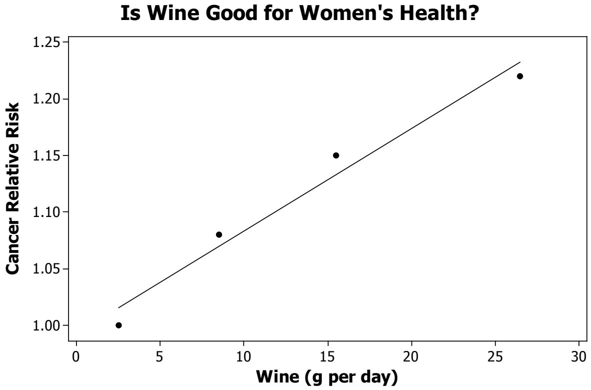
Chapter 26 – Inference for Regression

26.1 (a) A scatterplot of the data is provided, along with the least-squares regression line (students were not asked to add the line). We see that there is a strong, positive, linear relationship between wine intake and relative risk. From software, the correlation is $r = \sqrt{0.97} = 0.985$.



(b) If we knew it, the slope β would tell us how much the relative risk of breast cancer changes in women for each increase of 1 gram of wine per day (on average). The estimate of β is b = 0.009012 (see output provided). We estimate that an increase in intake of 1 gram per day increases relative risk of breast cancer by about 0.009. The estimate of α is a = 0.9931. According to our estimate, wine intake of 0 grams per day is associated with a relative risk of breast cancer of 0.9931 (about 1).

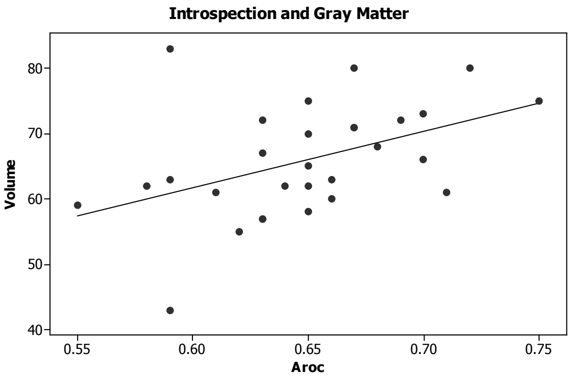
Regression Analysis: Risk versus Wine

```
The regression equation is Risk = 0.993 + 0.00901 Wine

 Predictor       Coef    SE Coef       T       P
 Constant     0.99309    0.01777   55.88   0.000
 Wine       0.009012   0.001112    8.10   0.015

S = 0.0198583   R-Sq = 97.0%   R-Sq(adj) = 95.6%
```

(c) The least-squares regression line is given by $\hat{y} = 0.9931 + 0.009x$. The provided table summarizes computed residuals, which sum to zero, as demonstrated. We also have $s^2 = 0.00079/2 = 0.000395$, which provides an estimate of σ^2. We estimate σ by $s = \sqrt{\frac{0.00079}{4-2}} = 0.01987$, which agrees (up to roundoff error) with S in the output given.

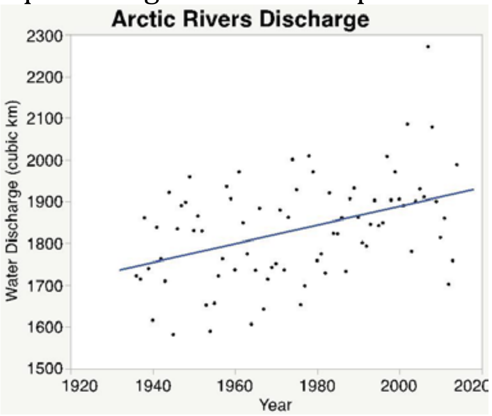
			Residual	
x	y	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$
2.5	1.00	1.0156	−0.0156	0.00024
8.5	1.08	1.0697	0.0103	0.00011
15.5	1.15	1.1328	0.0172	0.00030
26.5	1.22	1.2319	−0.0119	0.00014
			0	0.00079

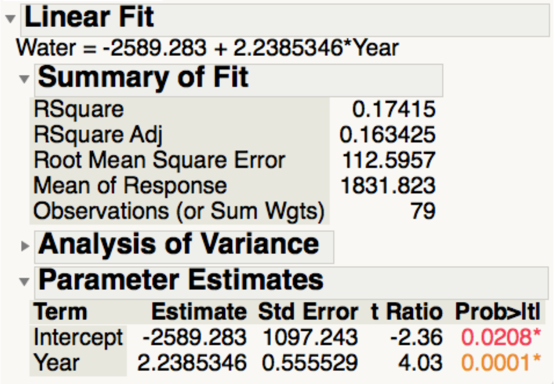
26.2 (a) A scatterplot of the data is provided, with the least-squares regression line added, which is asked for in part (c). From the output provided, $r^2 = 0.2006$. Our model explains about 20.1% of the observed variability in brain volume.



(b) Reading directly from the output in Figure 26.4, we estimate α by a = 10.0655. We estimate β by b = 86.0308. We estimate σ by $s = 7.9088$. Units are not defined in the problem. **(c)** The least-squares regression line is given by $\hat{y} = 10.0655 + 86.0308x$. This line has been added to the scatterplot in part (a).

26.3 (a) A scatterplot of discharge by year is provided, along with the fitted regression line, which is requested in part (b). Discharge seems to be increasing over time, but there is also a lot of variation in this trend. From the JMP output provided, $r^2 = 0.174$, so the least-squares regression line explains 17.4% of the total observed variability in Arctic discharge.





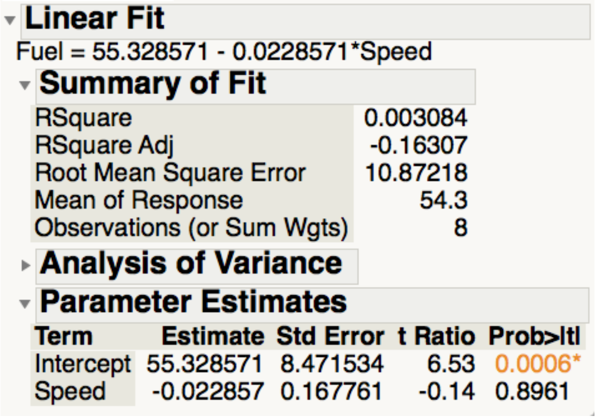
```
Linear Fit
Water = -2589.283 + 2.2385346*Year
  Summary of Fit
  RSquare                       0.17415
  RSquare Adj                  0.163425
  Root Mean Square Error       112.5957
  Mean of Response             1831.823
  Observations (or Sum Wgts)         79
  Analysis of Variance
  Parameter Estimates
  Term       Estimate  Std Error  t Ratio  Prob>|t|
  Intercept  -2589.283  1097.243    -2.36   0.0208*
  Year      2.2385346   0.555529     4.03   0.0001*
```

(b) The regression line has been added to the scatterplot provided in part (a). The least-squares regression line is given by $\hat{y} = -2589.283 + 2.239x$. We see from the output given that $s = 112.596$.

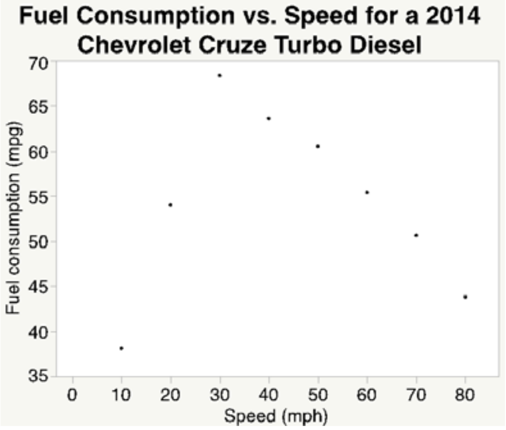
26.4 (a) $t = \frac{b}{SE_b} = \frac{0.009012}{0.001112} = 8.1$. **(b)** The sample size is $n = 4$, so df = 4 – 2 = 2. Using Table C, for a one-sided alternative, 0.005 < P < 0.01 (P = 0.0074 from technology; the P-value given in the Minitab output for Exercise 26.1 uses a two-tailed alternate). These data indicate strongly that breast cancer relative risk increases with additional wine consumption. Don't forget that these data involved averaging over individuals. The data points provided are averages at each value of wine intake. No doubt, there would be far more variation between individuals, and the statistical significance of the results will not be as strong for individuals as for averages.

26.5 Refer to the output provided with the solution to Exercise 26.3. We test $H_0\colon \beta = 0$ versus $H_\text{a}\colon \beta > 0$. We compute $t = \frac{b}{SE_b} = \frac{2.2385}{0.5555} = 4.03$. Here, df = n – 2 = 79 – 2 = 77. In referring to Table C, we round df down to df = 60. Using Table C, we obtain P < 0.0005. Using software, we obtain P = 0.000 (rounded to three decimal places). There is strong evidence of an increase in Arctic discharge over time.

26.6 The JMP output is provided. We have $H_0\colon \beta = 0$ versus $H_\text{a}\colon \beta \neq 0$. We observe $t = \frac{b}{SE_b} = \frac{-0.0229}{0.1678} = -0.14$. We clearly will not reject the null hypothesis. With df = 8 – 2 = 6, using Table C, we obtain P > 0.50. From software, P = 0.896. There is little evidence of a straight-line relationship between fuel use and speed. However, examining the provided scatterplot, we see that there is, in fact, a very strong (nonlinear) relationship between speed and fuel use. One should always plot data before performing a regression.



```
▾ Linear Fit
Fuel = 55.328571 - 0.0228571*Speed
  ▾ Summary of Fit
    RSquare                      0.003084
    RSquare Adj                  -0.16307
    Root Mean Square Error       10.87218
    Mean of Response                 54.3
    Observations (or Sum Wgts)          8
  ▸ Analysis of Variance
  ▾ Parameter Estimates
    Term       Estimate  Std Error  t Ratio  Prob>|t|
    Intercept  55.328571  8.471534     6.53   0.0006*
    Speed      -0.022857  0.167761    -0.14   0.8961
```



26.7 (a) Refer to the solution of Exercise 26.4. For testing H_0: $\beta = 0$ versus H_a: $\beta > 0$, we have t = 8.1 with df = 2. For the one-sided alternative suggested, we obtained 0.005 < P < 0.01. This test is equivalent to testing H_0: population correlation = 0 versus H_a: population correlation > 0. **(b)** Using software, r = 0.985. This can also be computed by referring to the Minitab output provided with Exercise 26.1, with $r = +\sqrt{r^2} = +\sqrt{0.97}$. Referring to Table E with n = 4, we find that 0.005 < P < 0.01, just as in part (a). These tests are equivalent.

26.8 Refer to the scatterplot provided with the solution to Exercise 4.47. We have $r =$ 0.878 and n = 13. Testing H_0: population correlation = 0 versus H_a: population correlation > 0, using Table E, P < 0.0005. There is overwhelming evidence of a positive linear relationship between social distress score and activity in the part of the brain known to be activated by physical pain.

26.9 Referring to Table C, t^* = 2.920 (df = 4 – 2 = 2, with 90% confidence). A 90% confidence interval for β is given by 0.009012 $\pm$ 2.920(0.001112) = 0.009012 $\pm$ 0.003247 = 0.00577 to 0.01226. With 90% confidence, the expected increase in relative risk of breast cancer associated with an increase in alcohol consumption by 1 gram per day is between 0.00577 and 0.01226.

26.10 There are n = 29 observations, so df = 29 – 2 = 27. From Table C, t^* = 2.052. From the output provided in Figure 26.4, b = 86.030829 and SE_b = 33.04842. A 95% confidence interval for the increase in Aroc per unit increase in volume is given by 86.030829 $\pm$ 2.052(33.04842) = 18.22 to 153.85. With 95% confidence, each unit increase in introspective ability (as measured by Aroc) is associated with an increase of between 18.22 and 153.85 units of gray-matter volume (as measured by Brodmann area).

26.11 Refer to the output provided in the solution to Exercise 26.3. We have b = 2.2385 and SE_b = 0.5555. With 79 observations, df = 77. Using Table C, we look under the row corresponding to df = 60 (the nearest smaller value of df in the table). We obtain t^* = 1.671 (t^* = 1.665 from software). A 90% confidence interval for β is given by 2.2385 $\pm$ 1.671(0.5555) = 2.2385 $\pm$ 0.9282 = 1.3103 to 3.1667 cubic kilometers per year (software: 1.3136 to 3.1634). With 90% confidence, the yearly increase in Arctic discharge is between 1.3103 and 3.1667 cubic kilometers. This confidence interval excludes zero, so there is evidence that Arctic discharge is increasing over time.

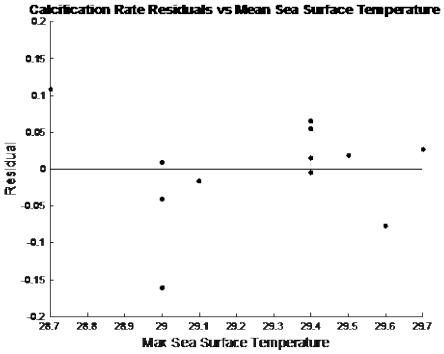
26.12 (a) If we wish to predict the relative risk of breast cancer for a single group of women for which $x^* = 10$, then we should use a prediction interval. This is denoted "95% PI" in the output shown in Figure 26.8 of the text. With 95% confidence, the relative risk of breast cancer for an individual woman drinking 10 grams of red wine per day is 0.98643 to 1.18000. **(b)** From the output shown in Figure 26.8 of the text, $\hat{\mu} = 1.08321$ and $SE_{\hat{\mu}} = 0.01057$. For df = 4 – 2 = 2 and 90% confidence, t^* = 2.920. A 90% confidence interval for the mean relative risk of breast cancer in all women drinking 10 grams of red wine per day is 1.08321 $\pm$ 2.920(0.01057) = 1.052 to 1.114.

26.13 (a) If $x^* = 0.65$, then our prediction for mean volume is $\hat{\mu} = 10.0655 + 86.0308(0.65) = 65.98552$. **(b)** We have $SE_{\hat{\mu}} = 1.47$. For df = 29 – 2 = 27 and 95% confidence, we have t^* = 2.052. A 95% confidence interval for mean brain gray-matter volume in people with 0.65 Aroc is given by 65.98552 $\pm$ 2.052(1.469) = 62.971 to 69.000.

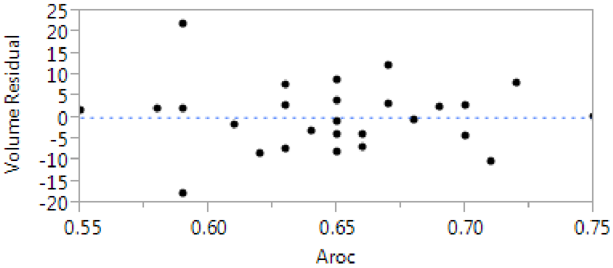
26.14 (a) A stemplot of the residuals is provided, where –1|6 represents –0.16. The distribution of residuals appears close to Normal, with one outlier in the left tail (a residual of –0.16).

```
−1 | 6
−1 |
−0 | 8
−0 | 421
 0 | 1123
 0 | 56
 1 | 1
```

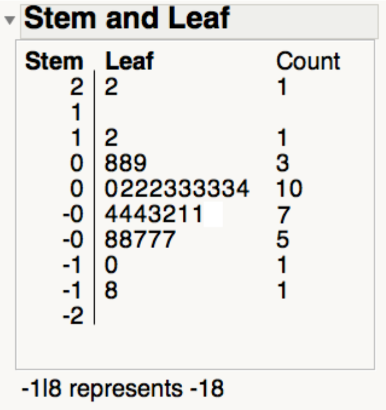
(b) The residual plot is also provided. There is not evidence of a clear deviation from a linear pattern; that is, there is no visible pattern in the residuals. While there appears to be roughly equal variability in the residuals about the "residual = 0" line, there is a slightly larger variance in residuals that have small maximum sea surface temperatures. (With such few observations, it is difficult to have "perfectly" equal variability.)



26.15 (a) The provided residual plot does not suggest any deviation from a straight-line relationship between brain volume and Aroc score, although there are two large (in absolute value) residuals near the left end of the plot. Both babies had Aroc scores of about 0.63, but one child's IQ was underpredicted (the positive residual) and one child's was overpredicted (the negative residual).



(b) The provided stemplot of residuals does not suggest that the distribution of residuals departs strongly from Normality. The value 22 from observation 5 may be an outlier; other than that, the residuals are symmetric and mound-shaped.



```
Stem and Leaf
Stem  Leaf        Count
   2 | 2          1
   1 |
   1 | 2          1
   0 | 889        3
   0 | 0222333334 10
  -0 | 4443211    7
  -0 | 88777      5
  -1 | 0          1
  -1 | 8          1
  -2 |

-1|8 represents -18
```

(c) It is reasonable to assume that the observations are independent, because we have 29 different subjects who are measured separately. **(d)** Other than the large residuals noted in part (a), there is no indication that variability changes; there are fewer babies with low Aroc scores, so there is naturally less variability on the left end of the plot.

26.16 (a) price $= 100.2 + 1.2186 \times$ appraised value. From the output provided, *a* = 100.2 and *b* = 1.2186.

26.17 (c) 0.766. With a positive association, $r = +\sqrt{r^2} = +\sqrt{0.587} = 0.766$.

26.18 (b) the average increase in selling price in a population of units when appraised value increases by $1000. Individual price increases vary, so answer option (c) is inappropriate. The population regression line provides all predicted mean prices, given appraised value.

26.19 (a) $H_0\colon \beta = 0$ versus $H_\mathrm{a}\colon \beta > 0$. This is a one-sided alternative because we wonder if larger appraisal values are associated with larger selling prices.

26.20 (c) less than 0.001. Note that the output shows *P* = 0.000 to three decimal places.

26.21 (c) 211.291. This is the value of *s*.

26.22 (c) 50. There are 52 observations, so df = 52 – 2 = 50.

26.23 (b) 1.2186 $\pm$ 0.2905. Using Table C and 50 degrees of freedom, t^* = 2.009, so the margin of error is 2.009(0.1446) = 0.2905.

26.24 (a) $646,500 and $1,503,700. The prediction interval is appropriate because she wants an interval for just her unit.

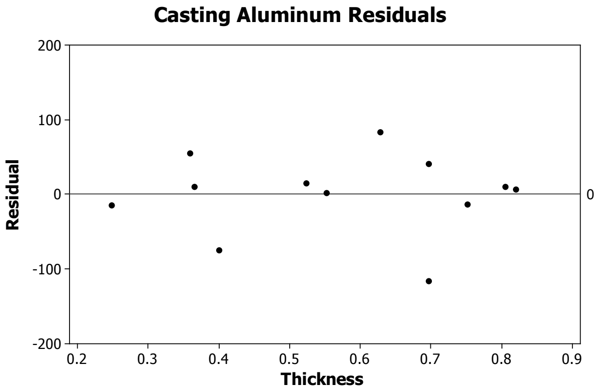
26.25 (a) Scientists estimate that each additional 1% increase in the percent of Bt cotton plants results in an average increase of 6.81 mirid bugs per 100 plants. **(b)** The regression model explains 90% of the variability in mirid bug density. That is, knowledge of the proportion of Bt cotton plants explains almost all of the variation in mirid bug density. **(c)** Recall that the test $H_0\colon \beta = 0$ versus $H_\mathrm{a}\colon \beta > 0$ is exactly the same as the test H_0: population correlation = 0 versus H_a: population correlation > 0. Because *P* < 0.0001, there is strong evidence of a positive linear relationship between the proportion of Bt cotton plants and the density of mirid bugs. **(d)** We may conclude that denser mirid bug populations are associated with larger proportions of Bt cotton plants. However, it seems plausible that a reduced use of pesticides (an indirect cause) instead of more Bt cotton plants (a direct cause) is the reason for this increase.

26.26 We test $H_0\colon \beta = 0$ versus $H_\mathrm{a}\colon \beta \neq 0$ and observe $t = \frac{b}{SE_b} = \frac{274.78}{88.18} = 3.116$ with df = 12 – 2 = 10. The two-sided *P*-value is between 0.01 and 0.02 (technology gives 0.0109). There is strong evidence of a linear relationship between thickness and gate velocity.

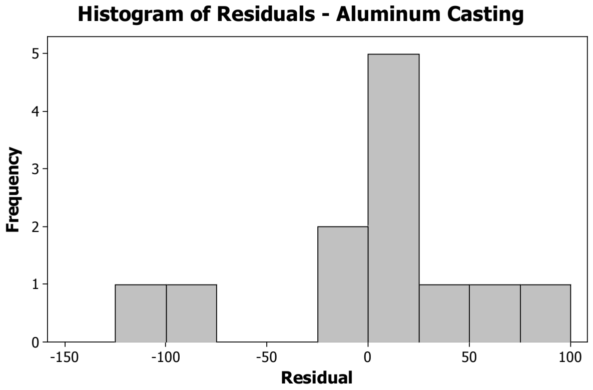
26.27 (For 90% intervals with df = 10, use t^* = 1.812.) **(a)** Use the estimated slope and standard error given in Figure 26.13. The confidence interval for β is $b \pm t^* SE_b = 274.78 \pm 1.812(88.18) = 274.78 \pm 159.78 = 115.0$ to 434.6 fps/inch. **(b)** This is the “90% CI” given in Figure 26.13: 176.2 to 239.4 fps. To confirm this, we can use the given values of $\hat{y} =$

207.8 and $SE_{\hat{\mu}} = 17.4$, labeled “Fit” and “SE Fit” in the output shown in Figure 26.13 of the text: $\hat{y} \pm t^* SE_{\hat{\mu}} = 207.8 \pm 1.812(17.4) = 176.3$ to 239.3 fps, which agrees with the output up to roundoff error.

26.28 (a) The provided scatterplot shows no obvious nonlinearity or change in variability.



(b) The provided histogram is unimodal and skewed to the left, and there are two very large, negative residuals (including observation 9, which is “flagged” in Figure 26.13).

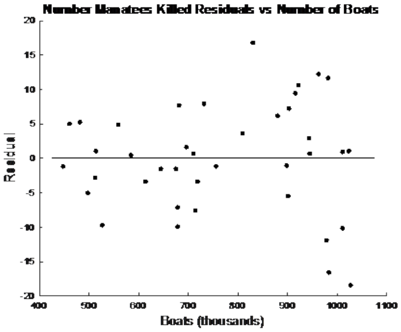


(c) Student opinions about whether this point is influential may vary. There are some changes, but they might not be considered substantial: the regression standard error is about 25% smaller and the prediction for *x* = 0.5 inch is about 8 fps larger, and the confidence interval is also narrower due to the reduced standard error.

26.29 (a) The provided stemplot confirms the comments from the text: there is little evidence of non-Normality in the residuals, and there don’t appear to be any outliers.

```
-1 | 87
-1 | 2000
-0 | 755
-0 | 333221110
 0 | 11111234
 0 | 55567889
 1 | 122
 1 | 7
```

(b) The provided scatterplot confirms the comments made in the text: There is no clear pattern, but the variability about the “residual = 0” line may be slightly greater when *x* is larger.



(c) Presumably, close inspection of a manatee’s corpse will reveal nonsubtle clues when the cause of death is from collision with a boat propeller. It seems reasonable that the kills are mostly not caused by pollution.

26.30 We test $H_0: \beta = 0$ versus $H_a: \beta > 0$. With df = 39 – 2 = 37, we have $t = \frac{b}{SE_b} = \frac{0.133051}{0.006954} = 19.133$, and P < 0.0005. There is overwhelming evidence that manatee kills increase with number of boats registered.

26.31 (a) This is a confidence interval for β. With df = 37, using Table C (and rounding degrees of freedom down to 30), we have t^* = 2.042, so a 95% confidence interval for β is $b \pm t^* SE_b = 0.133051 \pm 2.042(0.006954) = 0.133051 \pm 0.0142 = 0.11885$ to 0.14725 additional killed manatees per 1000 additional boats. (Using technology, df = 37, and t^* = 2.026, we have 0.11896 to 0.14714 additional killed manatees per 1000 additional boats.)
(b) With 900,000 boats, we predict $\hat{y} = -45.27 + 0.133051(900) = 74.4759$ killed manatees, which agrees with the output in Figure 26.14 under “Fit.” We need the prediction interval because we are forecasting the number of manatees killed for a single year. According to the output provided, a 95% prediction interval for the number of killed manatees is 58.06 to 90.89 kills if 900,000 boats are registered.

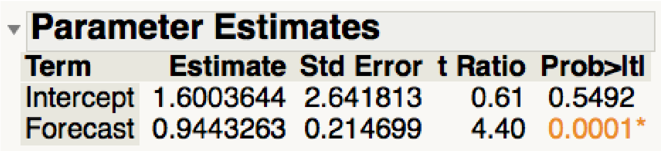
26.32 (a) The Minitab output for this analysis is provided. Recall that the test for $H_0: \beta = 0$ is equivalent to the test of H_0: population correlation = 0. We have $t = -4.64$ and $P <$ 0.0005 for testing $H_0: \beta = 0$ versus $H_a: \beta \neq 0$. For a one-sided test, P is half the size of Minitab’s value. For the test of H_0: population correlation = 0 against H_a: population correlation < 0, we have P < 0.00025. There is overwhelming evidence of a negative population correlation.

Regression Analysis: Fat versus NEA

```
The regression equation is Fat = 3.51 - 0.00344 NEA

 Predictor          Coef      SE Coef       T       P
 Constant         3.5051       0.3036   11.54   0.000
 NEA          -0.0034415    0.0007414   -4.64   0.000

S = 0.739853   R-Sq = 60.6%   R-Sq(adj) = 57.8%

 New Obs     Fit   SE Fit       95% CI            95% PI
       1   2.129    0.193   (1.714, 2.543)   (0.488, 3.769)
```

(b) To find this interval, we need SE_b, which is 0.0007414. With df = 14, t* = 1.761 for 90% confidence. A 90% confidence interval for β is $-0.0034415 \pm 1.761(0.0007414) = -0.0034415 \pm 0.00131 = -0.00475$ to −0.00213. **(c)** This question calls for a prediction interval. The Minitab output provided in part (a) gives the interval as 0.488 to 3.769 kg.

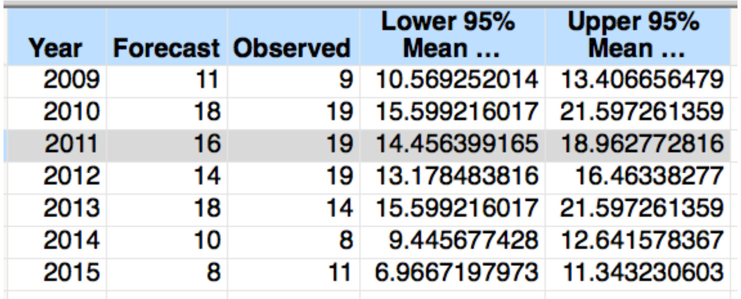
26.33 (a) We test H_0: population correlation = 0 against H_a: population correlation > 0; recall this is equivalent to a test for $\beta > 0$. We see that t = 4.40 with df = 32 – 2 = 30. So, the one-sided P-value is P = 0.0001/2 = 0.00005, using the provided JMP output. There is very strong evidence of a positive correlation between Gray's forecasted number of storms and the number of storms that actually occur.



Parameter Estimates

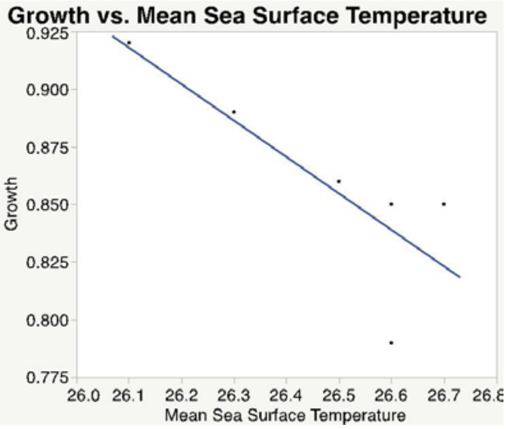
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.6003644	2.641813	0.61	0.5492
Forecast	0.9443263	0.214699	4.40	0.0001*

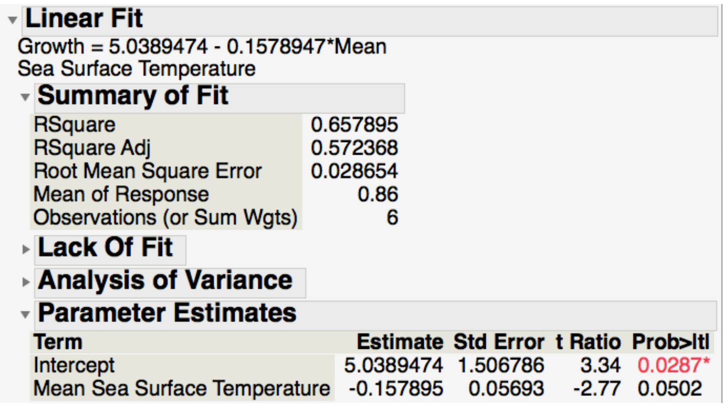
(b) The output provided gives the confidence interval for the mean number of storms in years for which Gray predicts 16 storms (use the line for 2011). Here $\hat{\mu} = 1.6004 + 0.9443(16) = 16.7092$ storms, and JMP gives the 95% confidence interval for the mean as 14.4564 to 18.9628 storms.



Year	Forecast	Observed	Lower 95% Mean ...	Upper 95% Mean ...
2009	11	9	10.569252014	13.406656479
2010	18	19	15.599216017	21.597261359
2011	16	19	14.456399165	18.962772816
2012	14	19	13.178483816	16.46338277
2013	18	14	15.599216017	21.597261359
2014	10	8	9.445677428	12.641578367
2015	8	11	6.9667197973	11.343230603

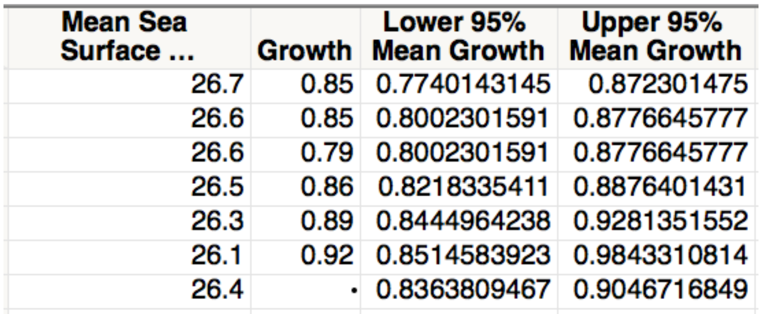
26.34 (a) The provided scatterplot reveals a fairly strong, negative, linear relationship between the mean sea surface temperature and coral growth. A formal test of $H_0: \beta = 0$ versus $H_a: \beta < 0$ reveals t = −2.77 with df = 6 – 2 = 4. The one-sided P-value is P = 0.0502/2 = 0.0251, using the provided JMP output. There is strong evidence of a negative linear relationship between the mean sea surface temperature and coral growth.





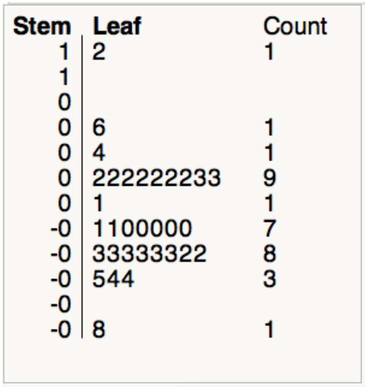
```
Linear Fit
Growth = 5.0389474 - 0.1578947*Mean
Sea Surface Temperature
Summary of Fit
RSquare                       0.657895
RSquare Adj                   0.572368
Root Mean Square Error        0.028654
Mean of Response                  0.86
Observations (or Sum Wgts)           6
Lack Of Fit
Analysis of Variance
Parameter Estimates
Term                          Estimate  Std Error  t Ratio  Prob>|t|
Intercept                     5.0389474  1.506786     3.34  0.0287*
Mean Sea Surface Temperature  -0.157895   0.05693    -2.77  0.0502
```

(b) The fit is $\hat{\mu} = 5.0389 - 0.1579(26.4) = 0.8703 \text{ cm/year}$. A 95% confidence interval for the mean coral growth per year when the mean sea surface temperature is 26.4°C is given by 0.8363 to 0.9047 cm/year (from the provided JMP output).



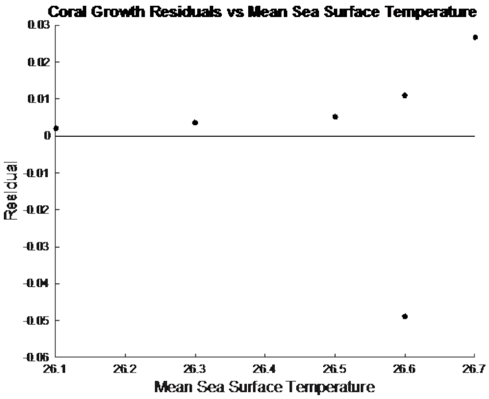
Mean Sea Surface ...	Growth	Lower 95% Mean Growth	Upper 95% Mean Growth
26.7	0.85	0.7740143145	0.872301475
26.6	0.85	0.8002301591	0.8776645777
26.6	0.79	0.8002301591	0.8776645777
26.5	0.86	0.8218335411	0.8876401431
26.3	0.89	0.8444964238	0.9281351552
26.1	0.92	0.8514583923	0.9843310814
26.4	•	0.8363809467	0.9046716849

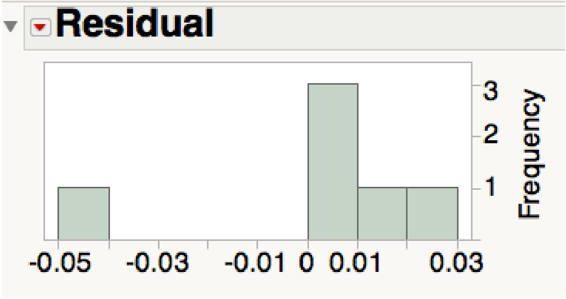
26.35 The stemplot is provided, where residuals are rounded to the nearest whole number. The plot suggests that the residuals may not follow a Normal distribution. Specifically, there is both a low outlier and a high outlier that seems extreme. This makes regression inference and interval procedures unreliable.



Stem	Leaf	Count
1	2	1
1		
0		
0	6	1
0	4	1
0	222222233	9
0	1	1
-0	1100000	7
-0	33333322	8
-0	544	3
-0		
-0	8	1

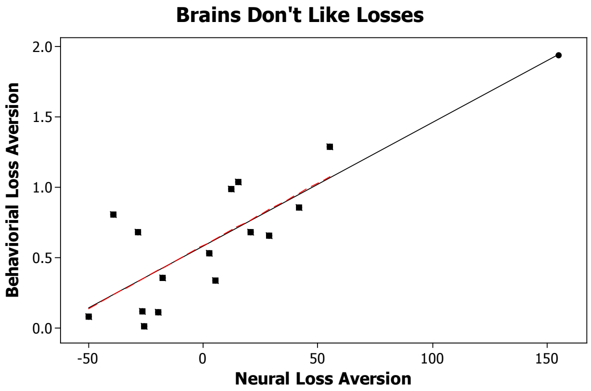
26.36 (The two requested plots are provided.)





(a) There is a potential outlier, but with only six observations, and considering the scatterplot provided in the solution to Exercise 26.34, there is no evidence of a systematic departure from nonlinearity in the relationship between mean sea surface temperature and coral growth. **(b)** There is some evidence (albeit difficult to detect with only six observations) that residuals are non-Normal. **(c)** It is not clear that observations are independent. For example, perhaps temperatures one year are correlated with temperatures the next year. **(d)** There appears to be a trend in the residual plot—there is small variation for residuals corresponding to small temperatures and large variation for residuals corresponding to large temperatures. However, the outlier may be causing this pattern to occur.

26.37 (a) Shown is the scatterplot with two (nearly identical) regression lines: one using all points and one with the outlier omitted. The Minitab output for both regression analyses is provided.



Regression output (all points)

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

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

S = 0.279729   R-Sq = 72.0%
```

Regression output (without outlier)

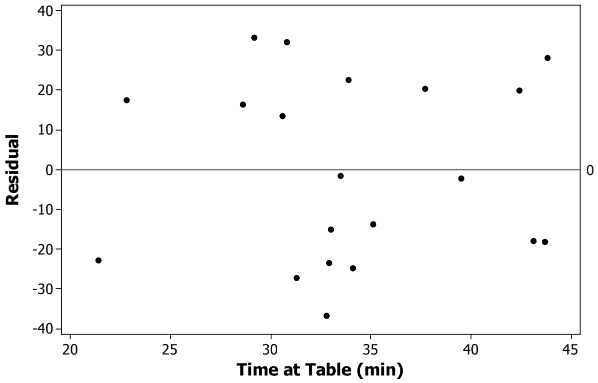
```
The regression equation is
Behave = 0.586 + 0.00891 Neural

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

S = 0.290252   R-Sq = 49.2%
```

(b) The correlation for all points is r = 0.8486. For testing the slope, t = 6.00, for which $P <$ 0.0005. **(c)** Without the outlier, r = 0.7014, the test statistic for the slope is t = 3.55, and P = 0.004. In both cases, there is strong evidence of a linear relationship between neural loss aversion and behavioral loss aversion. However, omitting the outlier weakens this evidence somewhat.

26.38 (a) The Minitab output and the scatterplot with superimposed regression line are shown. The regression equation is $\hat{y} = 560.65 - 3.0771x$, and the correlation is $r = -0.6492$. Generally, the longer a child remains at the table, the fewer calories he or she will consume. This relationship is moderately strong and linear.

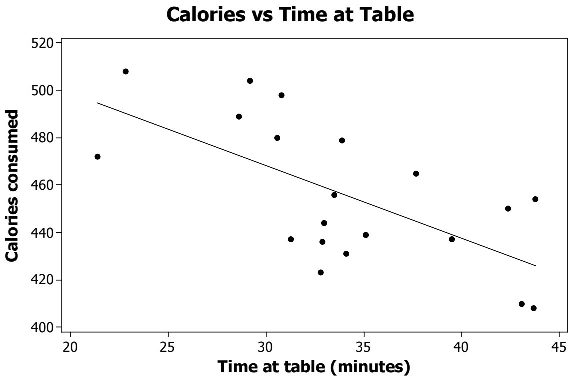
```
The regression equation is
Cal = 561 - 3.08 Time

 Predictor         Coef     SE Coef        T       P
 Constant        560.65       29.37    19.09   0.000
 Time           -3.0771      0.8498    -3.62   0.002

S = 23.3980   R-Sq = 42.1%
```

(b) All the conditions for inference appear to be upheld. First, it is reasonable to view the children as independent. The scatterplot in part (a) appears to be roughly linear and does not suggest that the standard deviation changes (the provided scatterplot of residuals against time spent at the table also supports this later observation). The provided stemplot suggests that the distribution of residuals has a slightly irregular appearance, but it is not markedly non-Normal.



```
-3 | 6
-2 | 7432
-1 | 8853
-0 | 21
 0 |
 1 | 3679
 2 | 028
 3 | 23
```

(c) The slope is significantly different from 0; $t = -3.62$ and P = 0.002. Software reports that $SE_b = 0.8498$. With df = 18, t^* = 2.101, so the 95% confidence interval for β is –3.0771 ± 2.101(0.8498) = –4.8625 to –1.2917 calories per minute.



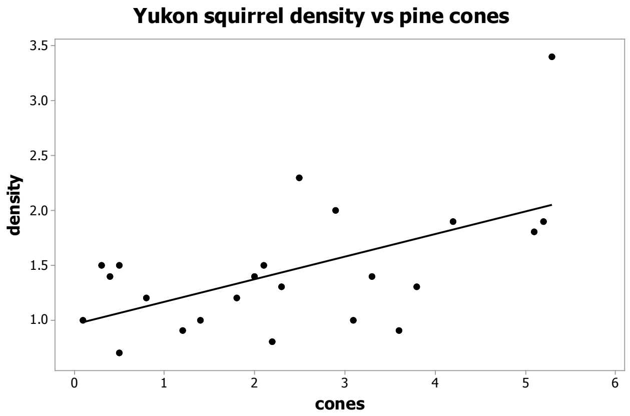
26.39 A stemplot is provided. The distribution is skewed right, but the sample is large so t procedures should be safe. We find $\bar{x} = 0.2781$ g/m^2 and s = 0.1803 g/m^2. Table C gives t* = 1.984 for df = 100 (rounded down from 115). The 95% confidence interval for μ is $0.2781 \pm 1.981(0.1803/\sqrt{116}) = 0.2449$ to 0.3113 g/m^2. (Using df = 115, we have t* = 1.981, and the 95% confidence interval for μ is identical.)

```
 0 | 0067778999
 1 | 000011112222333455555566677778889
 2 | 000001111223334446666677788889
 3 | 00000111222333456667788999
 4 | 01456667
 5 | 3589
 6 | 04
 7 |
 8 | 29
 9 | 0
10 | 5
```

26.40 Refer to the output provided with the solution to Exercise 26.38. We construct a prediction interval for the calories of a single child sitting at the table for 40 minutes, instead of a confidence interval for mean calories of all children sitting for 40 minutes. We have $\hat{y} = 560.65 - 3.0771(40) = 437.57$ calories. Then $SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}} = 23.4\sqrt{1 + \frac{1}{20} + \frac{(40 - 34.01)^2}{758.07}} = 24.51$ calories. A 95% prediction interval is given by $437.57 \pm 2.101(24.51)$ = 386 to 489 calories. Note that it would be preferable to simply ask for this interval from software. Doing so, the following output would be appended to the output provided with Exercise 26.38.

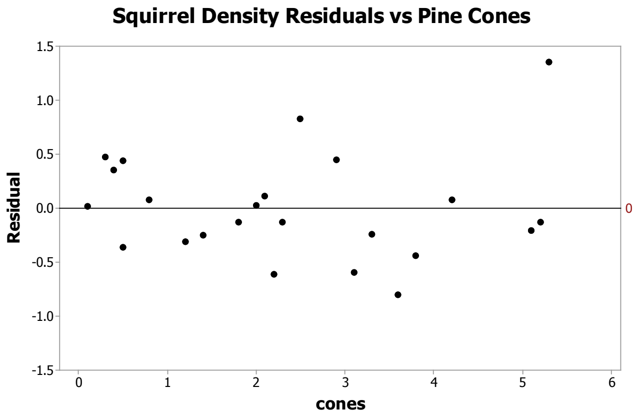
```
New Obs     Fit  SE Fit        95% CI              95% PI
      1  437.57    7.30  (422.23, 452.90)  (386.07, 489.06)
```

26.41 PLAN: We examine the relationship between pine cone abundance and squirrel density using a scatterplot and regression. SOLVE: The provided scatterplot indicates a positive relationship that is roughly linear, with what appears to be an outlier at the upper right of the graph. Regression output is shown. Regression gives predicted squirrel density as $\hat{y} = 0.961 + 0.205x$. The slope is significantly different from zero (t = 3.13, P = 0.005). To assess the evidence that more cones leads to more offspring, we should use the one-sided alternative, $H_{\mathrm{a}}\colon\ \beta > 0$, for which P is half as large (so P = 0.0025). The conditions for inference seem to be violated. The provided residual plot shows what appears to be increasing variability with increasing cone values, as well as the outlier already mentioned. The provided stemplot of the residuals indicates two large positive outliers; the distribution may be right-skewed. CONCLUDE: We seem to have strong evidence of a positive linear relationship between cone abundance and squirrel density; however, conditions for inference may not be satisfied.



Regression Analysis: density versus cones

```
Coefficients

 Term          Coef  SE Coef  T-Value  P-Value   VIF
 Constant     0.961    0.188     5.12    0.000
 cones       0.2053   0.0657     3.13    0.005  1.00

 S          R-sq
 0.501448   31.75%
```

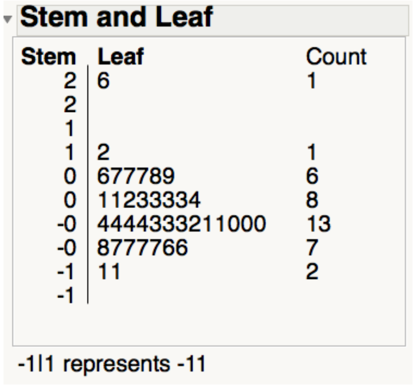


```
-0 | 76
-0 | 54
-0 | 33222
-0 | 111
 0 | 00001
 0 | 3
 0 | 444
 0 |
 0 | 8
 1 |
 1 | 3
```

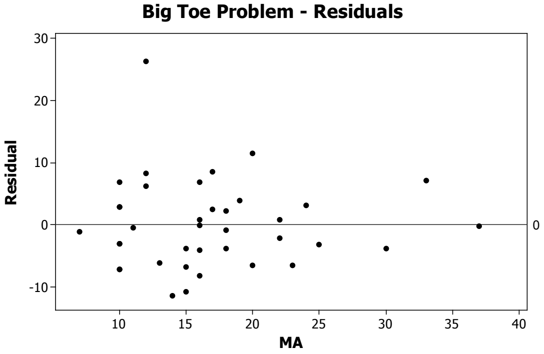
26.42 PLAN: We examine the relationship between HAV angle and MA angle using a scatterplot and regression. SOLVE: Refer to solutions to Exercises 7.49 and 7.51 for the scatterplot and the regression line fit to the data. Although there was an outlier in the data, the data show a roughly linear relationship. From Exercise 7.51, we have $\hat{y}$ = 19.723 + 0.3388*x*. The regression output shown indicates that for testing $H_0: \beta = 0$ versus $H_a: \beta \neq 0$, we find *t* = 1.90 and *P* = 0.065. We have some evidence of a linear relationship between MA angle and HAV, but not strong evidence. Note that perhaps the researchers were really interested in testing $H_0: \beta = 0$ versus $H_a: \beta > 0$ (as perhaps they felt that severe MA deformity is associated with larger MA angle). If so, then *P* = 0.033, which is half that for the two-sided test. We have strong evidence for such an assertion. An analysis of the residuals, both in the provided stemplot and the provided scatterplot against MA angle, shows the same outlier visible in the original scatterplot; this might make us hesitate to use inference procedures. Other than the outlier, there are no great causes of concern: the rest of the stemplot appears to be roughly Normal, and the scatterplot has no clear pattern, although there is some suggestion that the variability about the line is slightly greater for small MA angles. It seems reasonable to believe that observations are independent here because the data concern different patients. CONCLUDE: The correlation is significantly positive with the full data set and significantly different from zero with the outlier removed. In neither case is the relationship very useful for prediction, because the models explain less than 20% of the total variation in HAV.

Note: *If we remove the outlier, then $\hat{y}$ = 17.7 + 0.419*x. *In the absence of the outlier, there is strong evidence of a linear relationship between HAV and MA angle (*t *= 2.93,* P *= 0.006). Residual analysis shows little reason for concern, and all assumptions needed for inference appear to be met, as discussed above.*

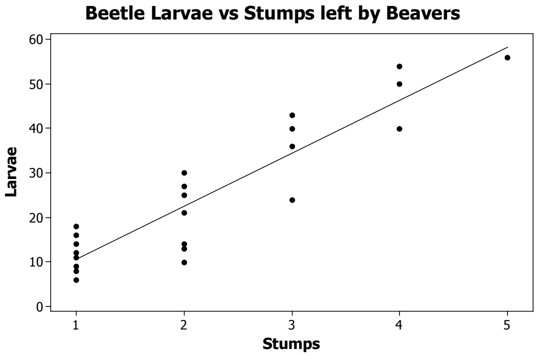
```
The regression equation is
HAV = 19.7 + 0.339 MA

 Predictor        Coef    SE Coef       T       P
 Constant       19.723      3.217    6.13   0.000
 MA             0.3388     0.1782    1.90   0.065

S = 7.22371   R-Sq = 9.1%   R-Sq(adj) = 6.6%
```



```
Stem and Leaf
Stem  Leaf             Count
   2 | 6                 1
   2 |
   1 |
   1 | 2                 1
   0 | 677789            6
   0 | 11233334          8
  -0 | 4444333211000    13
  -0 | 8777766           7
  -1 | 11                2
  -1 |

-1|1 represents -11
```

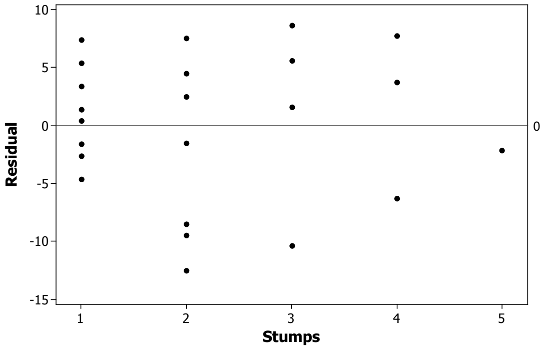


26.43 PLAN: We will examine the relationship between beaver stumps and beetle larvae using a scatterplot and regression. We specifically wish to test for a positive slope β and find a confidence interval for β. SOLVE: The provided scatterplot shows a positive linear association; the regression line is $\hat{y} = -1.286 + 11.894x$. This line is superimposed on the scatterplot. A stemplot of the residuals is shown and does not suggest non-Normality of the residuals, the provided residual scatterplot does not suggest nonlinearity, and the problem description makes clear that observations are independent. Regression output is shown. To test $H_0: \beta = 0$ versus $H_a: \beta > 0$, the test statistic is $t = 10.47$ (df = 21), for which Table C provides a one-sided P-value, $P < 0.0005$. For df = 21, $t^* = 2.080$ for 95% confidence, so with b and SE_b as given by Minitab, we are 95% confident that β is between 11.894 ± 2.080(1.136) = 9.531 and 14.257. CONCLUDE: We have strong evidence that beetle larvae counts increase with beaver stump counts. Specifically, we are 95% confident that each additional stump is (on average) accompanied by between 9.5 and 14.3 additional larvae clusters.



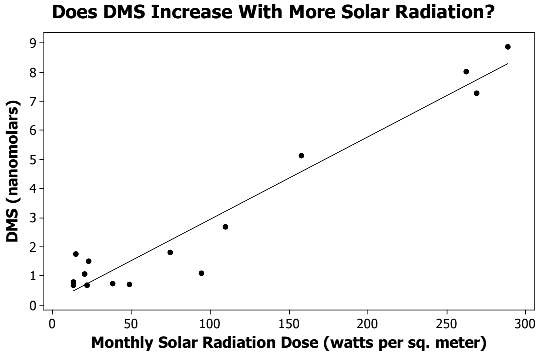
```
-1 | 300
-0 | 965
-0 | 3222
 0 | 0122344
 0 | 567789
```

```
The regression equation is Larvae = - 1.29 + 11.9 Stumps

 Predictor      Coef     SE Coef       T       P
 Constant     -1.286       2.853   -0.45   0.657
 Stumps       11.894       1.136   10.47   0.000

S = 6.41939   R-Sq = 83.9%   R-Sq(adj) = 83.1%
```

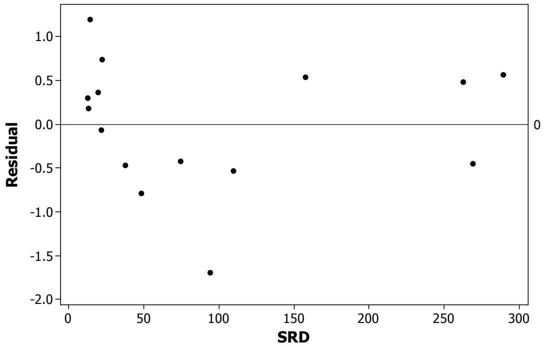


26.44 PLAN: We will examine the relationship between SRD and DMS using a scatterplot and regression. We will test H_0: $\beta = 0$ versus H_a: $\beta > 0$ and find a 90% confidence interval for β. SOLVE: The provided scatterplot shows a positive linear association; the regression line is $\hat{y}$ = 0.1385 + 0.0282x. This line is superimposed on the scatterplot. The provided stemplot of residuals may suggest slight non-Normality due to a low outlier, but not severely so. The provided residual scatterplot seems to provide evidence of nonlinearity (note the curve at the left end and decreasing variability). Observations are clearly independent. Regression output is shown. To test H_0: $\beta = 0$ versus H_a: $\beta > 0$, the test statistic is t = 14.03 (df = 13), for which Table C tells us that the one-sided P-value is P < 0.0005. For df = 13, t* = 1.771 for 90% confidence, so with b = 0.028219 and SE_b = 0.002011 as given by Minitab, we are 90% confident that β is between 0.028219 –

1.771(0.002011) and 0.028219 + 1.771(0.002011), or 0.0247 and 0.0318. CONCLUDE: We have strong evidence that DMS increases with SRD. Specifically, we are 90% confident that, on average, each additional unit increase in SRD raises surface DMS concentration by between 0.025 and 0.032 nanomolar.

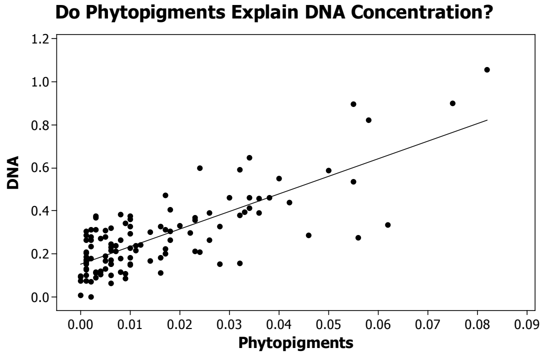


```
-1 | 6
-1 |
-0 | 75
-0 | 4440
 0 | 1334
 0 | 557
 1 | 2
```

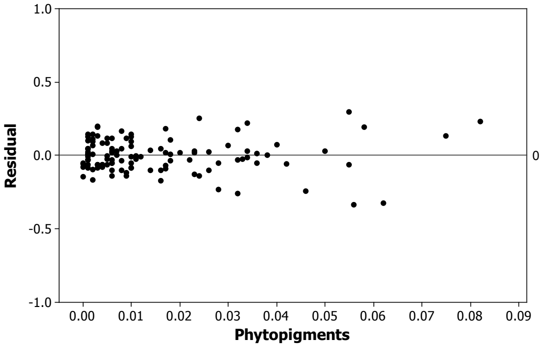


```
The regression equation is DMS = 0.138 + 0.0282 SRD

 Predictor        Coef    SE Coef       T       P
 Constant       0.1385     0.2757    0.50   0.624
 SRD          0.028219   0.002011   14.03   0.000

S = 0.759178   R-Sq = 93.8%   R-Sq(adj) = 93.3%
```

26.45 PLAN: Using a scatterplot and regression, we examine how well phytopigment concentration explains DNA concentration. SOLVE: The provided scatterplot shows a fairly strong, linear, positive association; the regression equation is $\hat{y}$ = 0.1523 + 8.1676x. This line is superimposed on the scatterplot. A provided stemplot of the residuals looks reasonably Normal, but the corresponding scatterplot that is also provided suggests that the variability about the line is greater when phytopigment concentration is greater. This may make regression inference unreliable, but we will proceed. Finally, observations are independent, from the context of the problem. Regression output is shown. The slope is significantly different from 0 (t = 13.25, df = 114, and P < 0.001). We might also construct a 95% confidence interval for β: 8.1676 ± 1.984(0.6163) = 6.95 to 9.39 (the 95% confidence interval is identical if we use df = 114). CONCLUDE: The significant linear relationship between phytopigment and DNA concentrations is consistent with the belief that organic matter settling is a primary source of DNA. Starting from a measurement of phytopigment concentration, we could give a fairly accurate prediction of DNA concentration, because the linear relationship explains about $r^2 = 60.6\%$ of the variation in DNA concentration. We are 95% confident that each additional unit increase in phytopigment concentration increases DNA concentration by between 6.95 and 9.39 units (on average).

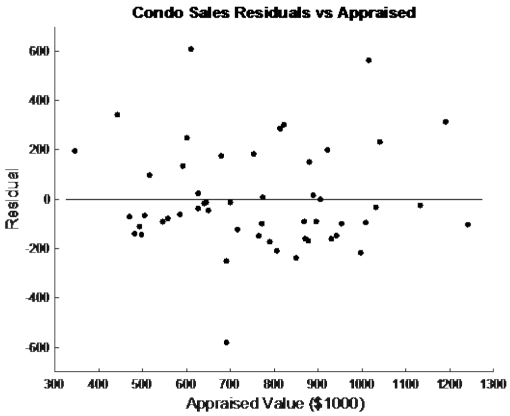


```
-3 | 32
-2 | 5
-2 | 42
-1 | 76
-1 | 443321000
-0 | 99998888876666666655555
-0 | 4443333322222100000
 0 | 000011111112222233333444444
 0 | 66678899
 1 | 0011112233444
 1 | 678999
 2 | 13
 2 | 59
```

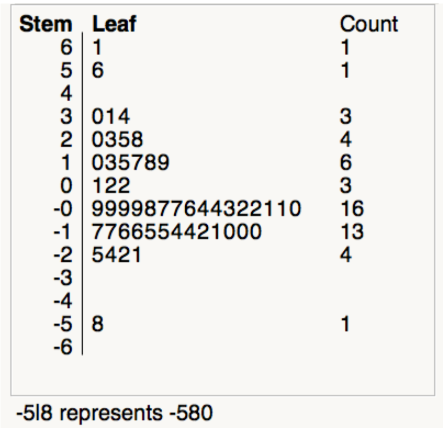


```
The regression equation is DNA = 0.152 + 8.17 Phyto

 Predictor       Coef     SE Coef       T       P
 Constant    0.15231     0.01419   10.73   0.000
 Phyto        8.1676      0.6163   13.25   0.000

S = 0.113612   R-Sq = 60.6%
```

26.46 (a) The residual scatterplot is provided. As usual, we add a horizontal line at residual zero, which is the mean of the residuals. This line corresponds to the regression line in the plot of selling price against appraised value. The residuals show a random scatter about the line, with roughly equal variability across their range (although there are two with large, positive residuals and one large, negative residual). This is what we expect when the conditions for regression inference hold.

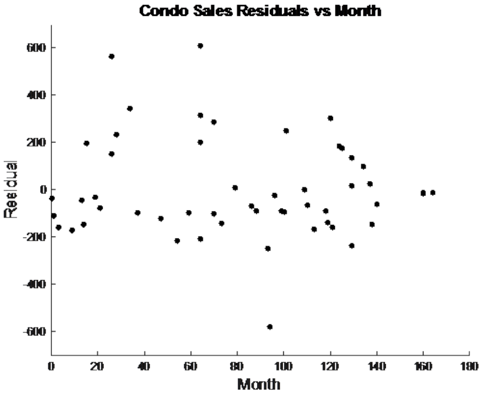


(b) The stemplot is shown. We again see the large residuals (outliers) on either end of the distribution. Otherwise, there are no strong deviations from Normality.



```
Stem Leaf               Count
   6 1                  1
   5 6                  1
   4
   3 014                3
   2 0358               4
   1 035789             6
   0 122                3
  -0 9999877644322110   16
  -1 7766554421000      13
  -2 5421               4
  -3
  -4
  -5 8                  1
  -6

 -5|8 represents -580
```

(c) The plot of residuals against the month of the sale is shown. The pattern of steadily rising residuals for the first 36 months shows that predicted prices are too high for early sales and too low for later sales. This is what we expect if selling prices are rising and appraised values aren't updated quickly enough to keep up. For months 70 to 100, predicted values tend to be a bit high (there are many "small" negative residuals). This may be a result of the crash in Florida real estate values.



26.47 (a) The mean is $\bar{x} = -0.00333$, and the standard deviation is $s = 1.0233$. For a standardized set of values, we expect the mean and standard deviation to be (up to roundoff error) 0 and 1, respectively. **(b)** The provided stemplot does not look particularly symmetric, but it is not strikingly non-Normal for such a small sample.

```
-2 | 2
-1 |
-1 | 4
-0 |
-0 | 32
 0 | 01122
 0 | 7
 1 | 0
 1 | 5
```

(c) The probability that a standard Normal variable is as extreme as this is about 0.0272.

26.48 The *t* statistic given in Figure 26.7 is $t = \frac{a}{SE_a} = \frac{-0.01270}{0.01264} = -1.00$. The *P*-value is 0.332, so we do not have enough evidence to conclude that the intercept α differs from zero.

26.49 For df = 14 and a 95% confidence interval, we use *t** = 2.145, so the interval is –0.01270 ± 2.145(0.01264) = –0.0398 to 0.0144. This interval does contain zero.

26.50 and **26.51** are Web-based exercises.