Chapter 2 Solutions

2.1 False. The slope of the line depends on the scale for Y and X as well as the correlation between them. If we make all of the X values smaller (say by dividing each by 1000), the slope will get bigger (multiplying by 1000 to compensate), even though the correlation doesn't change at all.

2.2 True. The total degrees of freedom $(n-1)$ are split to 1 for the model and $n-2$ for the error term.

2.3 True. When finding a confidence interval for the slope, as the sample size increases, the degrees of freedom for the t-distribution increases. This in turn causes the t^* value required to achieve a given confidence level to decrease slightly.

2.4 False. A very weak predictor might explain little or no variability at all, making $SSModel$ very small. Since $SSTotal = SSModel + SSE$, the sum of squares for the error, SSE gets larger as $SSModel$ gets smaller. For any model with $r^2 < 50\%$, we will have $SSModel < SSE$.

2.5 False. The size of a typical error is measured by the standard deviation of the error term, $\hat{\sigma}_\epsilon$, which is also a term in computing the margin of error for a prediction interval. So the interval will get wider as $\hat{\sigma}_\epsilon$ increases. Also, if we have larger errors, we have less accuracy in the predictions, so we need a wider interval to capture a new observation.

2.6 True. We need a wider interval to capture an individual value than to capture the mean response for a particular value of the predictor. This can be seen in the extra "`1+`" term that appears under the square root in the formula for computing a prediction interval that is not present when computing a confidence interval for the mean response.

2.7 True. The coefficient of determination is r^2, so a larger correlation (in magnitude) will give a larger value for r^2.

2.8 (c) The correlation of $r = 0.6$ means that $r^2 = 0.6^2 = 0.36$, or 36% of variability in the response Y is explained by the linear model based on the predictor X.

2.9 a. A high value of r^2 could occur, for example, with a scatterplot that shows a steep but obviously curved relationship. So a model based on a transformation might give an even better fit than the linear model.

b. A low r^2 does not necessarily imply that another model would provide a better description of the relationship. For example, we could generate data from a linear model with a large variance in the error term. This could produce a low r^2, but the linear model is still the "correct" form of the relationship.

2.10 The width of a prediction interval depends on $SE_{\hat{y}} = \hat{\sigma}_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum (x - \overline{x})^2}}$.

a. Increasing the sample size will reduce $SE_{\hat{y}}$ because the $1/n$ term gets smaller and $\sum (x - \overline{x})^2$ increases with more terms in the sum. Also, the t^* value for the interval will be slightly smaller with more degrees of freedom.

b. Increasing the variability in the predictor values makes $\sum (x - \overline{x})^2$ larger, so $SE_{\hat{y}}$ is smaller and the interval is narrower.

c. Increasing the variability of the response (σ_ϵ) tends to make the estimate $\hat{\sigma}_\epsilon$ larger, which increases $SE_{\hat{y}}$ and makes the interval wider.

d. Choosing a value for x^* that is farther from $\overline{x}$ increases the value of $(x^* - \overline{x})^2$, thus increasing $SE_{\hat{y}}$ and making the interval wider.

2.11 a. We test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. The test statistic is

$$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} = \frac{15.5}{3.4} = 4.56$$

We use a t-distribution with $40 - 2 = 38$ d.f. to find the P-value $= 2P(t_{38} > 4.56) = 0.00005$. This very small P-value gives strong evidence to reject H_0 and conclude that the slope of the regression model is different from zero. The data suggest that the slope is positive.

b. With 38 d.f., the t^* value for 95% confidence is 2.024. Thus the confidence interval for the slope is

$$\hat{\beta}_1 \pm t^* SE_{\hat{\beta}_1} = 15.5 \pm 2.024(3.4) = 15.5 \pm 6.88 = (8.62, 22.38).$$

We can be 95% sure that the slope of the linear model for the entire population is between 8.62 and 22.38.

2.12 a. We test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. The test statistic is

$$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} = \frac{5.3}{2.8} = 1.89$$

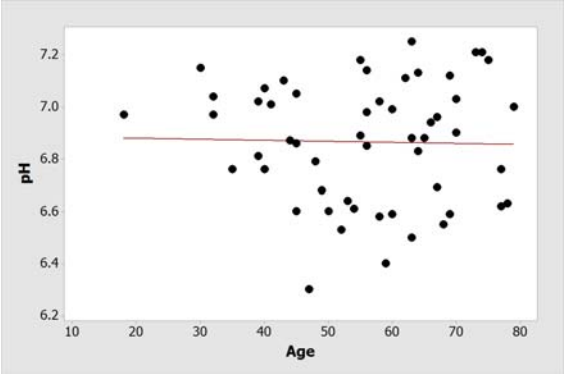
We use a t-distribution with $82 - 2 = 80$ d.f. to find the P-value $= 2P(t_{80} > 1.89) = 0.062$. This moderate P-value does not give enough evidence to reject H_0 and conclude that the slope of the regression model is different from zero.

b. With 80 d.f., the t^* value for 95% confidence is 1.99. Thus the confidence interval for the slope is

$$\hat{\beta}_1 \pm t^* SE_{\hat{\beta}_1} = 5.3 \pm 1.99(2.8) = 5.3 \pm 5.572 = (-0.272, 10.872).$$

We can be 95% sure that the slope of the linear model for the entire population is between -0.272 and 10.872.

2.13 a. The scatterplot shows a very weak negative linear relationship between pH and age.

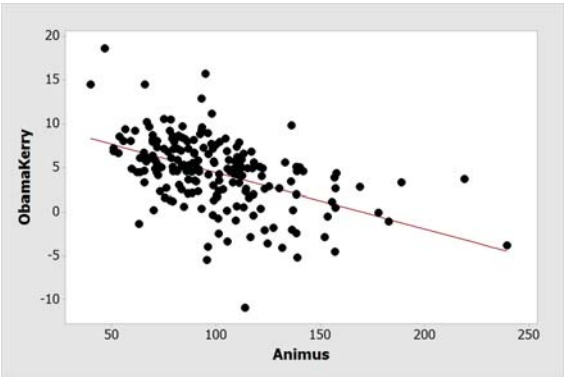


b. We test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$. Computer output shows

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.8881113  0.1321194   52.13   <2e-16 ***
Age         -0.0003905  0.0022944   -0.17    0.866
```

The small test statistic ($t = -0.17$) and large P-value (0.866) do not provide evidence to reject H_0. We fail to reject the hypothesis that there is no linear relationship between pH and age.

2.14 a. Yes, the scatterplot given below shows that there is a negative linear relationship.



b. We test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$. Computer output shows

```
Term          Coef  SE Coef  T-Value  P-Value
Constant    10.849    0.857    12.66    0.000
Animus    -0.06397  0.00827    -7.74    0.000
```

The large (in absolute value) test statistic ($t = -7.74$) and tiny P-value (approx. 0) provides strong evidence to reject H_0. We reject the hypothesis that there is no linear relationship between *ObamaKerry* and *Animus*.

2.15 a. We test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$. Computer output shows

```
The regression equation is Calories = 87.4 + 2.48 Sugar

Predictor    Coef  SE Coef      T      P
Constant   87.428    5.163  16.93  0.000
Sugar      2.4808   0.7074   3.51  0.001
```

The large test statistic (t = 3.51) and small P-value (0.001) provide evidence to reject H_0 and the data suggest that there is probably a positive relationship between sugar content and calories in cereals. We could also do this test using the ANOVA table that shows $F = 12.30$ and P-value = 0.001.

b. Using $SE_{\hat{\beta}_1} = 0.7074$ from the output and $t^* = 2.032$ with 34 d.f., the confidence interval for the slope is

$$2.4808 \pm 2.032(0.7074) = 2.4808 \pm 1.437 = (1.044, 3.918)$$

We are 95% confident that the average increase in calories for each extra gram of sugar in cereals is between 1.044 and 3.918 calories.

2.16 a. The hypotheses are $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$, where β_1 is the slope for the regression model to predict textbook price based on number of pages. The information below was obtained by running this model with statistical software for the data in **TextPrices**.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.42231   10.46374  -0.327    0.746
Pages        0.14733    0.01925   7.653 2.45e-08
```

The relevant test statistic is $t = 7.653$ and the P-value is 2.5×10^{-8}, or roughly zero. There is strong evidence to reject H_0 and conclude that the number of pages is related to the price of a textbook. We could also do this test using the ANOVA table that shows $F = 58.57$ and the same P-value.

b. Using the information from the computer output and $t^* = 2.048$ with $n - 2 = 28$ d.f., the confidence interval for the slope is

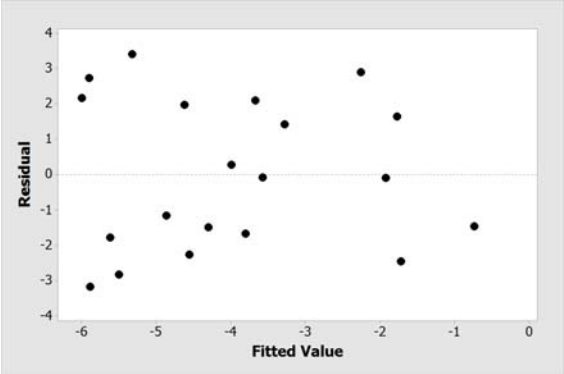
$$0.14733 \pm 2.048(0.01925) = 0.14733 \pm 0.03942 = (0.1079, 0.1868)$$

We are 95% confident that, as the number of pages goes up by 1, the average price of a textbook goes up by between 0.1078 and 0.1868, or between 11 cents and 19 cents, roughly.

2.17 a. Yes. According to the computer output below, $t = 3.18$ and the P-value = 0.005.

```
Term        Coef  SE Coef  T-Value  P-Value
Constant  -2.436    0.686    -3.55    0.002
APC        1.344    0.422     3.18    0.005
```

b. Here is a plot of residuals against fitted values. There is no pattern in the plot, which is good.



c. The estimated slope is 1.344 and the standard error is 0.422.

d. The 90% CI is (0.61,2.08). The interval does not contain zero, because the P-value for the test, from part (a), is less than 0.10.

2.18 Here is some output for fitting a linear model to predict $Weight$ using $WingLength$.

```
Predictor      Coef  SE Coef      T      P
Constant     1.3655   0.9573   1.43  0.156
WingLength  0.46740  0.03472  13.46  0.000
```

a. To test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$, we use the t-statistic, $t = 13.46$, from the output with a P-value that is essentially zero. This small P-value gives strong evidence that the slope is different from zero and that there is some relationship between $Weight$ and $WingLength$.

b. To find a 95% confidence interval for the slope, we use $\hat{\beta}_1 = 0.4674$ and $SE_{\hat{\beta}_1} = 0.03472$ from the output together with $t^* = 1.981$ for a t-distribution with 114 degrees of freedom.

$$\hat{\beta}_1 \pm t^* SE_{\hat{\beta}_1} = 0.4674 \pm 1.981(0.03472) = 0.4674 \pm 0.0688 = (0.399, 0.536)$$

We are 95% sure that the slope of the model to predict sparrow weight based on wing length is between 0.399 and 0.536.

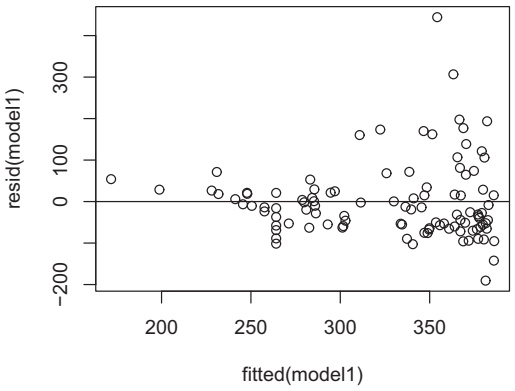
c. No, the 95% confidence interval does not include zero. This makes sense because of the duality between confidence intervals and hypothesis tests: The P-value for the test is small, which is consistent with the confidence interval excluding zero as a plausible value for the slope.

2.19 a. Computer output below shows the equation to be $\widehat{adj2007} = 388 - 54.4distance$. So each mile closer to the bike trail is associated with a mean price increase of about \$54,000.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  388.204     14.052  27.626  < 2e-16 ***
distance     -54.427      9.659  -5.635 1.56e-07 ***
---
Residual standard error: 92.13 on 102 degrees of freedom
Multiple R-squared: 0.2374,Adjusted R-squared: 0.2299
F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

b. The 90% confidence interval: We are 90% confident that the mean price increase for each mile closer to a bike trail is between \$38,390 and \$70,460.

c. The residual versus fitted values plot (RailsTrailsCh2Q1c.eps) shows a lack of constant variance, which could call the validity of (b) into question.

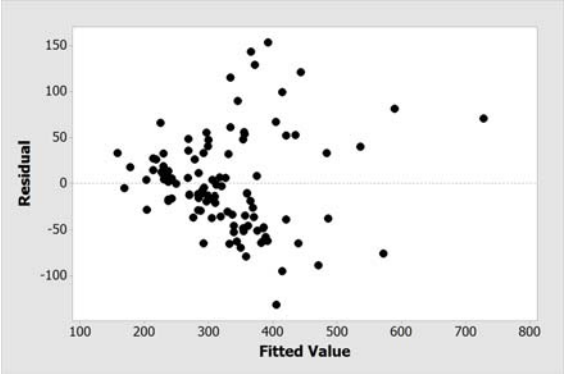


2.20 a. The fitted slope of 162.526 answers this question. Each additional thousand square feet of floor space is associated with an average selling price increase of about \$162,000.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   72.973     15.541   4.695 8.32e-06 ***
squarefeet   162.526      9.351  17.381  < 2e-16 ***
---
Residual standard error: 53 on 102 degrees of freedom
Multiple R-squared: 0.7476,Adjusted R-squared: 0.7451
F-statistic: 302.1 on 1 and 102 DF,  p-value: < 2.2e-16
```

b. We are 90% confident that each additional thousand square feet of floor space is associated with an average price increase of between \$147,000 and \$178,050.

c. The residuals-versus-fitted values plot suggests increasing variance for larger fitted values (i.e., bigger homes). This calls (b) into some question.



2.21 $r^2 = \frac{SSModel}{SSTotal} = \frac{110}{150} = 0.733$. This means that 73.3% of the variability in this response variable Y is explained by the linear model based on the predictor X.

2.22 $r^2 = \frac{SSModel}{SSTotal} = \frac{38}{102} = 0.373$. This means that 37.3% of the variability in this response variable Y is explained by the linear model based on the predictor X.

2.23 Here is some output for fitting a linear model to predict *Price* using *Year* after removing the first four data points.

```
The regression equation is Price = - 1647 + 0.841 Year

Predictor      Coef  SE Coef       T      P
Constant   -1647.17    46.86  -35.15  0.000
Year        0.84098  0.02357   35.68  0.000

S = 1.73702   R-Sq = 98.5%   R-Sq(adj) = 98.5%
```

a. In the output, we see that `R-Sq = 98.5%`, which means that 98.5% of the variation in stamp prices over these years is explained by the year.

b. We can test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$ using the t-statistic (35.68) and P-value (essentially zero) from the regression output. This large t-statistic and very small P-value give very strong evidence of some relationship between the price of stamps and year.

c. Here is the ANOVA table from the computer output.

```
Source          DF      SS      MS        F      P
Regression       1  3841.2  3841.2  1273.09  0.000
Residual Error  19    57.3     3.0
Total           20  3898.6
```

The large F-statistic ($F = 1273.09$) and P-value that is essentially zero give strong evidence that *Year* has some value in predicting the price of stamps.

2.24 Here is some output for fitting a linear model to predict $LogMrate$ using $LogBodySize$ based on this sample of caterpillars.

```
The regression equation is LogMrate = 1.31 + 0.916 LogBodySize

Predictor       Coef  SE Coef      T      P
Constant     1.30655  0.01356  96.33  0.000
LogBodySize  0.91641  0.01235  74.20  0.000

S = 0.175219   R-Sq = 94.8%   R-Sq(adj) = 94.8%
```

a. We see in the output that the fitted linear model is $\widehat{LogMrate} = 1.3066 + 0.9164 LogBodySize$.

b. We test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$ using the t-statistic (74.20) and P-value (essentially zero) from the regression output. This large t-statistic and very small P-value give very strong evidence of some relationship between the $LogMrate$ and $LogBodySize$.

c. Here is the ANOVA table from the computer output.

```
Source          DF      SS      MS        F      P
Regression       1  169.02  169.02  5505.26  0.000
Residual Error 303    9.30    0.03
Total          304  178.32
```

The large F-statistic ($F = 5505.26$) and P-value that is essentially zero give strong evidence that $LogBodySize$ has some value in predicting the log of the metabolic rate for this type of caterpillar.

d. The ratio is $\dfrac{SSModel}{SSTotal} = \dfrac{169.02}{178.32} = 0.9478$. This is just the computation for r^2, which tells us that 94.8% of the variability in log metabolic rate for these caterpillars is explained by the log of their body sizes.

2.25 a. The correlation coefficient between $Weight$ and $WingLength$ is $r = 0.7835$. To test $H_0 : \rho = 0$ versus $H_a : \rho = 0$, the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.7835\sqrt{116-2}}{\sqrt{1-0.7835^2}} = 13.46$$

Comparing this to a t-distribution with 114 degrees of freedom shows a P-value that is essentially zero. This gives strong evidence for some association between weight and wing length for sparrows.

b. The percent of variability in these sparrow weights that is explained by their wing lengths is $r^2 = 0.7835^2 = 0.614$ or 61.4%. You could also find this from the ANOVA table in the next part.

c. Here is some computer output with the ANOVA table for this model.

```
Analysis of Variance
Source           DF      SS      MS       F      P
Regression        1  355.05  355.05  181.25  0.000
Residual Error  114  223.31    1.96
Total           115  578.36
```

The large F-statistic ($F = 181.25$) and small P-value (essentially zero) provide strong evidence that wing length has some value for predicting sparrow weights.

d. The square root of the F-statistic from part (c) is $\sqrt{181.25} = 13.46$ which matches the t-statistic from part (a).

2.26 a. The correlation coefficient between $Width$ and $Year$ is $r = -0.247$. To test $H_0 : \rho = 0$ versus $H_a : \rho = 0$, the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.247\sqrt{252-2}}{\sqrt{(1-(-0.247)^2)}} = -4.03$$

Comparing this to a t-distribution with 250 degrees of freedom shows a P-value that is essentially zero. This gives strong evidence for some association between width and year for these kinds of leaves.

b. The percent of variability in these leaf widths that is explained by the year of measurement is $r^2 = (-0.247)^2 = 0.061$ or 6.1%. You could also find this from the ANOVA table in the next part.

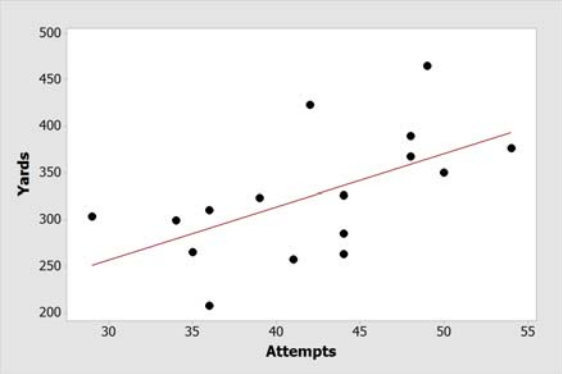
c. Here is some computer output with the ANOVA table for this model.

```
Analysis of Variance
Source       DF       SS       MS      F      P
Regression    1   32.911  32.9113  16.24  0.000
Error       250  506.764   2.0271
Total       251  539.676
```

The large F-statistic ($F = 16.24$) and small P-value (essentially zero) provide strong evidence that year has some value for predicting leaf width.

d. The square root of the F-statistic from part (c) is $\sqrt{16.24} = -4.03$ which matches the t-statistic from part (a).

2.27 a. Some output for predicting *Yards* based on *Attempts* for the data in **BreesPass** is shown below along with a scatterplot that includes the least squares line.




```
Coefficients:
    Estimate Std. Error t value Pr(>|t|)
(Intercept)   86.140     90.362   0.953   0.3566
Attempts       5.691      2.122   2.682   0.0179 *
 ---
Residual standard error: 56.34 on 14 degrees of freedom
Multiple R-squared:  0.3394,Adjusted R-squared:  0.2922
F-statistic: 7.191 on 1 and 14 DF,  p-value: 0.01789
```

The prediction equation is $\widehat{Yards} = 86.1 + 5.69 Attempts$.

b. Brees passed for a total of 5208 yards on 673 attempts, which gives an average of $5208/673 = 7.7$ yards per attempt. This is a bit off from the slope of the regression line ($\hat{\beta}_1 = 5.69$).

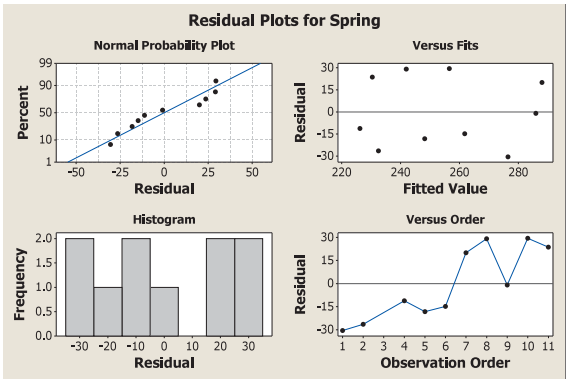
c. Using $r^2 = 0.339$ from the output, we can conclude that 33.9% of the variability in Brees's game yardage can be explained by the number of attempts.

2.28 Here is some output for fitting a linear model to predict *Spring* using *Fall* with the data for 2003 omitted.

```
The regression equation is Spring = 548 - 1.05 Fall

Predictor     Coef  SE Coef      T      P
Constant     548.0    106.7   5.13  0.001
Fall       -1.0483   0.3805  -2.75  0.025
```

```
S = 24.9411   R-Sq = 48.7%   R-Sq(adj) = 42.3%

Analysis of Variance
Source           DF      SS      MS     F      P
Regression        1  4721.1  4721.1  7.59  0.025
Residual Error    8  4976.5   622.1
Total             9  9697.6
```

a. Some residual plots for this model to predict spring enrollments based on fall enrollments (with the 2003 data omitted) are shown below. The normal probability plot shows some slight curvature in the upper tail, and it's hard to say much definitive about the histogram with such a small sample. The plot of the residuals against fitted values looks fairly typical, an unstructured pattern. The positive association in the plot of the residuals against order (or *AYear*) indicates that it might be beneficial to add *Ayear* to the model.



b. In the computer output, we see `R-sq = 48.7%`, so we find that 48.7% of the variability in spring enrollments over these years can be explained by the fall enrollments.

c. The ANOVA table is shown in the output above. The F-statistic is $F = 7.59$ and the P-value is 0.025.

d. We can test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$ using either the ANOVA F-statistic (from part (c)) or the t-test for the slope in the output ($t = -2.75$, P-value $= 0.025$). The P-value in each case is small (and are identical to each other), so we have evidence to show a relationship between fall and spring enrollments.

e. To find a 95% confidence interval for the slope, we use $\hat{\beta}_1 = -1.048$ and $SE_{\hat{\beta}_1} = 0.3805$ from the output together with $t^* = 2.306$ for a t-distribution with 8 degrees of freedom.

$$\hat{\beta}_1 \pm t^* SE_{\hat{\beta}_1} = -1.048 \pm 2.306(0.3805) = -1.048 \pm 0.877 = (-1.93, -0.171)$$

We are 95% sure that the slope of the model to predict spring enrollment based on fall enrollment is between −1.93 and −0.171. This 95% confidence interval does not include zero, which is consistent with the test that shows evidence that the slope differs from zero.

2.29 Here is some output for fitting a linear model to predict $Wall03$ using $Gdiam03$.

```
The regression equation is Wall03 = - 1.05 + 0.368 Gdiam03

Predictor     Coef  SE Coef      T      P
Constant   -1.0521   0.4010  -2.62  0.009
Gdiam03    0.36821  0.02004  18.38  0.000

S = 1.50114   R-Sq = 36.3%   R-Sq(adj) = 36.2%
```

a. To test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$, we use the t-statistic, $t = 18.38$, from the output with a P-value that is essentially zero. This small P-value gives strong evidence that the slope is different from zero and that there is some relationship between the wall thickness and gall diameter in 2003.

b. In the output, we see that $\hat{\beta}_1 = 0.368$ and its standard error is $SE_{\hat{\beta}_1} = 0.0200$.

c. The size of a typical error is reflected in the regression standard error, $\hat{\sigma}_\epsilon = 1.50$.

d. No, the value of $r^2 = 36.3\%$ in the output shows that only 36.3% of the variability in wall thickness is explained by the gall diameter for these data from 2003.

e. Here is some computer output with 95% intervals for the wall thickness for goldenrod galls in 2003 when the diameter is 20 mm.

```
Gdiam03     Fit  SE Fit       95% CI            95% PI
     20  6.3122  0.0617  (6.1910, 6.4334)  (3.3615, 9.2629)
```

Based on the 95% CI, we are 95% confident that the mean wall thickness (in 2003) is between 6.19 mm and 6.43 mm when the gall diameter is 20 mm.

f. The correlation coefficient between $Wall03$ and $Gdiam03$ is $r = 0.602$. To test $H_0 : \rho = 0$ versus $H_a : \rho = 0$, the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.602\sqrt{595-2}}{\sqrt{1-0.602^2}} = 18.4$$

Comparing this to a t-distribution with 593 degrees of freedom shows a P-value that is essentially zero. This gives strong evidence for some association between wall thickness and diameter for goldenrod galls in 2003.

2.30 Here is some output for fitting a linear model to predict $Wall04$ using $Gdiam04$.

```
The regression equation is Wall04 = - 0.845 + 0.363 Gdiam04

Predictor     Coef  SE Coef      T      P
Constant   -0.8450   0.3577  -2.36  0.018
Gdiam04    0.36317  0.01719  21.12  0.000

S = 1.60835   R-Sq = 32.2%   R-Sq(adj) = 32.1%
```

a. To test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$, we use the t-statistic, $t = 21.12$, from the output with a P-value that is essentially zero. This small P-value gives strong evidence that the slope is different from zero and that there is some relationship between the wall thickness and gall diameter in 2004.

b. In the output, we see that $\hat{\beta}_1 = 0.363$ and its standard error is $SE_{\hat{\beta}_1} = 0.0172$.

c. The size of a typical error is reflected in the regression standard error, $\hat{\sigma}_\epsilon = 1.61$.

d. No, the value of $r^2 = 32.2\%$ in the output shows that only 32.2% of the variability in wall thickness is explained by the gall diameter for these data from 2004.

e. Here is some computer output with 95% intervals for the wall thickness for goldenrod galls in 2004 when the diameter is 20 mm.

```
Gdiam03     Fit  SE Fit       95% CI             95% PI
     20  6.4184  0.0533  (6.3137, 6.5231)  (3.2603, 9.5765)
```

Based on the 95% CI, we are 95% confident that the mean wall thickness (in 2004) is between 6.31 mm and 6.52 mm when the gall diameter is 20 mm.

f. The correlation coefficient between $Wall04$ and $Gdiam04$ is $r = 0.567$. To test $H_0 : \rho = 0$ versus $H_a : \rho = 0$, the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.567\sqrt{942-2}}{\sqrt{1-0.567^2}} = 21.1$$

Comparing this to a t-distribution with 940 degrees of freedom shows a P-value that is essentially zero. This gives strong evidence for some association between wall thickness and diameter for goldenrod galls in 2004.

2.31 Here is some output for fitting a linear model to predict $Hgt97$ using $Hgt90$.

```
The regression equation is Hgt97 = 307 + 2.32 Hgt90

Predictor     Coef  SE Coef      T      P
Constant   307.439    9.841  31.24  0.000
Hgt90       2.3224   0.4920   4.72  0.000
```

```
S = 78.7921   R-Sq = 2.7%   R-Sq(adj) = 2.6%
```

a. We can test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$ using the t-statistic (4.72) and P-value (essentially zero) from the regression output. This very small P-value gives strong evidence of some relationship between the pine heights in 1990 and 1997.

b. Based on the `R-sq=2.7%` value in the output, we see that only 2.7% of the variability in height of the trees in 1997 is explained by their heights in 1990.

c. Using technology, the ANOVA table for this regression model is

```
Source           DF       SS      MS      F      P
Regression        1   138344  138344  22.28  0.000
Residual Error  807  5010010    6208
Total           808  5148354
```

d. Based on information from the ANOVA table, the coefficient of determination is

$$r^2 = \frac{SSModel}{SSTotal} = \frac{138,344}{5,148,354} = 0.0269 \text{ or about } 2.7\%$$

e. No. Even though there is a significant linear association (and no problems with the conditions), the model only explains a very small fraction of the variation in the heights of the trees in 1997.

2.32 Here is some output for fitting a linear model to predict $Hgt97$ using $Hgt96$.

```
The regression equation is
Hgt97 = 40.6 + 1.10 Hgt96

Predictor     Coef  SE Coef       T      P
Constant    40.591    2.524   16.08  0.000
Hgt96      1.09606  0.00873  125.49  0.000

S = 18.4653   R-Sq = 94.9%   R-Sq(adj) = 94.9%
```

a. We can test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$ using the t-statistic (125.49) and P-value (essentially zero) from the regression output. This huge t-statistic and very small P-value give very strong evidence of some relationship between the pine heights in 1996 and 1997.

b. Based on the `R-sq = 94.9%` value in the output, we see that 94.9% of the variability in height of the trees in 1997 is explained by their heights in 1996.

c. Using technology, the ANOVA table for this regression model is

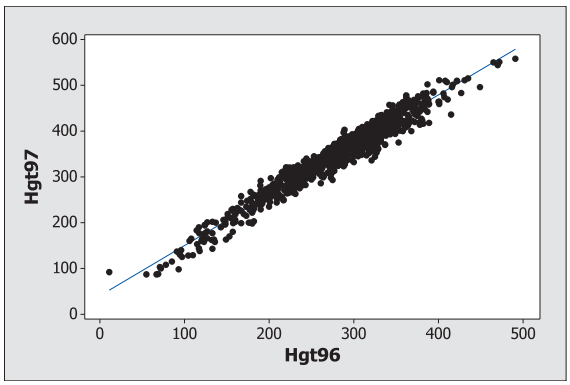
```
Source           DF       SS       MS         F      P
Regression        1  5369261  5369261  15747.14  0.000
Residual Error  852   290504      341
Total           853  5659765
```

Note: The SSTotal differs from the previous exercise due to cases lost due to missing data.

d. Based on information from the ANOVA table, the coefficient of determination is

$$r^2 = \frac{SSModel}{SSTotal} = \frac{5,369,261}{5,659,765} = 0.9487 \text{ or about } 94.9\%$$

e. Yes. The linear model appears to be a very effective way to summarize the relationship between the heights of pine tree seedlings over this year of growth. The scatterplot shows a strong linear trend with a slope that is clearly different from zero with almost 95% of the variability explained.



2.33 Here is some output for fitting a linear model to predict $Hgt97$ using $Hgt96$.

```
The regression equation is  Hgt97 = 40.6 + 1.10 Hgt96

Predictor      Coef  SE Coef       T      P
Constant     40.591    2.524   16.08  0.000
Hgt96       1.09606  0.00873  125.49  0.000

S = 18.4653   R-Sq = 94.9%   R-Sq(adj) = 94.9%
```

a. To find a 95% confidence interval for the slope, we use $\hat{\beta}_1 = 1.096$ and $SE_{\hat{\beta}_1} = 0.00873$ from the output together with $t^* = 1.963$ for a t-distribution with 852 degrees of freedom.

$$\hat{\beta}_1 \pm t^* SE_{\hat{\beta}_1} = 1.096 \pm 1.963(0.00873) = 1.096 \pm 0.017 = (1.079, 1.113)$$

We are 95% sure that the slope of the model to predict 1997 height based on 1996 height is between 1.079 and 1.113.

b. The value of 1 is not in our confidence interval. Thus we would reject the null hypothesis that the slope is equal to 1. Since both confidence limits are above 1, we are fairly sure the slope is bigger than one that indicates that the pine trees are growing.

c. No, the intercept has no practical meaning in this situation. We are never going to predict the heights of trees planted before 1990 that had no height in 1996.

2.34 a. Larger moths will produce more eggs, so the association is expected to be positive.

b. The correlation coefficient between $Eggs$ and $BodyMass$ for this dataset is $r = 0.441$.

c. To test $H_0 : \rho = 0$ versus $H_a : \rho = 0$, the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.441\sqrt{39-2}}{\sqrt{1-0.441^2}} = 2.99$$

Comparing this to a t-distribution with 37 degrees of freedom shows a P-value of 0.005. This gives strong evidence for some association between the body mass and number of eggs.

d. Here is some output for fitting a linear model to predict $Eggs$ using $BodyMass$.

```
The regression equation is Eggs = 24.4 + 79.9 BodyMass

Predictor   Coef  SE Coef     T      P
Constant   24.38    45.38  0.54  0.594
BodyMass   79.86    26.69  2.99  0.005

S = 44.7537   R-Sq = 19.5%   R-Sq(adj) = 17.3%
```

The fitted model is $\widehat{Eggs} = 24.38 + 79.86 BodyMass$.

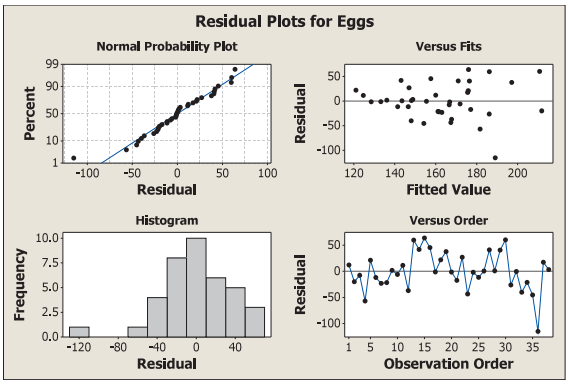
e. The last moth in the dataset (case #39) has a body mass of 1.668 but no eggs. All of the rest of the moths have many more eggs.

2.35 Here is some output for fitting a linear model to predict $Eggs$ using $BodyMass$ after removing the case with zero eggs.

```
The regression equation is Eggs = 29.6 + 79.2 BodyMass

Predictor   Coef  SE Coef     T      P
Constant   29.56    37.28  0.79  0.433
BodyMass   79.24    21.92  3.62  0.001

S = 36.7496   R-Sq = 26.6%   R-Sq(adj) = 24.6%
```

a. From the output we see that, without the zero egg case, the fitted model is $\widehat{Eggs} = 29.56 + 79.24 BodyMass$.

b. Some residual plots for this model (with the zero egg case removed) are shown as follows. The normal probability plot is linear, with the exception of one small observation. The histogram is centered at zero, with the unusually small residual clearly visible in the lower tail. The plot of the residuals against the fitted values shows only a few small fitted values, and the unusual residual (less than -100) is obvious again.

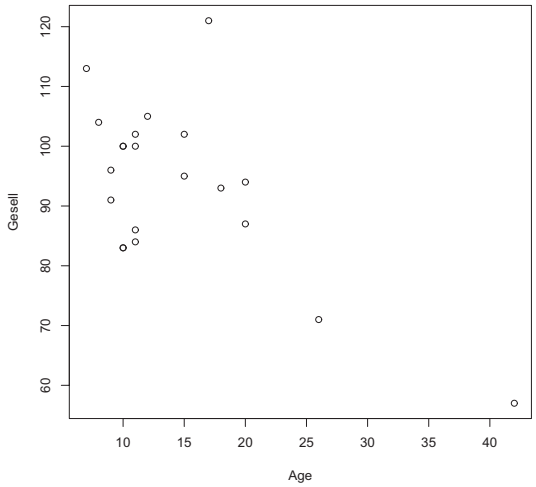


c. The estimated slope changes only slightly from 79.86 to 79.24.

d. Using the output from this model and the previous exercises, we see that eliminating the case with zero eggs increases the coefficient of determination from $r^2 = 19.5\%$ to $r^2 = 26.6\%$.

2.36 a. We might expect children who start talking early to have relatively high *Gesell* scores, which would imply a negative relationship between these variables.

b. The scatterplot shows a somewhat negative relationship between *Gesell* score and *Age* of first speaking, with a possible outlier at $Age = 42$.



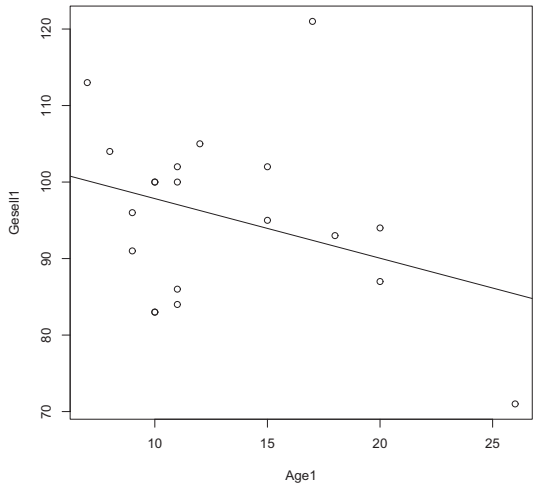
c. Software produces the output below for fitting this regression model.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.8738     5.0678  21.681 7.31e-15 ***
Age          -1.1270     0.3102  -3.633  0.00177 **

Residual standard error: 11.02 on 19 degrees of freedom
Multiple R-squared:  0.41,      Adjusted R-squared: 0.3789
F-statistic:  13.2 on 1 and 19 DF,  p-value: 0.001769
```

The prediction equation is $\widehat{Gesell} = 109.87 - 1.127Age$. The value of $r^2 = 0.41$ indicates that 41% of the variability in *Gesell* scores for these 21 children can be explained by the linear model based on *Age*. The small P-value = 0.00177 for the test of slope indicates that we have strong evidence that this is a statistically significant relationship and *Age* is a useful predictor of *Gesell* score.

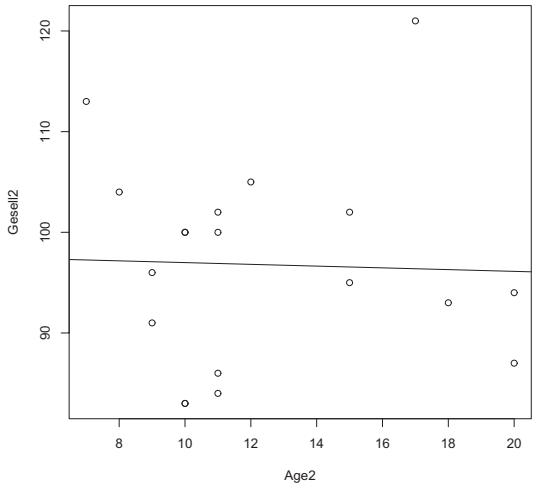
d. The largest residual is 30.3, which applies to child #19. This child had the highest *Gesell* score (121) but took slightly longer than average (17 months compared to $\overline{x} = 14.4$ months) to first speak.

2.37 With child #18 ($Age = 42$) removed, the scatterplot with least squares line is shown below. The equation of the least squares line is now $\widehat{Gesell} = 105.63 - 0.779Age$. This is less steep than the original fitted line and the slope is no longer significant (P-value = 0.149). The value of r^2 with the child at $Age = 42$ removed drops to only 11.2%.

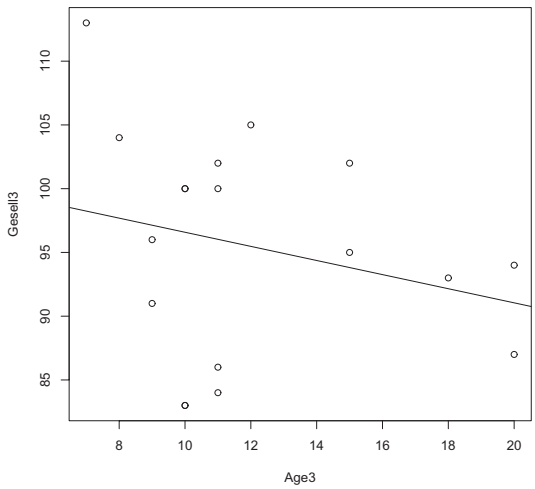


2.38 If we also remove child #2 ($Age = 26$), the regression line becomes almost flat, $\widehat{Gesell} = 97.86 - 0.087Age$. The P-value for testing for a significant slope is very large (0.89) and r^2 (0.0012)

is very small. There is little evidence of a linear relationship after the two largest *Age* values are removed.



2.39 Removing the child with $Gesell = 121$ (child #19 in the original dataset or child #17 in the dataset with the first two removed) produces a line with a steeper negative slope, $\widehat{Gesell} = 102.1 - 0.554Age$, but still not significant (P-value = 0.311). The r^2 value increases slightly to 0.064.



2.40 a. From software, the 90% confidence interval is ($281,833, $283,716).

b. From software, the 90% prediction interval is ($269,537, $296,012).

c. We are 90% confident that the mean sale price for a home listed at \$300,000 is between \$281,833 and \$283,716 (part (a) with interpretive statement). For a single home listed at \$300,000, we are 90% confident that its sale price would be between \$269,537 and \$296,012 (part (b) with interpretive statement).

2.41 a. From software, the 95% confidence interval is (1.767, 2.735).

b. From software, the 95% prediction interval is (−0.594, 5.097).

c. We are 95% confident that the mean leaf width in 2020 will be between 1.767 mm and 2.735 mm (part (a) with interpretive statement. For a single plant in 2020 we are 95% confident that its leaf width would be between −0.594 mm and 5.097 mm (part (b) with interpretive statement).

2.42 a. The following output gives a predicted value of −3.78.

```
Variable  Setting
APC            -1

     Fit    SE Fit         95% CI                95% PI
-3.78029  0.491819  (-4.81356, -2.74702)  (-8.48262, 0.922041)
```

b. The output in the solution to part (a) gives the interval as (−4.81,−2.75).

c. The output in the solution to part (a) gives the interval as (−8.48,0.92).

d. The prediction interval is intended to cover 95% of all people with APC measurements of −1, whereas the confidence interval is a statement about the average of all people with APC measurements of −1. There is a lot more uncertainty about where a single observation will land than there is about an average.

2.43 Here is some computer output with 95% intervals for the weight for sparrows with wing length of 20 mm.

```
WingLength      Fit  SE Fit       95% CI            95% PI
      20.0   10.714   0.285  (10.148, 11.279)  (7.884, 13.543)
```

a. Based on the output, the fitted value is 10.714 grams.

b. Based on the output, we are 95% sure that the mean weight of all sparrows with a 20-mm wing length is between 10.15 and 11.28 grams.

c. A 95% prediction interval for the weight of a sparrow with a wing length of 20 mm goes from 7.88 to 13.54 grams.

d. A wing length of 25 mm is closer to the average wing length of all sparrows (27.319) than a length of 20 mm, so the standard error of the prediction would be smaller.

2.44 a. Using software we request intervals for a predictor value of $x^* = 450$. Some output is shown as follows.

```
NewObs    Fit  SE Fit      95% CI           95% PI
      1  62.88    5.44  (51.73, 74.02)  (0.90, 124.85)

Values of Predictors for New Observations
NewObs  Pages
      1    450
```

For a confidence interval for the mean price, we use the part of the output labeled "95% CI," (51.73, 74.02). We are 95% sure that the mean price of all 450-page textbooks at the Cal Poly bookstore is between \$51.73 and \$74.02.

b. Using the same output above, we find that the 95% PI is (0.90, 124.85). We are 95% sure that a particular 450-page textbook will cost between \$0.90 and \$124.85 at the Cal Poly bookstore.

c. The midpoints of both intervals are the same, \$62.88, which is the predicted price for a 450-page textbook. Both intervals have the general form of $\hat{y} \pm$ some margin of error.

d. The confidence interval for the mean price is much narrower than the prediction interval for the price of an individual textbook. We need a much wider interval to capture most of the textbook prices than we do to just capture the mean of those prices.

e. The narrowest possible interval is when the number of pages is $x^* = \overline{x} = 464.5$ so that the term involving $(x^* - \overline{x})^2$ contributes nothing to the standard error, $SE_{\hat{y}}$.

f. Using software, we request intervals for a predictor value of $x^* = 1500$.

```
NewObs     Fit  SE Fit       95% CI            95% PI
      1  217.57   20.66  (175.25, 259.89)  (143.36, 291.78)XX
XX denotes a point that is an extreme outlier in the predictors.

Values of Predictors for New Observations
NewObs  Pages
      1   1500
```

The 95% prediction interval for textbook price when the number of pages is 1500 goes from \$143.36 to \$291.78. However, 1500 pages is much larger than any of the textbooks in the sample (as seen in the note in the output) and much of the prediction interval covers prices that are much larger than any of the prices in the sample. We should avoid this sort of extrapolation and thus would have less than 95% confidence that the interval would capture the price for a particular 1500-page textbook.

2.45 a. Using software we request intervals for a predictor value of $x^* = 150$. Some output is shown below.

```
Variable  Setting
Animus        150


    Fit    SE Fit         95% CI               95% PI
1.25341  0.490816  (0.285387, 2.22143)  (-5.78015, 8.28697)
```

For a confidence interval for the mean value of *ObamaKerry*, we use the part of the output labeled "95% CI," (0.285, 2.221). We are 95% sure that the mean value of *ObamaKerry* for all markets with and animus value of 150 is between 0.285 and 2.221.

b. Using the same output above, we find that the 95% PI is (−5.780, 8.287). We are 95% sure that in a particular market with an animus value of 150, the value of *ObamaKerry* will be between −5.780 and 8.827.

c. The midpoints of both intervals are the same, 1.253, which is the predicted value of $ObamaKerry$ for a market with animus value of 150. Both intervals have the general form of $\hat{y} \pm$ some margin of error.

d. The confidence interval for the mean value of *ObamaKerry* is much narrower than the prediction interval for the value of *ObamaKerry* of an individual market. We need a much wider interval to capture most of the *ObamaKerry* values than we do to just capture the mean of those markets.

e. The narrowest possible interval is when the animus value is $x^* = \overline{x} = 99.07$ so that the term involving $(x^* - \overline{x})^2$ contributes nothing to the standard error, $SE_{\hat{y}}$.

f. Using software, we request intervals for a predictor value of $x^* = 0$.

```
Variable  Setting
Animus          0


    Fit    SE Fit        95% CI              95% PI
10.8485  0.856986  (9.15834, 12.5388)  (3.67982, 18.0173)  XX

XX denotes an extremely unusual point relative to predictor levels used to fit the model.
```

The 95% prediction interval for *ObamaKerry* when the animus value is 0 goes from 3.680 to 18.017. However, an animus value of 0 is much smaller than any of the markets in the sample (as seen in the note in the output), and much of the prediction interval covers values that

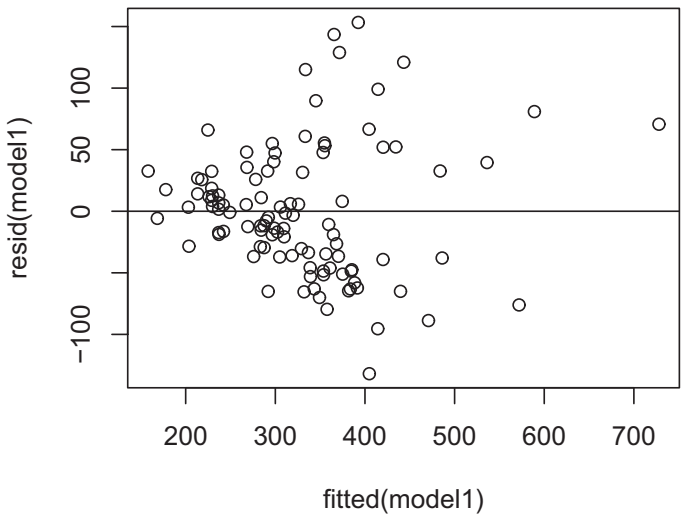
are larger than most of the values in the sample. We should avoid this sort of extrapolation and thus would have less than 95% confidence that the interval would capture the price for a particular market with animus value of 0.

2.46 a. The summary of the fitted model is given below in the computer ouput. Plugging 1.5 in for *squarefeet* (recall that units are in thousands), we get 316.7619 or (converting to dollars) $316,762 as the predicted selling price.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   72.973     15.541   4.695 8.32e-06 ***
squarefeet   162.526      9.351  17.381  < 2e-16 ***
---
Residual standard error: 53 on 102 degrees of freedom
Multiple R-squared: 0.7476,Adjusted R-squared: 0.7451
```

b. The prediction interval is (211,123, 422,401). Translating to dollars, we would predict with 95% confidence that the particular 1500 square-foot home would sell for between $211,123 and $422,401.

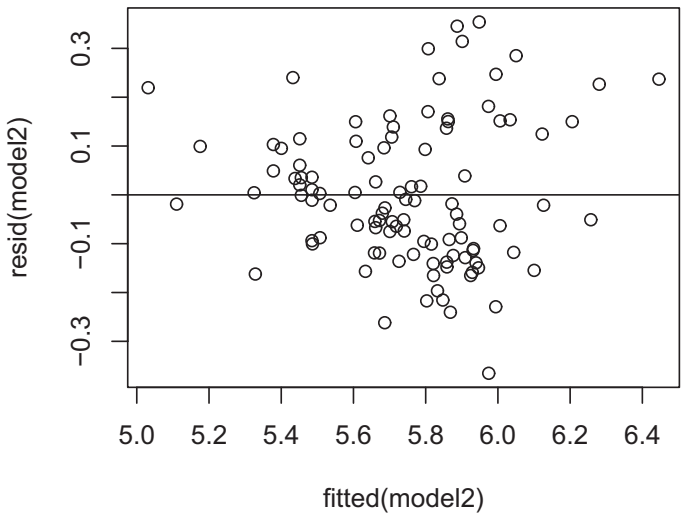
c. The residuals-versus-fitted values plot shows a slight lack of constant variance.



d. Using (natural) logs, we regress $\log(adj2007)$ on $\log(squarefeet)$ obtaining results given in the summary table below. The R^2 value is 0.7388 compared to 0.7476 for the un-logged model.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.47958    0.02137  256.41   <2e-16 ***
logsquarefeet  0.69334    0.04082   16.98   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1461 on 102 degrees of freedom
Multiple R-squared: 0.7388,Adjusted R-squared: 0.7362
F-statistic: 288.5 on 1 and 102 DF,  p-value: < 2.2e-16
```

The model conditions look good here; the residual plot adheres well to constant variance condition.



e. The 95% prediction interval is (5.469552, 6.051882). We first get the interval for the model with two logged variables. This interval is not in the thousands of dollar units of the original home prices. So we must exponentiate (anti-log) the interval. We can then claim to be 95% confident that a 1500 square foot home will sell for between 237,347 and 424,912. This interval is tighter than that obtained with the original variables. Since the logged variables give a model that adheres better to regression conditions, with similar R^2 value, we would put more trust in the prediction interval from part (d) over that of part (b).

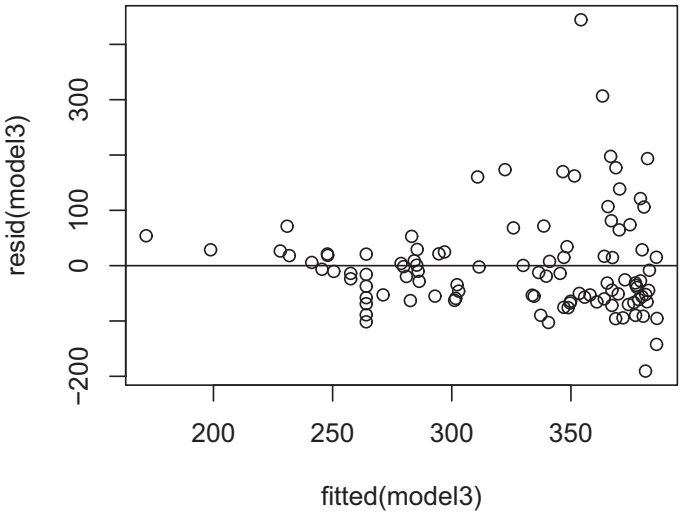
2.47 a. The model summary is given below. The fit is fairly weak, with only 23.74% of price variation explained by the regression onto *distance*. Each mile closer to a bike trail corresponds

to a price increase of about \$54,427. (Remember the units are miles for *distance*; thousands for selling price.) If we plug 0.5 mile in for distance we get a predicted selling price of $388.204 - 54.427(0.5) = 360,990$.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  388.204     14.052  27.626  < 2e-16 ***
distance     -54.427      9.659  -5.635 1.56e-07 ***
---
Residual standard error: 92.13 on 102 degrees of freedom
Multiple R-squared: 0.2374,Adjusted R-squared: 0.2299
F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

b. We are 90% confident that the home that resides half a mile from a trail will sell for between \$207,017 and \$514,964; the point estimate of price is \$360,990.

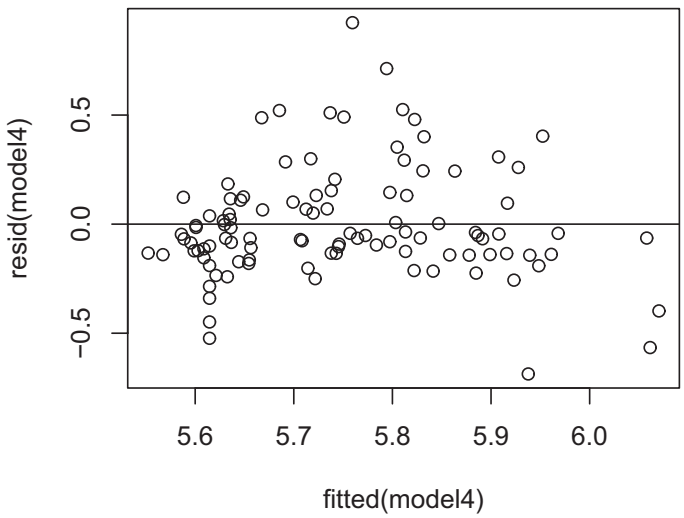
c. The residual plot shows a clear lack of the constant variance model condition, raising questions about (a) and (b).



d. The 90% prediction interval is (5.355955, 6.212704). This interval is in the log scale; we then exponentiate to get the interval back into dollars, so we can compare with the previous result and also, of course, just to interpret the interval in natural terms. We are 90% confident that this particular home, located half a mile from a trail, will sell for between \$211,866 and

$499,049. The estimated selling price is $325,164. This interval is a bit narrower than that obtained in (b).

e. The residual plot shows very good adherence now to the constant variance condition. Logging the data has been helpful.



2.48 a. The largest (absolute) residual for fitting a line to the data in **BreesPass** occurs in the 5^{th} game against CAR when Brees passed for 465 yards on 49 attempts. The predicted value is $\widehat{Yards} = 86.1 + 5.69(49) = 364.9$ yards, so the residual is $465 - 364.9 = 100.1$ yards. This was a particulary good game for Brees since he gained many more yards than would be expected for the number of attempts.

b. Using statistical software with a value for *Attempts* of $x^* = 40$, we get output for the prediction interval.

```
fit     lwr     upr
314     189     439
```

Thus we are 90% sure that Brees will pass for between 189 and 439 yards when he makes 40 throws in a game.

c. Another player could be better (more yards per attempt) or worse (fewer yards per attempt) than Brees, so we would not expect the same model to apply to all quarterbacks. Also, a

different player may get many fewer or many more attempts in a game, which would produce a problem with extrapolation from Brees's model.

2.49 a. From the output in the previous exercise, the fitted model is $\widehat{Spring} = 548.0 - 1.048Fall$. When the fall enrollment is 290 students, we have

$$\widehat{Spring} = 548.0 - 1.048(290) = 244 \text{ students}$$

b. Here is some output with regression intervals for this model when $Fall = 290$.

```
Fall     Fit  SE Fit       95% CI            95% PI
 290  244.00    8.81  (223.69, 264.31)  (183.01, 305.00)
```

Based on the `95% CI` in the output, we are 95% sure that the average spring enrollment for all years with a fall enrollment of 290 is between 223.7 and 264.3 students.

c. Using the `95% PI` from the output in part (b), when the fall enrollment in a particular year is 290, we are 95% sure that between 183 and 305 students will enroll in math classes the next spring.

d. Since the administrator wants an interval for the spring enrollment in a particular year, she should use the prediction interval in part (c).

2.50 a. The linear model from the previous exercise is $\widehat{LogMrate} = 1.3066 + 0.9164LogBodySize$. When $LogBodySize = 0$, the predicted $LogMrate$ is

$$\widehat{LogMrate} = 1.3066 + 0.9164(0) = 1.3066$$

Since the log in this example uses base 10, we convert to a prediction of the metabolic rate with

$$\widehat{Mrate} = 10^{\widehat{LogMrate}} = 10^{1.3066} = 20.26$$

This is the predicted metabolic rate for a caterpillar with $LogBodysize = 0$, which means its body size is $10^0 = 1$ gram.

b. The following output shows regression intervals for $LogMrate$ when $LogBodySize = 0$.

```
LogBodySize     Fit  SE Fit       95% CI            95% PI
          0  1.3066  0.0136  (1.2799, 1.3332)  (0.9607, 1.6524)
```

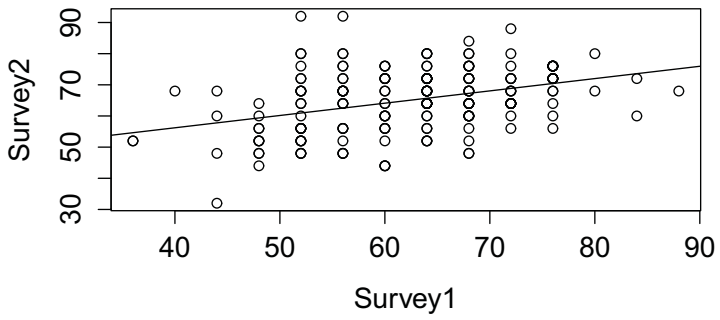
The width of the prediction interval is $1.6524 - 0.9607 = 0.6917$ and the width for the confidence interval for the mean is $1.3332 - 1.2799 = 0.0533$. The PI is $0.6917 - 0.0533 = 0.6384$, or about 12 times wider than the CI for the mean when $LogBodySize = 0$.

c. The output below shows regression intervals for $LogMrate$ when $LogBodySize = -2$.

```
LogBodySize     Fit  SE Fit       95% CI            95% PI
      -2    -0.5263  0.0185  (-0.5627, -0.4898)  (-0.8730, -0.1796)
```

The width of the prediction interval is $-0.1796-(-0.8730)=0.6934$ and the width for the confidence interval for the mean is $-0.4898-(-0.5627)=0.0729$. The PI is $0.6934-0.0729=0.6205$ or about 8.5 times wider than the CI for the mean when $LogBodySize=-2$.

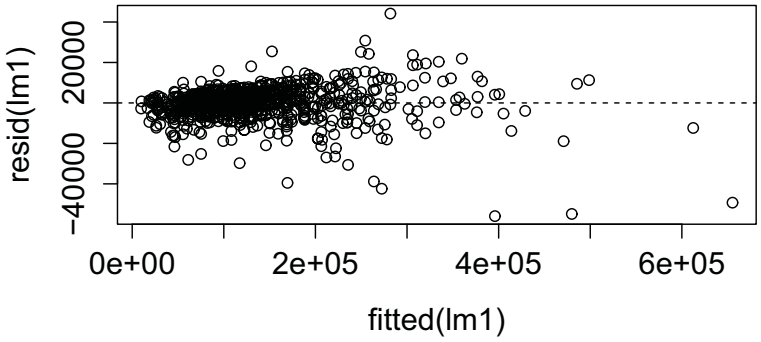
2.51 a. The scatterplot shows a clear positive linear trend, but the relationship is not strong. The correlation is only 0.38.

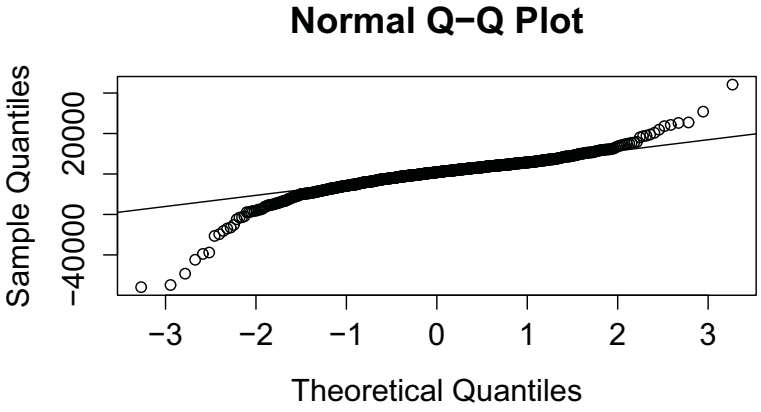


b. The fitted regression model summary is given below. The predicted equation is $\widehat{Survey2}=40.417+0.395Survey1$. The slope is clearly greater than zero.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.41661    4.43100   9.121  < 2e-16 ***
Survey1      0.39478    0.07004   5.637 6.07e-08 ***
---
Residual standard error: 9.338 on 193 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.1414,Adjusted R-squared:  0.1369
F-statistic: 31.77 on 1 and 193 DF,  p-value: 6.074e-08
```

c. The residual versus fitted values plot shows a good adherence to the linear trend with constant variance conditions, but the normal plot of residuals suggests a lack of normality in the error term.





d. To see if the $y = x$ model works, we look at both the intercept and the slope coefficients. Clearly we can infer that the intercept is statistically bigger than 0 (the 40.4166 and the P-value less than 2e-16). The slope is also clearly not 1, since with an estimated slope of 0.394 and a standard error of 0.07, 1 would be far outside even a very high confidence interval.

2.52 a. 95% confidence interval for slope:

$$0.9431 \pm 1.962526(0.003201)$$

$$0.9431 \pm 0.006282046$$

$$0.9431 \pm 0.0063$$

So we are 95% confidence that the true slope is between 0.9368 and 0.9494.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.448e+02  5.236e+02  -0.277    0.782
ListPrice    9.431e-01  3.201e-03 294.578   <2e-16 ***
---
Residual standard error: 8019 on 927 degrees of freedom
Multiple R-squared:  0.9894,Adjusted R-squared:  0.9894
```

b. We can test the null hypothesis that the true intercept is 0 by reading and interpreting the (Intercept) line of the summary output. The P-value in that line tests this very null and with a large P-value of 0.782, we have no evidence to reject the null; it is tenable that the true intercept is 0.

c. We define a variable as $fraction = SalePrice/ListPrice$. We then perform a one-sample t-test on *fraction* and the software gives output that includes the desired confidence interval; the output is given as follows. From this we are 95% confident that the mean ratio of sale price to list price is between 0.9311 and 0.9401. This interval is shifted to the left of the interval found in (a), probably because the non-significant-from-zero intercept is still estimated to a negative intercept which creates a line steeper (slightly) than the average fraction.

```
One Sample t-test

data:  fraction
t = 410.1117, df = 928, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.9311611 0.9401158
sample estimates:
mean of x
0.9356384
```

2.53 a. Software gives an interval of (64.13893, 67.68782). Cubing the endpoints of the interval gives ($263,854, $310,121).

b. This interval is wider than the one from the untransformed model.

2.54 a. Using software, we compute the correlations between all pairs of quantitative variables in the **BaseballTimes2017** dataset.

```
              Runs  Margin Pitchers Attendance   Time
Runs         1.000 -0.1776   0.7272     0.1252  0.745
Margin      -0.178  1.0000   0.0653    -0.2825 -0.165
Pitchers     0.727  0.0653   1.0000    -0.0222  0.648
Attendance   0.125 -0.2825  -0.0222     1.0000  0.319
Time         0.745 -0.1647   0.6478     0.3187  1.000
```

Looking across the last row, we see that the strongest correlation with $Time$ is number of $Runs$, $r = 0.745$.

b. Some output for predicting game $Time$ based on number of $Runs$ is shown below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   148.04      12.00   12.34  3.5e-08 ***
Runs            4.18       1.08    3.87   0.0022 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1  Residual standard error: 15.3 on
Multiple R-squared:  0.555,Adjusted R-squared:  0.518
F-statistic:   15 on 1 and 12 DF,  p-value: 0.00224
```

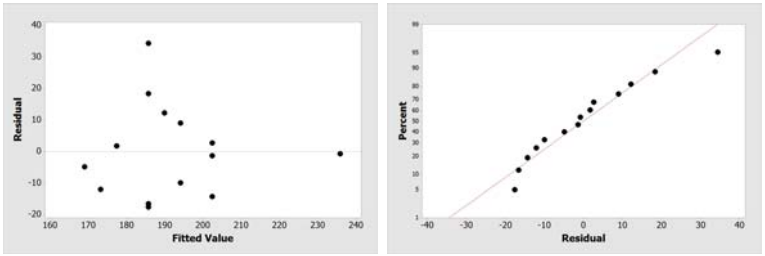
The prediction equation is $\widehat{Time} = 148.04 + 4.18Runs$. The slope of 4.18 means that for every extra run scored in a game the expected game time increases by about 4.2 minutes.

c. We use the hypotheses $H_0 : \rho = 0$ versus $H_a : \rho = 0$, where ρ is the correlation between $Time$ and $Runs$ for all (major league) baseball games. The correlation of $r = 0.745$ and sample size of 14 games mean that the t-statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.745\sqrt{14-2}}{\sqrt{1-0.745^2}} = 3.87$$

The P-value $= 2P(t_{12} > 3.87) = 0.002$, so we reject H_0 and conclude that there is a significant (positive) correlation between game $Time$ and number of $Runs$.

d. Two plots of the residuals are shown below. There is no pattern in the plot of residuals versus fitted values; however, the normal quantile plot shows a departure from normality. The upward curvature suggests a long right-hand tail for the distribution of the residuals. This calls into question the trustworthiness of the t-test from part (c). But the t-test is robust and the P-value is very small, so we can still be confident that $Time$ and $Runs$ are positively correlated.



2.55 a. Here are the correlations (with P-value below) for each of the potential predictors of $Nfrass$.

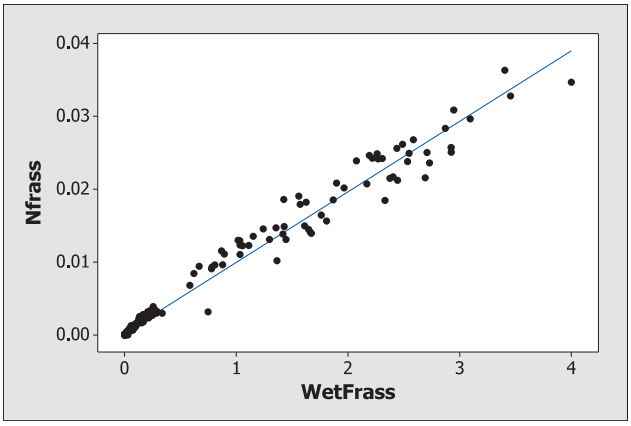
```
          Mass     Intake    WetFrass   DryFrass   Cassim   Nassim
Nfrass   0.886      0.931       0.990      0.983    0.892    0.841
         0.000      0.000       0.000      0.000    0.000    0.000
```

The strongest correlation is between $Nfrass$ and $WetFrass$ with $r = 0.990$

b. Here is some output (and a scatterplot) for the model to predict $Nfrass$ based on $WetFrass$.

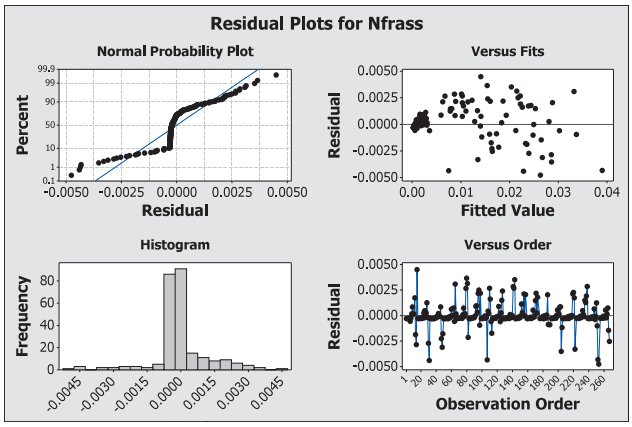
```
The regression equation is Nfrass = 0.000297 + 0.00967 WetFrass

Predictor        Coef     SE Coef        T      P
Constant   0.00029667  0.00008639     3.43  0.001
WetFrass   0.00967478  0.00008535   113.36  0.000

S = 0.00119672   R-Sq = 98.1%   R-Sq(adj) = 98.1%
```

The fitted line is $\widehat{Nfrass} = 0.000297 + 0.00967 WetFrass$.



c. We test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$ using the t-statistic (113.36) and P-value (essentially zero) from the regression output. This large t-statistic and very small P-value give very strong evidence of some relationship between the $Nfrass$ and $WetFrass$.

d. The coefficient of determination is 98.1%. Thus we are explaining a considerable amount of variation in $WetFrass$. There are a few concerns with the regression conditions. The histogram of the residuals is centered at zero, with some unusually small residuals. These unusually small values are also obvious on the normal probability plot, where the clear departures from a linear trend in the lower tail indicates a lack of normality in the residuals. We also see small variability in the residuals for small predicted values of $Nfrass$ that increases somewhat for larger predicted values. But this is a fairly large sample size and the scatterplot in part (b) shows that the linear model does a good job of summarizing the general trend in this relationship.



2.56 a. Here are the correlations (with P-value below) for each of the potential predictors of $Cassim$.

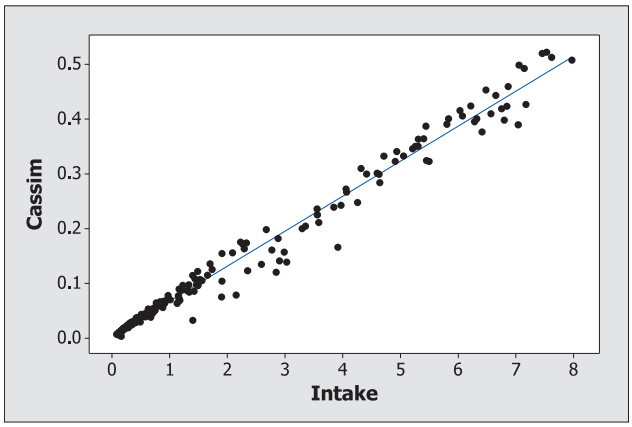
	Mass	Intake	WetFrass	DryFrass	NFrass	Nassim
Cassim	0.681	0.993	0.872	0.929	0.892	0.992
	0.000	0.000	0.000	0.000	0.000	0.000

The strongest correlation is between $Cassim$ and $Intake$ with $r = 0.993$

b. Here is some output (and a scatterplot) for the model to predict $Cassim$ based on $Intake$.

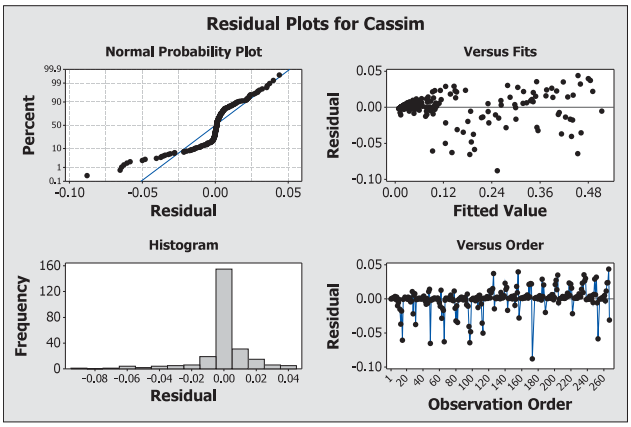
```
The regression equation is Cassim = 0.00379 + 0.0639 Intake

Predictor       Coef    SE Coef       T      P
Constant    0.003787   0.001317    2.88  0.004
Intake     0.0639029  0.0004908  130.21  0.000

S = 0.0165365   R-Sq = 98.5%   R-Sq(adj) = 98.5%
```

The fitted line is $\widehat{Cassim} = 0.0038 + 0.0639 Intake$.

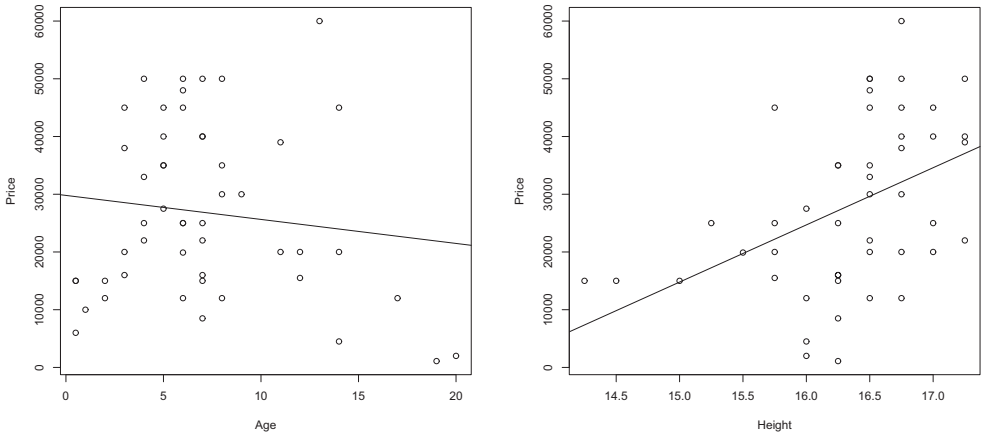


c. We test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ using the t-statistic (130.21) and P-value (essentially zero) from the regression output. This large t-statistic and very small P-value give very strong evidence of some relationship between the $Cassim$ and $Intake$.

d. The coefficient of determination is 98.5%. Thus we are explaining a considerable amount of variation in $Cassim$. There are a few concerns with the regression conditions. The histogram of the residuals is centered at zero, with some unusually small residuals. These unusually small values are also obvious on the normal probability plot, where the clear departures from a linear trend in the lower tail indicates a lack of normality in the residuals. We also see small variability in the residuals for small predicted values of $Cassim$ that increases somewhat for larger predicted values. But this is a fairly large sample size and the scatterplot in part (b) shows that the linear model does a good job of summarizing the general trend in this relationship.



2.57 We first consider scatterplots (with regression lines) for each of the potential predictors, Age and $Height$, of the response variable $Price$ of a horse.



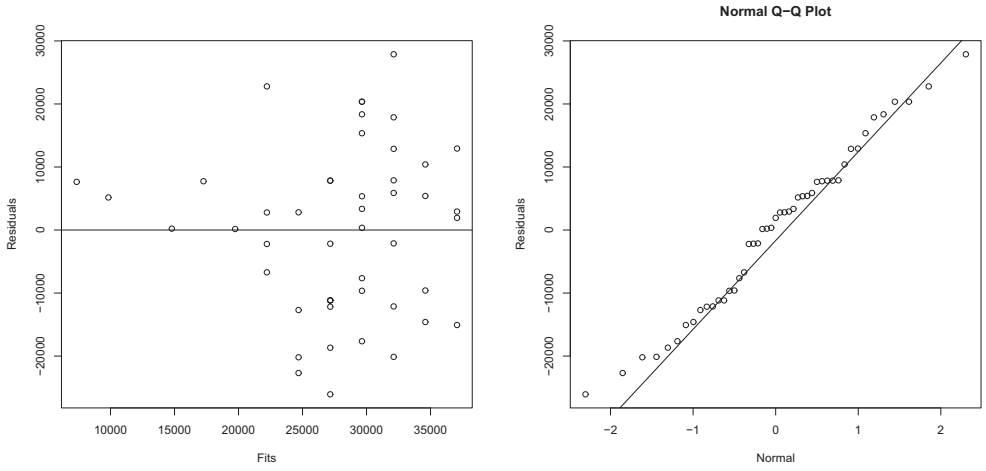
Age shows a weak (if any) linear relationship ($r = -0.127$) with *Price*, while *Height* shows a stronger association ($r = 0.443$) with *Price*. If we are using a single one of these predictors for all of the horses, *Height* would be the better option.

Some regression output for fitting the linear model to predict *Price* based on *Height* is shown below.

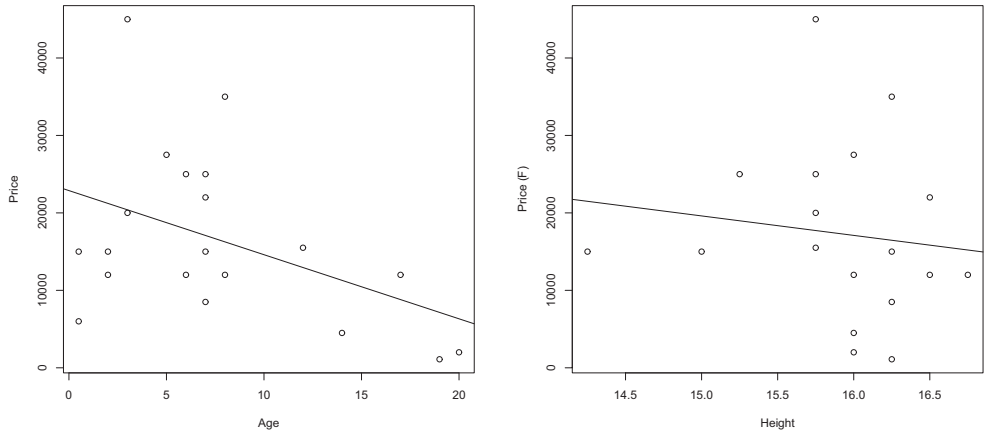
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -133791      48817  -2.741  0.00876 **
Height          9905       2987   3.316  0.00181 **
```

The estimated linear fit is $\widehat{Price} = -133791 + 9905 Height$. The small P-value (0.00181) for the t-test of slope indicates that there is strong evidence for a positive relationship between horse *Price* and *Height*.

A plot of residuals versus fits for the model based on *Height* shows a reasonably good scatter around the zero line, although there are very few horses with small predicted prices (below $20,000) and these all have positive residuals. A normal quantile plot of the residuals shows a relatively consistent linear trend, which gives support for the normality condition.



Female horses:
Most of the smaller (height) horses are female, so we consider a model for just the data on 20 female horses. Scatterplots now indicate a stronger relationship between *Price* and *Age* ($r = -0.439$) than *Height* ($r = -0.132$).

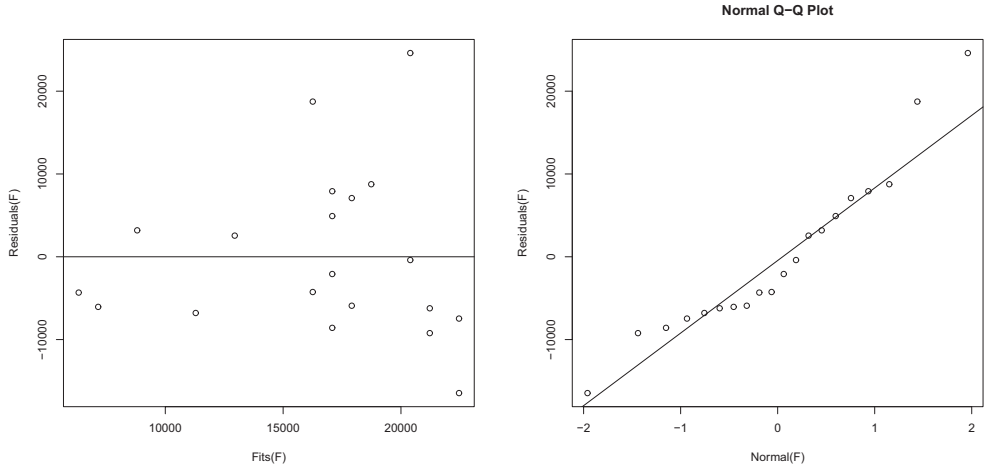


Fitting the regression for *Price* based on *Age* for the female horses gives

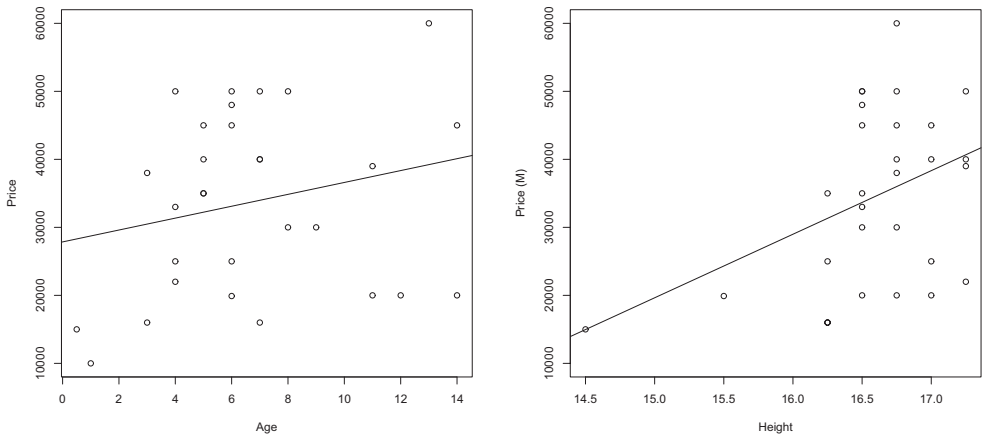
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  22877.8     3827.5   5.977 1.18e-05 ***
Age           -827.6      399.3  -2.073   0.0528
```

The estimated linear fit is $\widehat{Price} = 22877.8 - 827.6Age$. The P-value (0.0528) for the t-test of slope is not (quite) significant at a 5% level, leaving some doubt still about the strength of this linear relationship (although it's stronger than the *Price* versus *Height* relationship for female horses).

The residual versus fits plot for the *Price* versus *Age* model for female horses shows a reasonably random scatter on either side of the zero line with a possibility of variability increasing as the predicted prices increase. We see no serious concerns with lack of normality in the normal quantile plot for the residuals.



Male horses:
Data for the male horses alone looks more like the combined data; a stronger positive relationship between $Price$ and $Height$ ($r = 0.409$) and a smaller (but now positive) slope between $Price$ and Age ($r = 0.232$). We might have some concern about influential points with the two horses that are smaller in $Height$ than the rest of the male horses.

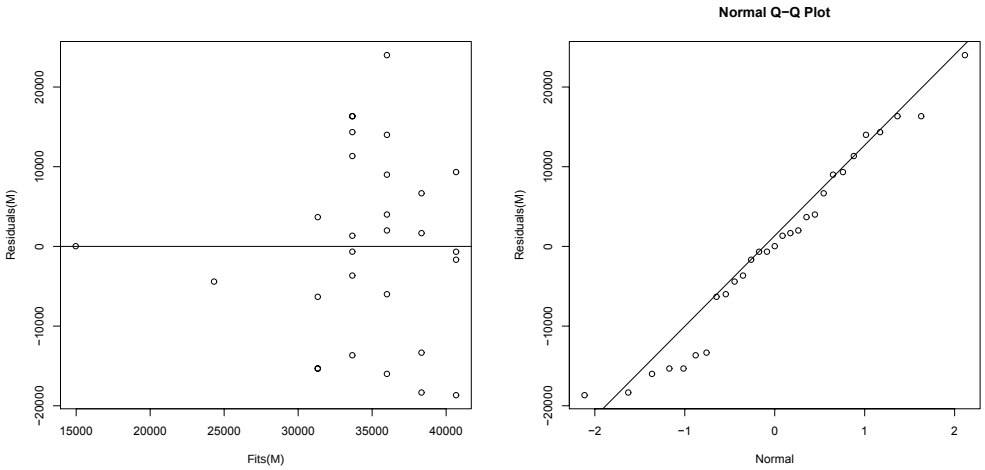


Fitting the regression for $Price$ based on$Height$ for the male horses gives

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -120546      66679  -1.808   0.0818 .
Height          9346       4016   2.327   0.0277 *
```

The estimated linear fit is $\widehat{Price} = -120,546 + 9346Height$. The P-value (0.0277) for the t-test of slope is significant at a 5% level, indicating some positive association between $Price$ and $Height$ for male horses. Note that the P-value is smaller (0.00181) for the combined data, but that model is based on a larger sample size and includes the smaller, less expensive female horses.

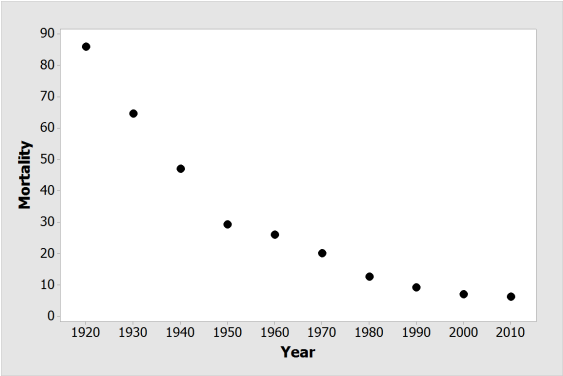
The residual versus fits plot for the $Price$ versus $Height$ model for male horses shows a reasonably random scatter on either side of the zero line. We note that the outlier small horse is predicted quite accurately, possibly a consequence of it having influence on the slope of the regression line. We see no serious concerns with lack of normality in the normal quantile plot for the residuals.



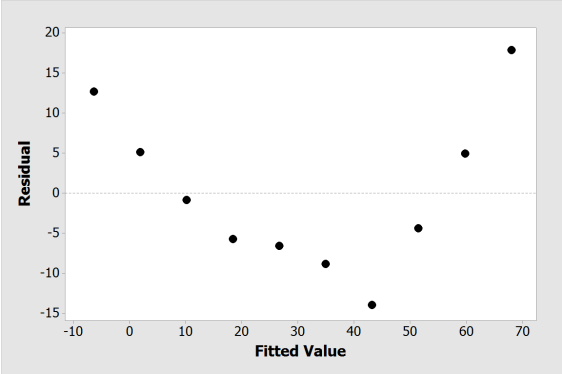
Given that the separate analyses show that Age is a stronger predictor of $Price$ for female horses (and in fact the correlation between $Price$ and $Height$ for the sample of female horses is negative), while $Height$ has a stronger (and positive) association with $Price$ for the sample of male horses, we should use the separate regression equations to predict the $Price$ of a horse, depending on its sex. Also the estimated standard deviation of the regression in the separate models (females, $s_\epsilon = 10,194$ and males, $s_\epsilon = 11,830$) is smaller than for the combined data ($s_\epsilon = 13,363$).

$\widehat{Price} = 22,877.8 - 827.6Age$ (for female horses)
$\widehat{Price} = -120,546 + 9346Height$ (for male horses)

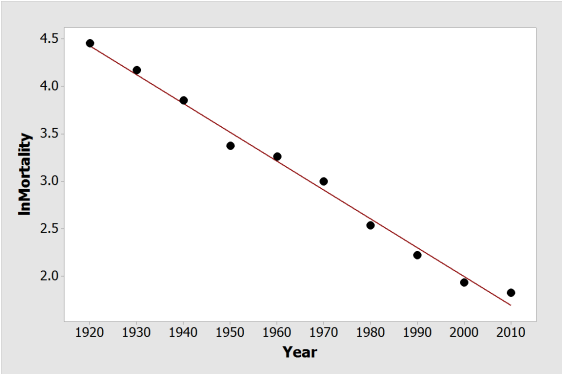
2.58 a. The scatterplot of $Mortality$ versus $Year$ shows a consistent decreasing trend, but the relationship is curved rather than linear.

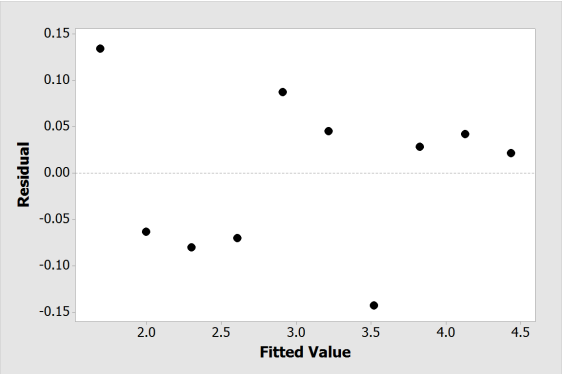


b. The prediction equation is $\widehat{Mortality} = 1656.46 - 0.827Year$. A plot of residuals versus fitted values illustrates the curved nature of the relationship. The linearity condition for a linear model is not satisfied.



c. A log transform for $Mortality$ to $lnMortality = log(Mortality)$ provides a more linear relationship with $Year$ as seen in the scatterplot and residual plot below. The curved pattern is no longer apparent in either of these plots.





d. The hypotheses are $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. Some computer output for fitting this model to $lnMortality$ is given below.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.8951     1.9734    31.9  1.0e-09 ***
Year         -0.0305     0.0010   -30.3  1.5e-09 ***
```

The test statistic is $t = -30.3$ and the P-value is $1.5 \cdot 10^{-9}$, or roughly zero, so the data suggest that there is strong evidence that $log(Mortality)$ (and thus $Mortality$ also) has gone down over time.

e. The estimated value for $Year = 2020$ is $\widehat{lnMortality} = 62.89 - 0.0305(2020) = 1.38$. We convert this to an estimate for $Mortality$ with $\widehat{Mortality} = e^{1.38} = 3.97$, or about 4 deaths per 1000 infants in 2020. We can also use technology to find a 95% prediction interval for $lnMortality$ when $Year = 2020$.

```
       fit     lwr    upr
1 1.38598 1.13127 1.6407
```

We exponentiate the endpoints of the interval to produce a prediction interval for $Mortality$ in 2020 that goes from $e^{1.1313} = 3.10$ to $e^{1.6407} = 5.16$.

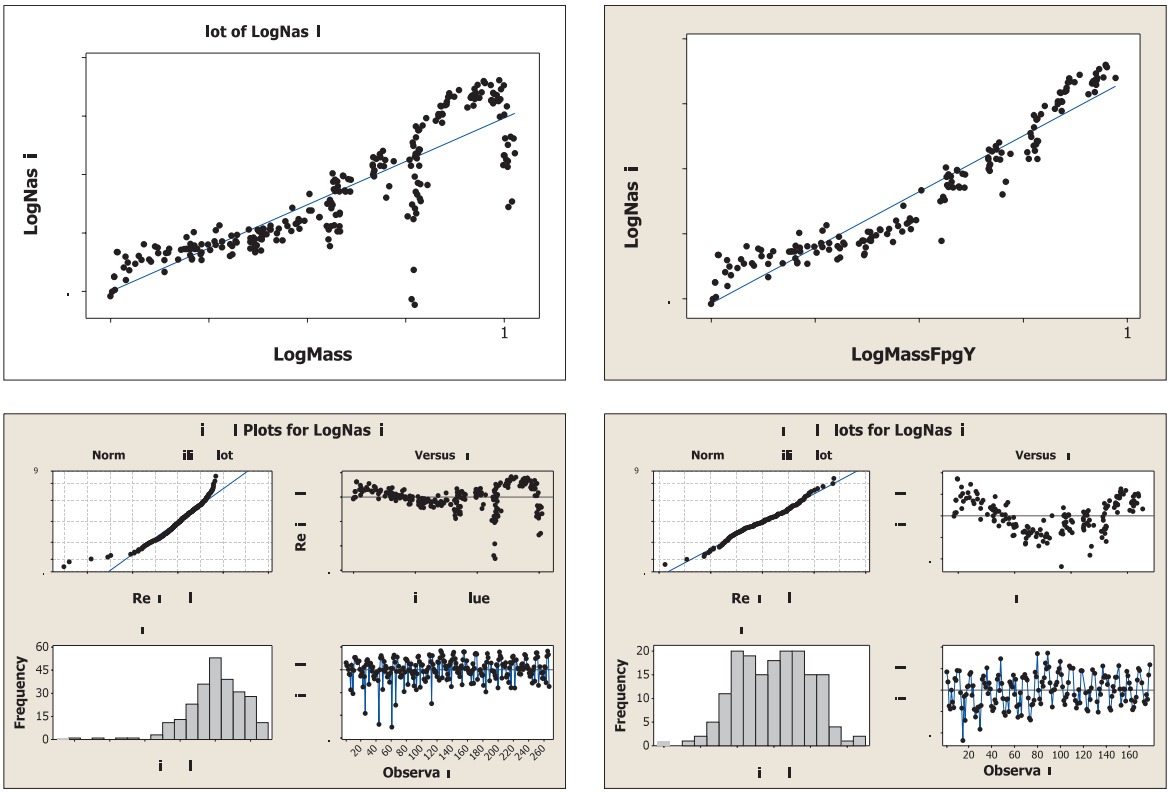
2.59 Following is some output for fitting a linear model to predict $LogNassim$ using $LogMass$ based on the entire sample of caterpillars.

```
The regression equation is LogNassim = - 1.89 + 0.371 LogMass

Predictor      Coef  SE Coef        T      P
Constant   -1.88738  0.01841  -102.53  0.000
LogMass     0.37096  0.01332    27.85  0.000

S = 0.250145   R-Sq = 75.5%   R-Sq(adj) = 75.5%
```

Here is corresponding output for fitting the model using only the caterpillars in the free growth period ($Fgp =$ "Y").

```
The regression equation is LogNassimFgpY = - 1.75 + 0.430 LogMassFgpY

Predictor        Coef   SE Coef        T      P
Constant     -1.74642   0.01489  -117.27  0.000
LogMassFgpY  0.429738  0.009946    43.21  0.000

S = 0.148628   R-Sq = 91.8%   R-Sq(adj) = 91.7%
```

We see that when the model is restricted to just caterpillars in the free growth period, the percent of variability in $LogNassim$ that is explained by the model goes up from 75.5% to 91.8%. Furthermore, the scatterplots and residual plots using the entire data (on the left in the following image) show more concerns with the conditions (especially several places with unusually low values of $LogNassim$ and much greater variability) than the plots (on the right) for the free growth caterpillars. While both relationships exhibit a bit of curvature in the residual versus fits plots, the normality plots look much better when the data include only caterpillars in the free growth period. (Recall that we examined a scatterplot of this relationship with different plotting symbols back in Chapter 1.)



2.60 a. Since the interval bounds include most (if not all) of the data values, they are most likely prediction intervals for individual values, rather than confidence intervals for the mean values.

b. Since the P-value is not very small (above a 5% significance level), there is not strong evidence against an assumption that the data are from a normal distribution. So a normality condition would be reasonable for these data.

c. The points in the normal probability plot follow the linear trend reasonably well, which is consistent with the conclusion of the test that the data could reasonably come from a normal distribution.

2.61 a. Based on the sample statistics and formulas for the slope and intercept, we have

$$\text{Slope:} \qquad \hat{\beta}_1 = r\frac{s_y}{s_x} = 0.701\left(\frac{104{,}807}{657}\right) = 111.826$$

$$\text{Intercept:} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 247{,}235 - 111.826(2009) = 22{,}576.6$$

This gives a fitted line of $\widehat{Gate} = 22{,}576.6 + 111.826Enroll$.

b. The coefficient of determination is $r^2 = 0.701^2 = 0.491$, so the enrollments can explain about 49.1% of the variation in the gate counts.

c. For a college with an enrollment of 1445, the predicted gate count is

$$\widehat{Gate} = 22{,}576.6 + 111.826(1445) = 184{,}165.2$$

d. The predicted value for a school with 2200 students is

$$\widehat{Gate} = 22{,}576.6 + 111.826(2200) = 268{,}593.8$$

If the actual gate count is 130,000 the residual is

$$Gate - \widehat{Gate} = 130{,}000 - 268{,}593.8 = -138{,}593.8$$

2.62 a. The R command plot(Calories Sugar, data=Cereal) shows that there is an upward trend between *Calories* and *Sugar*. The command cor(Calories,Sugar, data=Cereal) gives the correlation as 0.515.

b. The R command cor.test(Sugar+Calories, data=Cereal) gives the P-value as 0.0013 and the 95% CI as (0.225,0.722).

c. The P-value of 0.0013 is fairly close to the probability that the true correlation is negative.

d. The 95% CI is (0.225, 0.722), which is similar to the 95% credible interval of (0.21, 0.72.)