is. Here, $x = 170$ is approximately three standard deviations above the mean; so the predicted y value is approximately $0.8(3) = 2.4$ standard deviations above the mean.

12.30 GPAs and TV watching

a) The correlation of –0.35 indicates that there is a negative relation between the two variables.
$r^2 = 0.13$ indicates:
(i) only a small reduction of 13% in the overall prediction error when using the regression equation with time watching TV to predict GPA rather than using the mean GPA.
(ii) 13% of the variability observed in college GPA can be explained by its relationship with time watching TV.

b) (i) The student would be below the mean because the correlation and, hence, slope is negative.
(ii) $2(0.35) = 0.70$ standard deviations below the mean of college GPA. With regression to the mean, the predicted college GPA is closer to the mean GPA than time watching TV is to its mean.

12.31 GPA and study time

a) $r = 0.81$, there is a fairly strong, positive, linear association between GPA and study time. The more time a student studies, the higher their GPA is likely to be.

b) $r^2 = (0.81)^2 = 0.656$
(i) The overall prediction error when using the regression equation with study time to predict GPA is 66% smaller compared to using the sample mean GPA.
(ii) 66% of the variability observed in college GPA can be explained by the linear relationship with study time.

12.32 Placebo helps cholesterol?

a) Their mean cholesterol reading at time 2 should be $200 + 100(0.7) = 270$.

b) This does not suggest that placebo is an effective treatment; this decrease could occur merely because of regression to the mean. Subjects who are relatively high at one time will, on the average, be lower at a later time. So, if a study gives placebo to people with relatively high cholesterol (that is, in the right-hand tail of the blood cholesterol distribution), on the average we expect their values three months later to be lower.

12.33 Was the advertising strategy helpful?

The explanatory variable is mid-month sales and the response variable is end-month sales. We cannot conclude that the advertising strategy was successful. These bread types were very low selling to start with–two full standard deviations below the mean. This increase could have occurred because of regression to the mean. Subjects who are relatively low at one time will, on the average, be higher at a later time. So, for the bread types with relatively low sales (that is, in the left-hand tail of the distribution of mid-month sales), on average we expect their values on the end-month sales to be higher.

12.34 What's wrong with your stock fund?

This might be due to regression to the mean. Stocks that are relatively high one year will, on the average, be lower at a later time.

12.35 Golf regression

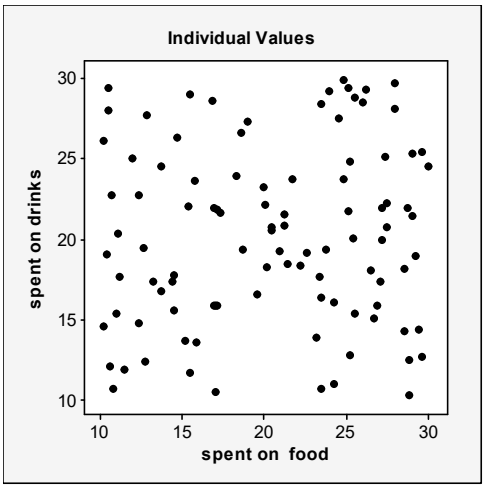
Regression to the mean suggests that we can expect that the five leaders would, on average, have higher scores the second time around.

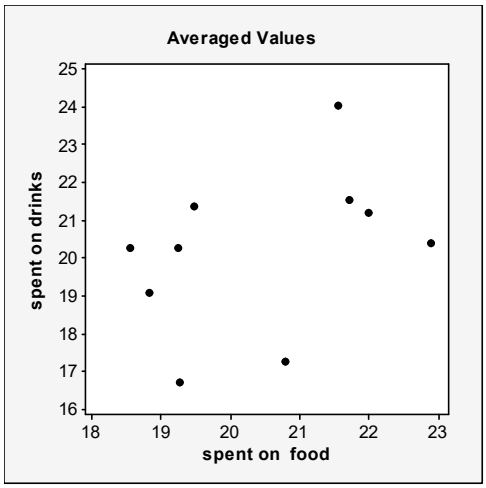
12.36 Car weight and mileage

There is a 75% reduction in error in predicting a car's mileage based on knowing the weight, compared to predicting by the mean mileage. This relatively large value means that we can predict a car's mileage quite well if we know its weight.

12.37 Food and drink sales

The scatterplots below demonstrate that at the individual level, the correlation is weaker. The first has a correlation of 0.087 (here, it was constructed using 100 pairs of dollar amounts, representing 10 transactions per day, for individuals, rather than 2500 individuals). The second has a correlation of 0.390 and is constructed the means of the dollar amounts per day). There is a great deal of variability in amounts spent for individuals, but not much variability in mean amounts spent for days. The summary values for days fall closer to a straight line. (Note that the *x* and *y* axes of the second scatterplot have a restricted range compared to the first.)



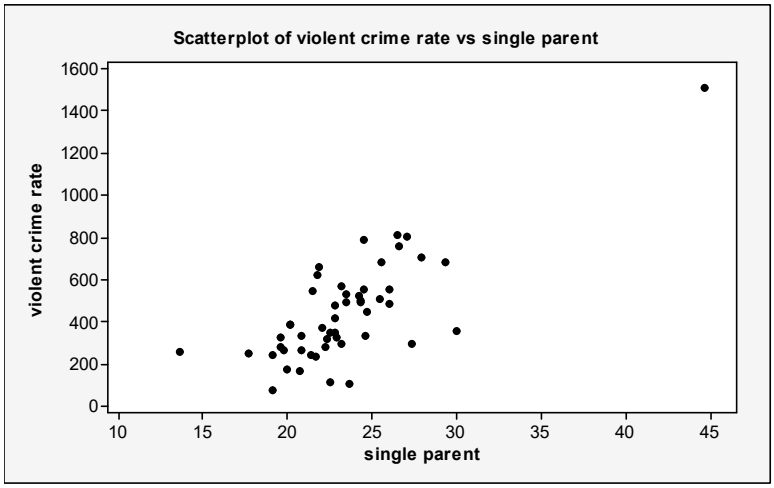


12.38 Yale and UConn

The correlation between high-school and college GPA would likely be higher at the University of Connecticut than at Yale. Yale would have a restricted range of high school GPA values, with nearly all of its students clustered very close to the top. UConn would have a wider range of high-school GPAs. The correlation tends to be smaller when we sample only a restricted range of *x* values than when we use the entire range.

12.39 Violent crime and single-parent families

a)



The scatterplot shows a likely positive correlation between these variables, with one extreme outlier.

12.39 (continued)

b) Technology confirms the change from 0.77 to 0.59. The correlation drops so dramatically because it depends in magnitude on the variability in scores. Without the outlier, the scores concentrate more narrowly at the lower end of the scale, leading to a weaker correlation. By contrast, when all the scores, including the outlier, are included, there is a wider range of values, and we would likely see a stronger correlation for them. In general, the correlation tends to be smaller in absolute value when we sample only a restricted range of x values than when we use the entire range.

💻12.40 Correlations for the strong and for the weak

a) From technology, the correlation between number of 60-pound bench presses before fatigue (BP60) and maximum bench press (maxBP) is 0.80 for females and 0.91 for male. Although both are strong, positive correlations, the correlation for males is stronger.

b) (i) Using only the x values below the median of 10 for females, the correlation is only 0.47. Using only the x values below the median of 17 for males, the correlation is 0.93.

(ii) Using only the x values above the median of 10 for females, the correlation is 0.67. Using only the x values above the median of 17 for males, the correlation is 0.57. They are so different because the correlation usually is smaller in absolute value when the range of predictor values is restricted.

Section 12.4: How the Data Vary Around the Regression Line

12.41 Poor predicted sales

a) The entry under "Marketing Spend", 150.00, is the marketing spend in thousands of pounds in month 11. The predicted sales in thousands of pounds for this month, however, 1895.20, are in the column "Fit". In the residual column, we see the difference between the actual and predicted sales, 29.10. Finally, in the column titled "StdResid", we see the standardized residual, 2.56 which is the residual divided by the standard error that describes the sampling variability of the residuals; it does not depend on the units used to measure the variable. The "R" designates a standardized residual above 2.0.

b) We would expect 5% of standardized residual to have an absolute value above 2.0. Thus, it is not surprising that three would have an absolute value above 2.0.

12.42 Loves TV and exercise

a) The residual, 48.8, is the difference between the actual reported minutes of exercise, and the predicted reported minutes of exercise for this student. To find the predicted value: $48.8 = 60 - \hat{y} \Rightarrow$

$\hat{y} = 60 - 8.8 \Rightarrow \hat{y} = 11.2.$

b) The standardized residual is the residual divided by standard error. It is the number of standard errors the residual falls from 0 (the mean or the predicted score). This student falls 6.41 standard errors above what would be predicted for the number of minutes per day she/he watches TV.

12.43 Bench press residuals

a) This figure provides information about the distribution of standardized residuals, and hence the conditional distribution of maximum bench press.

b) The conditional distribution in (a) seem to be approximately normal.

12.44 Predicting house prices

a) Using the residual *df* of 98, $98 = n - 2 \Rightarrow n + 89 + 2 = 100$. The sample size was 100.

b) The sample predicted mean selling price was $\hat{y} = 9.2 + 77(1.53) = 127.010,$ or \$127,010.

c) The estimated residual standard deviation of y is the square root of MS Error = 1349. The square root of 1349 is 36.7.

d) The prediction interval is: $\hat{y} \pm 2s = 127.010 \pm 2(36.729),$ or (53.6, 200.5).

12.45 Predicting annual salary

a) The value under "Fit", 148.0, is the predicted annual salary in thousands of dollars for those residents who had 15 years of education.

b) The 95% confidence interval of (129, 162) is the range of plausible values for the population mean of annual salary in thousands of dollars for those residents who had 15 years of education.

12.45 (continued)

c) The 95% prediction interval of (70, 210) is the range of plausible values for individual observations (annual salary in thousands of dollars) for those residents who had 15 years of education.

12.46 CI versus PI

A 95% prediction interval predicts where we can expect individual observations to fall for a given value of *x*. A 95% confidence interval provides a range of plausible values for the population mean for *y* at a given value of *x*. We would expect the PI to be wider than the CI because we can predict the population mean more precisely than we can predict an individual value.

12.47 ANOVA table for leg press

a) The residual standard deviation is the square root of Mean Square Error. The square root of 1303.72 is 36.1. This is the estimated standard deviation of maximum leg presses for female athletes who can do a fixed number of 200-pound leg presses.

b) The standard deviation is 36.1; with $x = 22$, $\hat{y} = 233.89 + 5.27(22) = 349.92$, so the 95% prediction interval for female athletes is $\hat{y} \pm 2s = 349.83 \pm 2(36.1)$, or (277.6, 422.0).

12.48 Predicting leg press

a) MINITAB got the "Fit" of 365.66 from: $\hat{y} = 233.89 + 5.27x = 233.89 + 5.27(25) = 365.6$ (the difference between 365.6 and 365.7 is due to differences in rounding).

b) The 95% confidence interval is approximately $\hat{y} \pm t_{.025}(se) = 356.66 \pm 2.004(5.02)$, or (355.6, 375.7); these are the plausible values for the population mean of *y* values at $x = 25$.

c) With 95% confidence, we predict that the maximum leg press for an individual athlete who can leg press 200 pounds 25 times is between 293 pounds and 439 pounds.

12.49 Variability and *F*

a) The Total SS is the sum of the regression SS and the residual SS. The residual SS represents the error in using the regression line to predict *y*. The regression SS summarizes how much less error there is in predicting *y* using the regression line compared to using $\bar{y}$.

b) The sum of squares around the mean divided by $n - 1$ is 192,787/56 = 3442.6, and its square root is 58.7. This estimates the overall standard deviation of *y* values, whereas the residual *s* estimates the standard deviation of *y* values at a fixed value of *x*.

c) The *F* test statistic is 92.87; its square root is the *t*-statistic of 9.64.

12.50 Assumption violated

a) Each student's scatterplot will be different, but should show less variability on *y* when *x* is low, and more variability on *y* when *x* is large. This reflects the fact that people with lower incomes are limited in how much they can give to charity; the amounts they give to charity will be, by necessity, low. Those with higher incomes have the ability to donate much more to charity, although many will choose to give low or moderate amounts; thus, the variability in amount donated to charity will be much higher among those with high incomes.

b) A 95% prediction interval would not work well at very small or very large *x* values because the prediction interval will have similar widths for each *x* value. Thus, it will predict a wider range than exists for those with very low incomes, and a narrower range than exists for those with very high incomes.

12.51 Understanding an ANOVA table

a) The MS values are calculated by dividing SS by DF. The top MS will be 500,000/1 = 500,000, and the bottom MS value will be 200,000/35 = 8571.42. *F* is the ratio of the two mean squares, 500,000/8571.42 = 58.34.

b) The *F* test statistic is an alternative test statistic for testing H_0: $\beta = 0$ against H_a: $\beta \neq 0$.

12.52 Predicting cell phone weight

a) The necessary assumptions are a linear relationship between the mean weight of cell phone and the capacity of its battery; a random sample of cell phones; and a normal conditional distribution of weight of phone for given battery capacity with constant standard deviations. The 95% confidence interval is (127, 137). With 95% confidence, the mean weight of a phone with 1500 mAh of battery capacity falls between 127 grams and 137 grams.

b) The 95% prediction interval is (89, 175). With 95% confidence, the weight of a cell phone with 1500 mAh battery capacity falls between 89 grams and 175 grams.

c) A prediction interval predicts where individual observations fall for a given value of x. A confidence interval provides a range of plausible values for the population mean for y at a given value of x.

12.53 Cell phone ANOVA

From Minitab:

```
Analysis of Variance

Source          DF     SS     MS      F      P
Regression       1  13682  13682  29.24  0.000
Residual Error  76  35560    468
Total           77  49242
```

a) Total SS is the sum of the Residual SS and Regression SS, so Total SS = 49242 = 13682 + 35560. The total SS is a measure of the overall variability in y (phone weight) but also measures the overall error when using $\bar{y}$ to predict y. The residual SS measures the overall prediction error when using $\hat{y}$ to predict y. Their difference equals the regression SS, which tells us how much the prediction error decreases when using $\hat{y}$ instead of $\bar{y}$ to predict y.

b) $s = 21.6$, which is the square root of 468, the mean square error. It estimates the standard deviation of cell phone weight at given battery capacity value and describes a typical value of the residual.

c) $s_y = 25.3$; This describes the variability in cell phone weight over the entire range of battery capacity values, not just the variability in cell phone weight at a particular capacity value.

Section 12.5: Exponential Regression: A Model for Nonlinearity

12.54 Savings grow exponentially

a) $y = \alpha\beta^x = 100(1.10)^1 = 110$

b) $y = \alpha\beta^x = 100(1.10)^5 = 161.05$

c) $y = \alpha\beta^x = 100(1.10)^x$

d) The first year after which you'll have more than \$200 is the 8th: $y = \alpha\beta^x = (100)(1.10)^8 = 214.36$.

12.55 Growth by year versus decade

a) $(1.036)^{20} = 2.0$

b) $(1.05)^{20} = 2.65$. The female population would double in two decades if there is an *'addition'* of 5% per year, but the effect here is multiplicative, not additive.

12.56 Moore's law today

a) $\hat{y} = 151.61(1.191)^{2015-1994} = 5955$

b) The number of components is predicted to increase by a factor of 1.191, which is the estimated value for β in the exponential regression model. This corresponds to a 19.1% increase year over year.

c) A linear trend on the log scale and the high correlation indicate that the regression model is appropriate.

12.57 U.S. population growth

a) $\hat{y} = 81.14(1.339)^0 = 81.14$ million; $\hat{y} = 81.14(1.339)^{11} = 323.26$ million

b) 1.1339 is the multiplicative effect on $\hat{y}$ for a one-unit increase in x.

c) This suggests a very good fit of data to model. The high correlation indicates a linear relation between the log of the y values and the x values.

12.58 Future shock

a) $(1.15)^5 = 2$; The population size after five decades is predicted to be 2.0 times the original population size.

b) $(1.15)^{10} = 4.0$; The population size after ten decades is predicted to be 4.0 times the original population size.

c) $(1.15)^{20} = 16.4$; The population size after twenty decades is predicted to be 16.4 times the original population size.

12.59 Age and death rate

a) 0.34 is the prediction for $\hat{y}$ when $x = 0$; 1.081 is the multiplicative effect on $\hat{y}$ for a one-unit increase in x.

b) (i) $0.34(1.081)^{25} = 2.4$

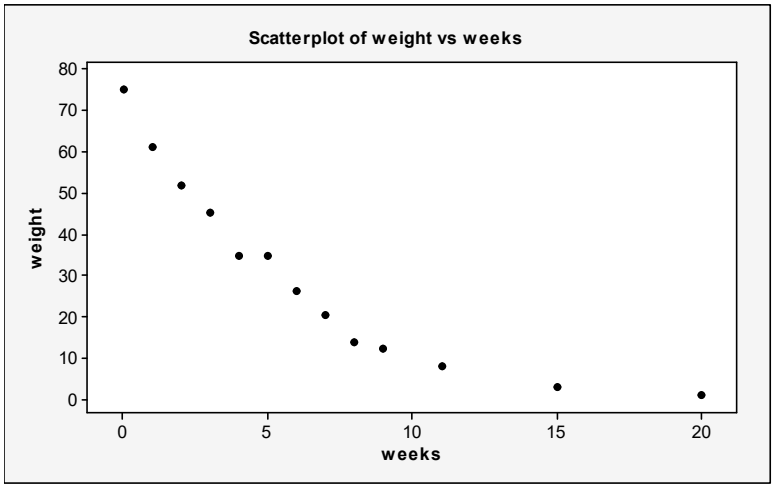
(ii) $0.34(1.081)^{55} = 24.7$

(iii) $0.34(1.081)^{80} = 172.8$

c) $(1.081)^9 = 2.01$; the predicted death rate doubles every nine years.

12.60 Leaf litter decay

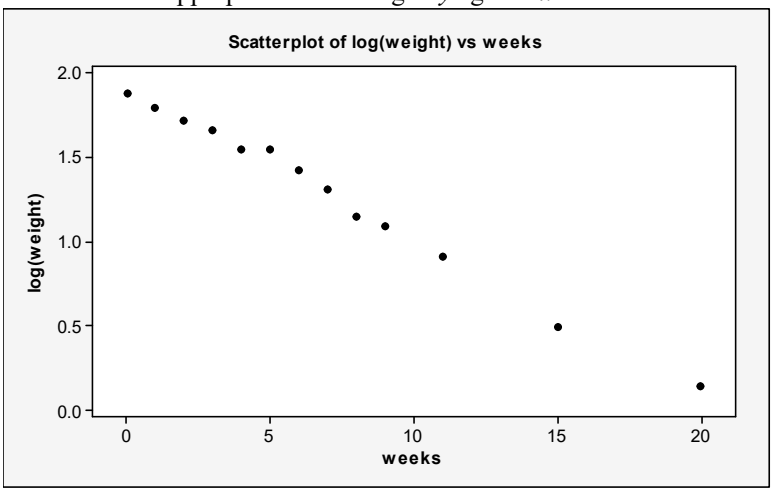
a) A straight-line model is inappropriate because the scatterplot shows that the relation between the two variables is curvilinear.



b) From technology: weight = 55.0 – 3.59weeks
The predicted weight after 20 weeks is 55.0 – 3.59(20) = –16.8; this does not make sense because a weight cannot be negative.

12.60 (continued)

c) A straight-line model seems appropriate for the log of y against x.



d) (i) $\hat{y} = 80.6(0.813)^0 = 80.6$

(ii) $\hat{y} = 80.6(0.813)^{20} = 1.3$

e) The coefficient 0.813 indicates that the predicted weight multiplies by 0.813 each week.

12.61 More leaf litter

a) The exponential model is more appropriate because the log of y values and the x values have a relation that is closer to a straight line than is the relation of x and y.

b) $(0.813)^3 = 0.54$, $(0.813)^4 = 0.44$; The half life is predicted to be between three and four weeks.

$(0.813)^{3.34} = 0.5$ gives a more specific predicted half-life of 3.34 weeks.

Chapter Problems: Practicing the Basics

12.62 Academic performance and participation in extracurricular activities

a) At fixed values of x, there is variability in the values of y so we cannot specify individual y values using x but we can try to specify the mean of those values and how that mean changes as x changes.

b) Because y values vary at a fixed value of x, the model has a σ parameter to describe the variability of the conditional distribution of those y values at each fixed x.

12.63 Theory exam–practical exam correlation

a) $b = r\left(s_y / s_x\right) = 0.60\left(80/60\right) = 0.8$

b) $a = \overline{y} - b\overline{x} = 360 - (0.8)(270) = 144$, so $\hat{y} = 144 + 0.8x$.

c) Treating theory exam score as x and practical exam score as y: $b = r\left(s_y / s_x\right) = 0.60\left(60/80\right) = 0.45$,

$a = \overline{y} - b\overline{x} = 270 - (0.45)(360) = 108$, so $\hat{y} = 108 + 0.45x$.

12.64 Stem cells

a) This is the interpretation of $r^2 = (0.804)^2 = 0.65$, or 65%.

b) This refers to the proportional reduction in overall prediction error, which is $r^2 = 0.65$.

12.65 Tall people

The response variable is height of children and the explanatory variable is height of parents. Very tall parents would tend to have children who are tall, but not as tall as they are. The prediction equation is based on correlation. Because the correlation between two variables is never greater than the absolute value of 1, a y value tends to be not so far from its mean as the x value is from its mean. This is called regression toward the mean.

12.66 Income and education in Florida

a) Technology gives us a correlation of 0.79.
   (i) This is a positive correlation. As one variable increases, so does the other tend to increase.
   (ii) The magnitude of the correlation is close to 1, which is a strong correlation. These two variables are highly related.

b) A county that is one standard deviation above the mean on x is predicted to be 0.79 standard deviations above the mean on y (see calculations below). This exemplifies regression to the mean; the y value is predicted to be fewer standard deviations from its mean than x is from its mean.
The regression equation is: predicted income = –4.6 + 0.419education. Since the mean for education is 69.49 and the standard deviation is 8.86, one standard deviation above the mean, therefore, is 78.35. We would predict an income of –4.6 + 0.419(78.35) = 28.229 (income is in thousands of dollars; hence, $28,229). The mean for income is 24.51 and its standard deviation is 4.69. $28,229 is 0.79 standard deviations above the mean.

12.67 Bedroom residuals

a) $\hat{y} = 33,778 + 31,077(3) = 127,009$; The residual is 338,000 – 127,009 = 210,991. This house sold for $210,991 more than predicted.

b) A residual divided by its *se* is a standardized residual, which measures the number of standard errors that a residual falls from 0. It helps us identify unusual observations. The standardized residual of 4.02 indicates that this observation is 4.02 standard errors higher than predicted.

12.68 Bedrooms affect price?

a) The regression parameter is the population slope that is estimated by the slope of the sample, 31,077. It represents the change in predicted selling price when the number of bedrooms increases by 1.

b) The 95% confidence interval is $b \pm t_{.025}(se) = 31,077 \pm 1.984(8049)$, or (15,108, 47,046).

c) A difference of two bedrooms would double both ends of the confidence interval because the slope is for an increase of 1, and here we have an increase of 2. It would now be: (30,216, 94,092). The difference between the means is double the slope.

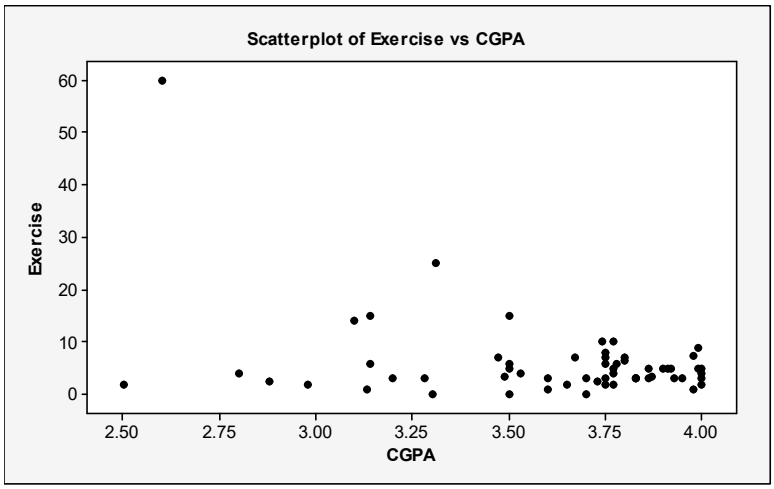
12.69 Types of variability

a) The residual standard deviation of y refers to the variability of the y values at a particular x value, whereas the standard deviation refers to the variability of all of the y values.

b) The fact that they're not very different indicates that the number of bedrooms is not strongly associated with selling price. The variability of y values at a given x is about the same as the variability of all of the y observations. We can see this by considering the r^2 of 0.13. The error using $\hat{y}$ to predict y is only 13% smaller than the error using $\bar{y}$ to predict y.

12.70 Exercise and college GPA

a)



12.70 (continued)

The observation with x of 2.60 and y of 60 is an outlier. It likely makes the slope and correlation more negative than they would be otherwise.

b) From technology, the regression equation is: Exercise = 35.7 – 8.24CGPA. The standardized residual of for the point (2.60, 60) is 6.35. The exercise score for this individual is 6.35 standard errors higher than predicted.

c) The new regression equation is: Exercise = 9.20 – 1.15CGPA. As expected, the outlier made the slope more negative than without the outlier.

12.71 Bench press predicting leg press

a) The 95% confidence interval provides a range of plausible values for the population mean of y when $x = 80$. The plausible values range from 338 to 365 for the *mean* of y values for *all* female high school athletes having $x = 80$.

b) The prediction interval provides a range of predicted y values for an individual observation when $x = 80$. For *all* female high school athletes with a maximum bench press of 80, we predict that 95% of them have maximum leg press between about 248 and 455 pounds. The 95% prediction interval is for a single observation y, whereas the confidence interval is for the *mean* of y.

12.72 Leg press ANOVA

a) The estimated standard deviation of the maximum leg press values of those with a maximum bench press of 80 is the residual standard deviation or the square root of the residual *MS*. The square root of 2624 is 51.2.

b) The approximate 95% prediction interval is: $\hat{y} \pm 2s = 351.2 \pm 2(51.2)$, or (248.8, 453.7).

12.73 Savings growth

a) $2000(2)^6 = €128,000$

b) $2000(2)^9 = €1,024,000$

c) The equation based on decade instead of year is $y = 2000(2)^x$.

12.74 Florida population

a) The approximate rate of growth per year is 3.6%.

b) (i) The predicted population size in 1830 is: $\hat{y} = 46(1.036)^0 = 46$ thousand

(ii) In 2000: $\hat{y} = 46(1.036)^{170} = 18,790$ thousand (i.e., almost 19 million)

c) For 2100: $\hat{y} = 46(1.036)^{270} = 645,493$ thousand. The same formula will not likely hold up between 2000 and 2100. Eventually, we would expect population size within a constrained area to level off.

12.75 World population growth

a) In 1900: $\hat{y} = 1.424(1.014)^0 = 1.42$ billion

In 2000: $\hat{y} = 1.424(1.014)^{110} = 6.57$ billion

b) The fit of the model corresponds to a rate of growth of 1.4% per year because multiplying by 1.014 adds an additional 1.4% each year.

c) (i) The predicted population size doubles after 50 years because $(1.014)^{50} = 2.0$, the number by which we'd multiply the original population size.

(ii) It quadruples after 100 years since $(1.014)^{100} = 4.0$.

d) The exponential regression model is more appropriate for these data because the log of the population size and the year number are more highly correlated ($r = 0.99$) than are the population size and the year number.

12.76 Match the scatterplot

a) These values correspond to scatterplot 1. The point to the upper right of the data decreases the magnitude of the correlation and slope; it is an outlier with respect to the fitted line and has a large residual.

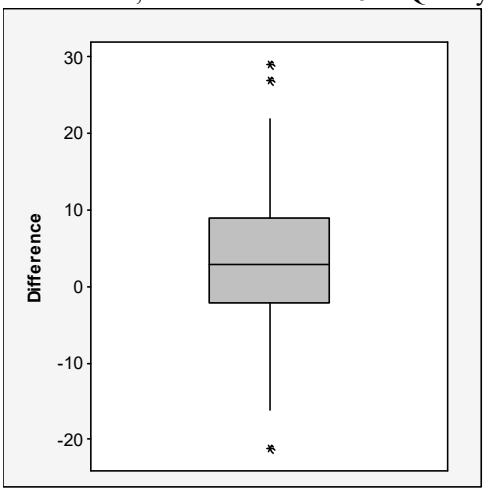
b) These values correspond to scatterplot 3. The point to the left of the data increases slope in magnitude and correlation; this point is an outlier in the *x* direction.

c) These values correspond to scatterplot 2. There are no outliers.

Chapter Problems: Concepts and Investigations

12.77 Softball data

a) The 3 outlying points represent outliers; values more than 1.5 × IQR beyond either Q1 or Q3.



b) From Minitab: Difference = –9.125 + 1.178Run
The difference is positive when –9.125 + 1.178RUNS > 0, which is equivalent to RUNS > 9.125/1.178, or RUNS > 7.7; thus, the team scores more runs than their opponents with eight or more runs.

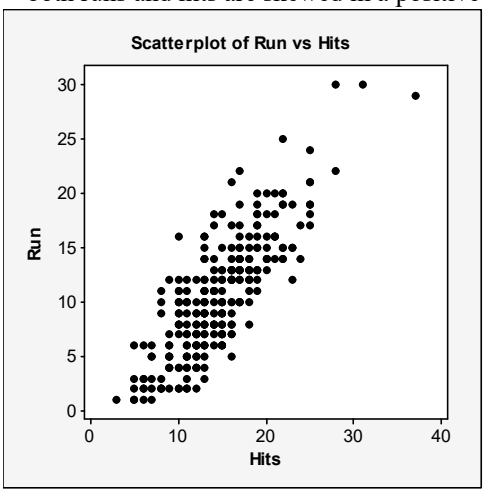
c) Runs, hits, and difference are positively associated with one another. Errors are negatively associated with those three variables.
From Minitab:

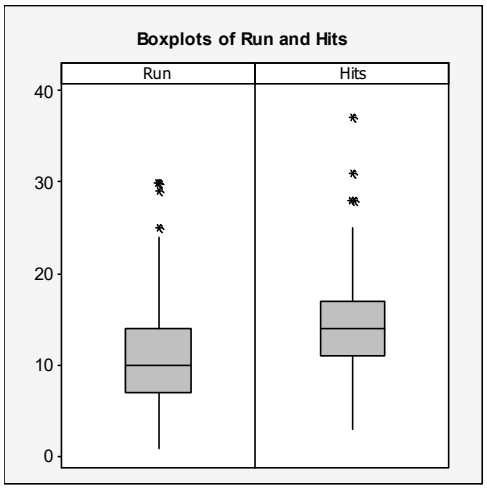
	Run	Hits	Errors
Hits	0.819		
Errors	-0.259	-0.154	
Difference	0.818	0.657	-0.501

d) From technology, the P-value of 0.000 for testing that the slope equals 0 provides extremely strong evidence that DIFF and RUNS are associated.

12.78 Runs and hits

a) A scatterplot indicates that a straight-line regression model seems appropriate. Box plots indicate that both runs and hits are skewed in a positive direction.





b) Reports will vary, but should include the following descriptive statistics about the individual variables.
From Minitab:

```
Variable    N  N*    Mean  SE Mean  StDev  Minimum      Q1
    Median      Q3
Run       277   0  10.596    0.325  5.414    1.000   7.000
Hits      277   0  14.643    0.298  4.958    3.000  11.000

Variable    Median      Q3  Maximum
Run         10.000  14.000   30.000
Hits        14.000  17.000   37.000
```

Reports will vary, but should include the following statistics about their relationship.
From Minitab:

```
Pearson correlation of Run and Hits = 0.819; P-Value = 0.000

The regression equation is Run = - 2.49 + 0.894 Hits
Predictor     Coef  SE Coef      T      P
Constant   -2.4940   0.5845  -4.27  0.000
Hits       0.89395  0.03781  23.64  0.000
S = 3.11481   R-Sq = 67.0%   R-Sq(adj) = 66.9%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        1  5422.6  5422.6  558.92  0.000
Residual Error  275  2668.1     9.7
Total           276  8090.7
```

12.78 (continued)

c) The report should include a discussion of unusual observations from the report of standardized residuals that are greater than 2 in absolute value (listed below).

From Minitab:

Obs	Hits	Run	Fit	SE Fit	Residual	St Resid
32	14.0	18.000	10.021	0.189	7.979	2.57R
53	8.0	11.000	4.658	0.313	6.342	2.05R
82	12.0	2.000	8.233	0.212	-6.233	-2.01R
119	14.0	17.000	10.021	0.189	6.979	2.24R
120	15.0	18.000	10.915	0.188	7.085	2.28R
123	22.0	25.000	17.173	0.335	7.827	2.53R
156	16.0	5.000	11.809	0.194	-6.809	-2.19R
171	13.0	16.000	9.127	0.197	6.873	2.21R
200	37.0	29.000	30.582	0.866	-1.582	-0.53 X
206	28.0	30.000	22.536	0.539	7.464	2.43RX
211	13.0	16.000	9.127	0.197	6.873	2.21R
230	17.0	19.000	12.703	0.207	6.297	2.03R
233	9.0	12.000	5.551	0.284	6.449	2.08R
246	17.0	22.000	12.703	0.207	9.297	2.99R
258	16.0	21.000	11.809	0.194	9.191	2.96R
263	10.0	16.000	6.445	0.257	9.555	3.08R

d) See Minitab output in (b).

12.79 GPA and TV watching

The two-page report will be different for each student, but should interpret results from the following output from technology.

From Minitab:

```
Pearson correlation of high_sch_GPA and TV = -0.268

The regression equation is high_sch_GPA = 3.44 - 0.0183 TV
Predictor       Coef   SE Coef      T      P
Constant     3.44135   0.08534  40.32  0.000
TV         -0.018305  0.008658  -2.11  0.039
S = 0.446707   R-Sq = 7.2%   R-Sq(adj) = 5.6%

Analysis of Variance
Source          DF       SS      MS     F      P
Regression       1   0.8921  0.8921  4.47  0.039
Residual Error  58  11.5737  0.1995
Total           59  12.4658
```

12.80 Female athletes' speed

The two-page report will be different for each student, but should interpret results from the following output from technology.

From Minitab:

```
Pearson correlation of 40-YD (sec) and WT (lbs) = 0.367

The regression equation is 40-YD (sec) = 5.29 + 0.00536 WT
(lbs)
Predictor      Coef   SE Coef      T      P
Constant     5.2920    0.2611  20.27  0.000
WT (lbs)   0.005363  0.001831   2.93  0.005
S = 0.342616   R-Sq = 13.5%   R-Sq(adj) = 11.9%
```

12.80 (continued)

From Minitab:

```
Analysis of Variance
Source          DF      SS      MS     F      P
Regression       1  1.0068  1.0068  8.58  0.005
Residual Error  55  6.4562  0.1174
Total           56  7.4630
```

12.81 Football point spreads

a) If there is no bias in the Las Vegas predictions, the predictions should exactly match the observations. In other words, $y = x$, which translates to the true y-intercept equaling 0 and the true slope equaling 1.

b) No. Based on the results in the table, the P-value for testing that the true y-intercept is 0 is quite large so that we are unable to conclude that the y-intercept differs from 0. The P-value for testing that the slope is equal to 0 is approximately 0 so that we reject the null hypothesis of the true slope equaling 0. The least squares fit for the slope is quite close to 1, namely 1.0251.

12.82 Iraq war and reading newspapers

Regression toward the mean indicates that for any particular pre-war RBS, the predicted during-war RBS will be relatively closer to its mean than the pre-war RBS will be to its mean. Thus, the finding that the during-war scores are not as extreme as the pre-war scores for the light and heavy readers may merely be reflecting the tendency for regression toward the mean.

12.83 Sports and regression

There are many examples that students could use in their response. Here's one example. If a Major League baseball player has an amazing year with many homeruns, he's unlikely to have that many the following year. If we look at the top ten homerun hitters for a given year, most of them are likely to have fewer homeruns the following year. Once a hitter reaches a given level, he's more likely to come back toward the average; it's hard to go up at that point.

12.84 Regression toward the mean paradox

Regression toward the mean does not imply that, over many generations, there are fewer and fewer very short people and very tall people. To reverse the logic, most very tall people had parents who were a bit shorter than they, and most very short people had parents who were a bit taller. Extreme observations will always occur, but then the offspring of the extremely tall or short, the next generation, will likely be closer to the mean.

12.85 Height and weight

As the range of values reflected by each sample is restricted, the correlation tends to decrease when we consider just students of a restricted range of ages. Using the two samples will increase the ranges for height and weight, and likely increase the correlation.

12.86 Mileage and weight

Correlation tends to decrease as the ranges of the values of the sample decrease. Sports cars have a smaller range of weight and a smaller range of mileage.

12.87 Dollars and pounds

a) The slope would change because it depends on units. The slope would be 2 times the original slope.

b) The correlation would not change because it is independent of units.

c) The t statistic would not change because although the slope doubles, so does its standard error. (The results of a test should not depend on the units we use.)

12.88 All models are wrong

a) All models are wrong because reality is never so simple as to follow, for example, exactly a straight line for how the mean of y changes as x changes, with exactly the same variability on y at all values of x.

b) Some models are not useful because they are applied inappropriately. For example, a linear model might be applied inappropriately to data that have a curvilinear pattern.

12.89 *df* for *t* tests in regression

a) $df = n$ – the number of parameters; for the model, $\mu_y = \alpha + \beta x$, there are two parameters α and β, and so $df = n - 2$.

b) When the inference is about a single mean, there is only one parameter, and therefore, $df = n - 1$.

12.90 Assumptions

a) To describe the relationship between two variables, the assumption is that the population mean of *y* has a straight-line relationship with *x.*

b) To make inferences about the relationship, we assume that the population mean of *y* has a straight-line relationship with *x*, that the data were gathered using randomization, and that population *y* values at each *x* value have a normal distribution with the same standard deviation at each *x* value. The third assumption is least critical, especially when the sample size is large.

12.91 Assumptions fail?

a) The percentage of unemployed workers would likely fluctuate quite a bit between 1900 and 2005, and this would not be a linear relationship.

b) Annual medical expenses would likely be quite high at low ages, then lower in the middle, then high again, forming a parabolic, rather than linear, relationship.

c) The relation between these variables is likely curvilinear. Life expectancy increases at first as per capita income increases, and then gradually levels off.

12.92 Lots of standard deviations

a) s_y is the standard deviation of all *y* values around the mean of all *y v*alues, $\bar{y}$.

b) s_x is the standard deviation of all *x* values around the mean of all *x v*alues, $\bar{x}$.

c) The residual standard deviation *s* is the standard deviation of all *y* values at a particular *x* value. It summarizes the sample variability around the regression line.

d) *se* of the slope estimate *b* describes the variability of the sampling distribution that measures how that estimate varies from sample to sample of size *n*.

12.93 Decrease in home values

a) The statement is referring to additive growth, but this is multiplicative growth. There is an exponential relation between these variables.

b) $\hat{y} = \$175{,}000(0.966)^{10} = \$123{,}825$; $(\$123{,}825 - \$175{,}000)/\$175{,}000 = -0.292$, so the percentage decrease for the decade is about 29.2%.

12.94 Population growth

a) $(1.0123)^{10} = 1.13$, which indicates a growth rate of about 13%.

b) At a growth rate of 1.23%, we'd multiply the initial population by 1.0123 each year. The exponent of *x* is necessary to indicate the exponential growth over *x* number of years.

12.95 Multiple choice: Interpret *r*

The best response is (b).

12.96 Multiple choice: Correlation invalid

The best response is (a).

12.97 Multiple choice: Slope and correlation

The best response is (d).

12.98 Multiple choice: Regress *x* on *y*

The best response is (a).

12.99 Multiple choice: Income and height

The best response is (b).

12.100 True or false

a) True
b) True
c) False
d) True
e) False

♦♦12.101 Golf club velocity and distance

a) We would expect that at 0 impact velocity, there would be 0 putting distance. The line would pass through the point having coordinates (0, 0).
b) If x doubles, then x^2 (and hence the mean of y) quadruples. For example, if x goes from 2 to 4, then x^2 goes from 4 to 16.

♦♦12.102 Why is there regression toward the mean?

a) For every standard deviation (represented in the units of the variable) that x changes, $\hat{y}$ will change by that many units times the slope.
b) An increase of 1 standard deviation in x is s_x units which, from (a), results in a change in $\hat{y}$ of $s_x b$ units, but this is equal to $r_x y$, or r standard deviations in y.

♦♦12.103 r^2 and variances

Because $r^2 = \dfrac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$, and dividing each term by approximately n (actually, $n-1$ and $n-2$) gives the variance estimates, it represents the relative difference between the quantity used to summarize the overall variability of the y values (i.e. the variability of the marginal distribution of y) and the quantity used to summarize the residual variability (i.e. the variance of the conditional distribution of y for a given x). These go in the numerator of the respective variance estimates, and their denominators are nearly identical ($n-1$ and $n-2$). Therefore, the estimated variance of the conditional distribution of y for a given x is approximately 30% smaller than the estimated variance of the marginal distribution of y.

♦♦12.104 Standard error of slope

a) A smaller numerator (s) will lead to a smaller value (se). If the standard error of the sample slope is smaller, it is a better estimate of the population slope. This is true because se of the slope estimate b describes the spread of the sampling distribution that measures how that estimate varies from sample to sample of size n. If it varies little from sample to sample, then it is a more precise estimate of the population slope.
b) The residual standard deviation decreases when the typical size of the residuals is smaller. If the residuals are smaller, that indicates that the observations are closer to the prediction equation.
c) As the sample size increases, the denominator of the expression increases, and se decreases. When the x values are more highly spread out, the denominator also increases, and se decreases.

♦♦12.105 Regression with an error term

a) Error is calculated by subtracting the mean from the actual score, y. If this difference is positive, then the observation must fall above the mean.
b) $\varepsilon = 0$ when the observation falls exactly at the mean. There is no error in this case.
c) Because the residual is $e = y - \hat{y}$, $e = y - \hat{y} \Rightarrow y = \hat{y} + e \Rightarrow y = (a + bx) + e = a + bx + e$. As $\hat{y}$ is an estimate of the population mean, e is an estimate of ε.
d) It does not make sense to use the simpler model, $y = \alpha + \beta x$, that does not have an error term because it is improbable that every observation will fall exactly on the regression line; it is improbable that there will be no error.

♦♦12.106 Rule of 72

a) The actual number of years necessary for the investment to reach 2000 is the value of x for which the exponential regression with multiplicative effect of 1.06 gives a predicted value of 2000.
b) Using the rule $\log(a^x) = x\log(a)$; $1.06^x = 2 \Rightarrow \log(1.06^x) = \log(2) \Rightarrow x\log(1.06) = \log(2) \Rightarrow$ $x\log(1.06) = \log(2)/\log(1.06) = 12$.

12.106 (continued)

c) (i) 72/1 = 72 years
(ii) 72/18 = 4 years

Chapter Problems: Student Activities

12.107 Analyze your data

Responses will be different based on the data files for each class and the variables chosen by each instructor.