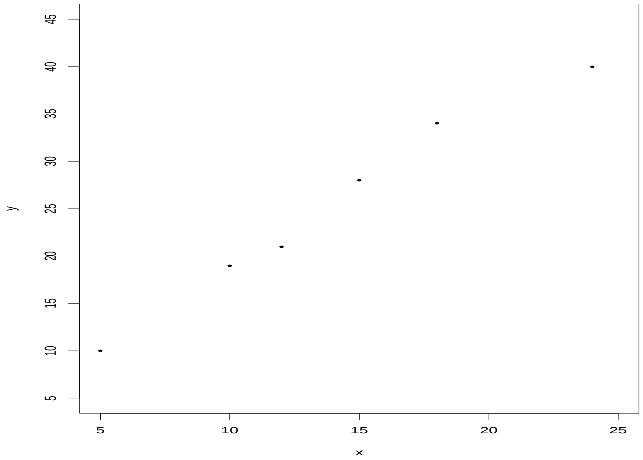
Chapter 11: Linear Regression and Correlation

11.1 A scatterplot of the data is given here:



11.2 The calculations are give here:

i	x_i	y_i	$(x_i - 14)^2$	$(x_i - 14)(y_i - 25.33)$
1	5	10	81	137.97
2	10	19	16	25.32
3	12	21	4	8.66
4	15	28	1	2.67
5	18	34	16	34.68
6	24	40	100	146.70
Total	84	152	218	356

$\bar{x} = 84/6 = 14 \qquad \bar{y} = 152/6 = 25.33$

$S_{xx} = \sum_i (x_i - 14)^2 = 218$

$S_{xy} = \sum_i (x_i - 14)(y_i - 25.33) = 356$

$\hat{\beta}_1 = S_{xy}/S_{xx} = 356/218 = 1.633$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 25.33 - (1.633)(14) = 2.468$

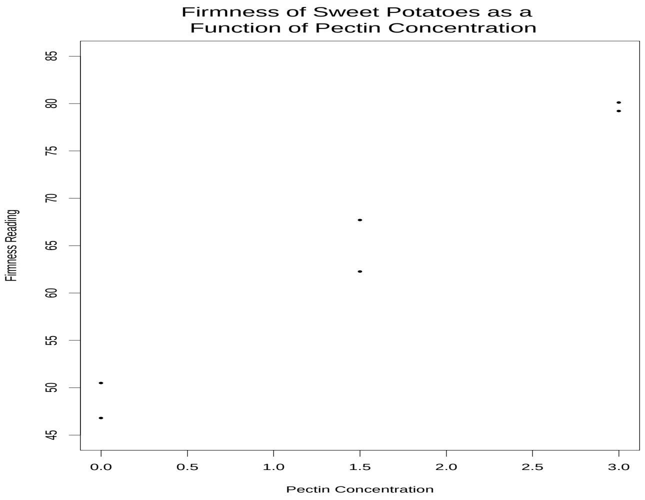
$\hat{y} = 2.468 + 1.633x$

11.3 a. $\hat{y} = 4.698 + 1.97x$

b. yes, the plotted line is relatively close to all 10 data points

c. $\hat{y} = 4.698 + (1.97)(35) = 73.65$

11.4 a. A scatterplot of the data is given here:



b. The calculations are give here:

i	x_i	y_i	$(x_i - 1.5)^2$	$(x_i - 1.5)(y_i - 64.43)$
1	0	8.5	2.25	20.895
2	0	8.4	2.25	26.445
3	3	7.9	0	0
4	3	8.1	0	0
5	6	7.8	2.25	23.505
6	6	7.6	2.25	22.155
7	9	7.3	0	0
8	9	7.0	0	0
9	12	6.8	2.25	23.505
10	12	6.7	2.25	22.155
Total	60	386.6	9	93

$\bar{x} = 60/10 = 6$ $\quad$ $\bar{y} = 386.6/6 = 64.43$

$S_{xx} = \sum_i (x_i - 1.5)^2 = 9$

$S_{xy} = \sum_i (x_i - 1.5)(y_i - 64.43) = 93$

$\hat{\beta}_1 = S_{xy}/S_{xx} = 93/9 = 10.33$
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 64.43 - (10.33)(1.5) = 48.935$
$\hat{y} = 48.935 + 10.33x$

11.5 $\hat{y} = 48.935 + 10.33(1.0) = 59.265$

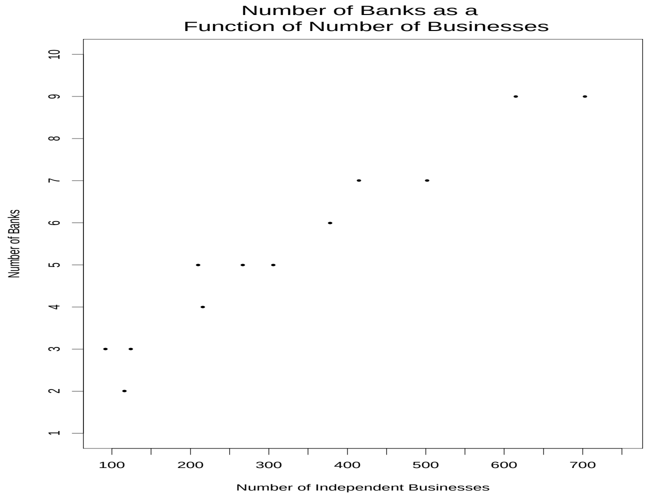
11.6 The Time Needed is increasing as the Number of Items increases but the rate of increase is less for larger values of Number of Items than it is for smaller values of Number of Items. A square root or logarithmic transformation is suggested from the plot.

11.7 a. For the transformed data, the plotted points appear to be reasonably linear.

b. The least squares line is $\hat{y} = 3.10 + 2.76\sqrt{x}$
(Using rounded values $3.097869 \approx 3.10$ and $2.7633138 \approx 2.76$).
That is, Estimated Time Needed $= 3.10 + 2.76\sqrt{\text{Number of Items}}$.

11.8 The Root Mean Square Error (RMSE) is 2.923232. This is the square root of the ”average” sum of squares of the y-values about the least squares line. This is an assessment of how ”close” the data values are to the least squares line in the y-direction.

11.9 a. A scatterplot of the data is given here:



From a scatterplot of data, a straight-line relationship between y and x seems reasonable.

b. The intercept is 1.766846 (1.77) and the slope is 0.0111049 (0.0111).

c. An interpretation of the estimated slope is as follows. For an increase of 1 business in a zip code area, there is an increase of 0.0111 in the average number of banks in the zip code.

d. From the output, the sample residual standard deviation is given by “Root MSE”, 0.5583.

11.10 From the scatterplot, there does not appear to be an increase or decrease in the variability of y as x increases. It appears to be somewhat constant over the range of the x values. Increasing variability would appear as a funnel shape (or back to back funnels) in the scatterplot.

11.11 a. The **LOWESS** smooth is basically a straight line, with a slight bend at the larger values of Income. Thus, it would appear that the relation is basically linear.

b. The points with high leverage are those that are far from the average value of the x-values along the x-axis. In this case there are two points with very high income values. They are the last two values in the data listing. The plot shows that the point with Income=65.0 and Price=110.0 is a considerable distance from the **LOWESS** line. This point therefore has high influence. The other high-leverage point Income=70.0 and Price=185.0 falls reasonably close to the **LOWESS** line. This point would not have such a large influence on the fitted line.

11.12 a. Slope = 1.80264 and Intercept = 47.15048 yielding $\hat{y} = 47.15 + 1.80x$ for the least squares line.

b. The estimated slope is approximately 1.80. This implies that as yearly income increases \$1000, the average sale price of the home increases \$1800. The estimated Intercept would be the average sale price of homes bought by someone with 0 yearly income. Because the data set did not contain any data values with yearly income close to 0, the Intercept does not have meaningful interpretation.

c. The residual standard deviation is "Root MSE" = 14.445.

11.13 The estimated slope changed considerably from 1.80 to 2.46. This resulted from excluding the data value having high-influence. Such a data point twists the fitted line towards itself and hence can greatly distort the value of the estimated slope.

11.14 a. From the computer output for Exercises 11.6 and 11.7, the standard error of the estimated slope $SE(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1} = 0.216418$. A 90% C.I. for β_1 is given by

$$\hat{\beta}_1 \pm t_{.05,98} SE(\hat{\beta}_1) \Rightarrow 2.7633 \pm (1.645)(0.216418) \Rightarrow (2.55, 2.98)$$

b. As the number of items increases, there is no change in the average Time needed to assemble for shipment.

c. $H_a : \beta_1 > 0$.

d. The p-value for testing $H_a : \beta_1 > 0$ is less than 0.0001. Thus, there is significant evidence that the slope is greater than 0. In order to conduct the test of hypothesis, the data from a series of populations having normal distributions with a common variance and with mean responses following the equation:

$\mu_{y|x} = \beta_o + \beta_1 \sqrt{x}$

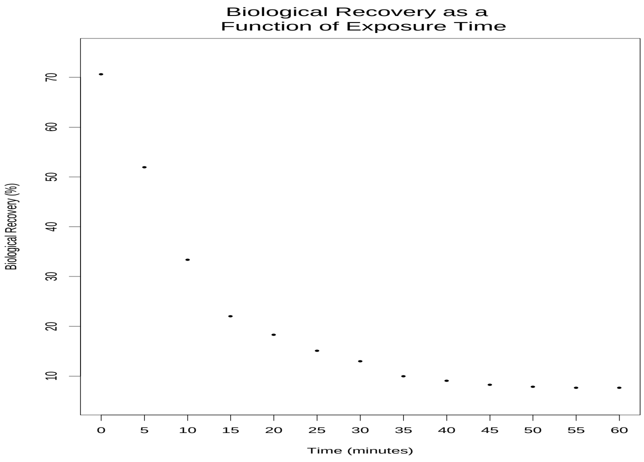
11.15 p-value < 0.0001

11.16 The original data and the log base 10 of recovery are given here:

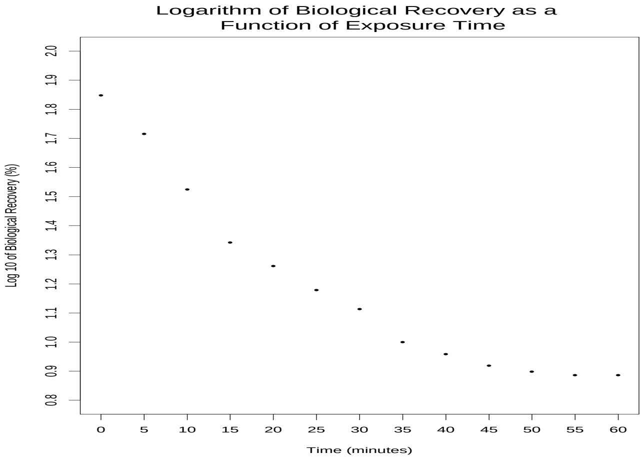
```
Data Display

 Cloud  Time   Recovery   LogRecovery
   1      0      70.6        1.849
   2      5      52.0        1.716
   3     10      33.4        1.524
   4     15      22.0        1.342
   5     20      18.3        1.262
   6     25      15.1        1.179
   7     30      13.0        1.114
   8     35      10.0        1.000
   9     40       9.1        0.959
  10     45       8.3        0.919
  11     50       7.9        0.898
  12     55       7.7        0.886
  13     60       7.7        0.886
```

a. Scatterplot of the data is given here:



b. Scatterplot of the data using $log_{10}(y)$ is given here:



11.17 Minitab Output is given here:

```
Regression Analysis: LogRecovery versus Time


The regression equation is
LogRecovery = 1.67 - 0.0159 Time

Predictor        Coef     SE Coef          T        P
Constant      1.67243     0.05837      28.65    0.000
Time        -0.015914    0.001651      -9.64    0.000

S = 0.1114      R-Sq = 89.4%     R-Sq(adj) = 88.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      1.1523      1.1523     92.93    0.000
Residual Error    11      0.1364      0.0124
Total             12      1.2887
```

a. $\hat{y} = 1.67 - 0.0159x$

b. $s_\epsilon = 0.1114$

c. $SE(\hat{\beta}_o) = 0.05837 \qquad SE(\hat{\beta}_1) = 0.001651$

11.18 $H_o : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$

Test Statistic: $|t| = 9.64$

p-value $= 2P(t_{11} > 9.64) < 0.0001 < 0.05 \Rightarrow$ Reject H_o and conclude there is significant evidence that β_1 is not 0.

11.19 a. Yes, the data values fall approximately along a straight-line.

b. $\hat{y} = 12.51 + 35.83x$

11.20 a. 1.069

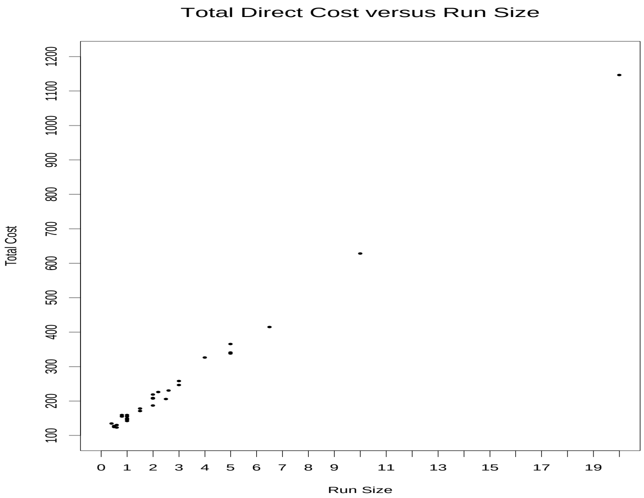
b. 6.957

c. $H_o : \beta_1 \leq 0$ versus $H_a : \beta_1 > 0$

Test Statistic: $t = 5.15$

p-value $= P(t_{10} > 5.15) = 0.0002 \Rightarrow$ Reject H_o and conclude there is significant evidence that there is a positive linear relationship.

11.21 a. Scatterplot of the data is given here:



An examination of the scatterplot reveals that a straight-line equation between total cost and run size may be appropriate. There is a single extreme point in the data set but no evidence of a violation of the constant variance requirement.

b. $\hat{y} = 99.777 + 5.1918x$

The residual standard deviation is $s = \sqrt{148.999} = 12.2065$

c. A 95% C.I. for the slope is given by $\hat{\beta}_1 \pm t_{0.025,28} SE(\hat{\beta}_1) \Rightarrow \quad 5.1918 \pm (2.048)(0.586455) \Rightarrow (5.072, 5.312)$

11.22 a. $t = 88.53$

b. From the output p-value = 0.0000, which can be interpreted as p-value < 0.0001. This is the p-value for a two-sided test of hypotheses. In the context of this problem, the appropriate alternative hypothesis would be that the slope is greater than 0 since the total cost of a job should increase as the size of the job increases. Thus, for $H_a : \beta_1 > 0$, p-value $= P(t > 88.53) < 0.0001$.

11.23 a. $F = 7837.26$ with p-value = 0.0000.

b. The F-test and two-sided t-test yield the same conclusion in this situation. In fact, for this type of hypotheses, $F = t^2$.

11.24 $\hat{y} = 1.67243 - (0.015914)(30) = 1.195$

95% C.I. on $\mu_{x=30}$: $\quad \hat{y} \pm t_{.025,11} s_\epsilon \sqrt{\frac{1}{n} + \frac{(30-\bar{x})^2}{S_{xx}}} \Rightarrow$

$1.195 \pm (2.201)(0.1114)\sqrt{\frac{1}{13} + \frac{(30-30)^2}{4550}} \Rightarrow \quad 1.195 \pm 0.068 \Rightarrow \quad (1.127, 1.263)$

11.25 95% prediction interval for log biological recovery percentage at x=30 is given by

$\hat{y} \pm t_{.025,11} s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(30-\bar{x})^2}{S_{xx}}} \Rightarrow$

$1.195 \pm (2.201)(0.1114)\sqrt{1 + \frac{1}{13} + \frac{(30-30)^2}{4550}} \Rightarrow \quad 1.195 \pm 0.0.254 \Rightarrow \quad (0.941, 1.449)$

The prediction interval is somewhat wider than the confidence interval on the mean.

11.26 a. $\hat{y} = -1.733333 + 1.316667x$

b. The p-value for testing $H_o : \beta_1 \leq 0$ versus $H_a : \beta_1 > 0$ is

p-value $= P(t_{10} \geq 6.342) < 0.0001 \Rightarrow$ Reject H_o and conclude there is significant evidence that the slope β_1 is greater than 0.

11.27 a. 95% Confidence Intervals for $E(y)$ at selected values for x:

$x = 4 \Rightarrow (2.6679, 4.3987)$
$x = 5 \Rightarrow (4.2835, 5.4165)$
$x = 6 \Rightarrow (5.6001, 6.7332)$
$x = 7 \Rightarrow (6.6179, 8.3487)$

b. 95% Prediction Intervals for y at selected values for x :

$x = 4 \Rightarrow (1.5437, 5.5229)$
$x = 5 \Rightarrow (2.9710, 6.7290)$
$x = 6 \Rightarrow (4.2877, 8.0456)$
$x = 7 \Rightarrow (5.4937, 9.4729)$

c. The confidence intervals in part (a) are interpreted as “We are 95% confident that the average weight loss over many samples of the compound when exposed for 4 hours will be between 2.67 and 4.40 pounds.” Similar statements for other hours of exposure.

The prediction intervals in part (b) are interpreted as “We are 95% confident that the weight loss of a single sample of the compound when exposed for 4 hours will be between 1.54 and 5.52 pounds.” Similar statements for other hours of exposure.

11.28 a. $\hat{y} = 99.77704 + 51.9179x \Rightarrow$ When x=2.0, $\hat{E}(y) = 99.77704 + (51.9179)(2.0) = 203.613$ as is shown in the output.

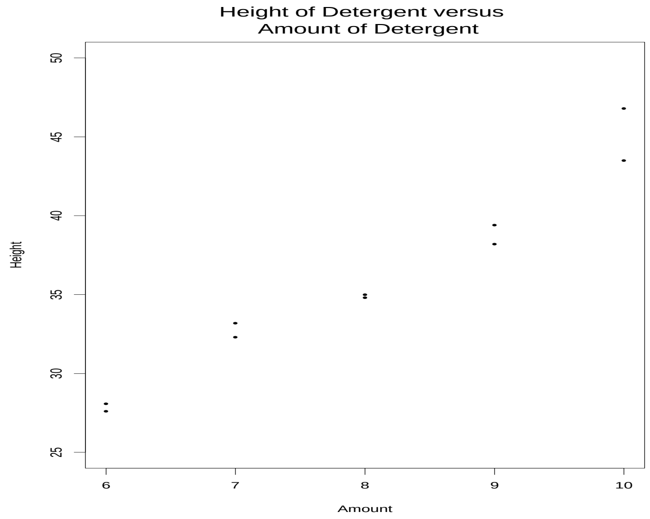
b. The 95% C.I. is given in the output as $(198.902, 208.323)$

11.29 No, because $x = 2.0$ is close to the mean of all x-values used in determining the least squares line.

11.30 a. The 95% P.I. is given in the output as $(178.169, 229.057)$

b. Yes, because \$250 does not fall within the 95% prediction interval. In fact, it is considerably higher that the upper value of \$229.057.

11.31 a. Scatterplot of the data is given here:



b. The SAS output is given here:

```
Model: MODEL1
Dependent Variable: Y

                         Analysis of Variance

                         Sum of         Mean
Source          DF      Squares       Square      F Value       Prob>F

Model            1    330.48450    330.48450      169.213       0.0001
Error            8     15.62450      1.95306
C Total          9    346.10900

    Root MSE       1.39752     R-square       0.9549
    Dep Mean      35.89000     Adj R-sq       0.9492
    C.V.           3.89390

                         Parameter Estimates

                 Parameter      Standard    T for H0:
 Variable  DF     Estimate         Error   Parameter=0    Prob > |T|

 INTERCEP   1     3.370000    2.53872138         1.327        0.2210
 X          1     4.065000    0.31249500        13.008        0.0001
```

$\hat{y} = 3.37 + 4.065x$

c. The residual plot is given here:

```
          Plot of RESID*PRED.  Symbol used is '*'.

     |
     |
   3 +
     |
     |                                                   *
     |
     |
     |
     |
   2 +
     |
     |
     |
     |              *
     |
R    |
E  1 +
S    |
I    |
D    |
U    |              *
A    |  *
L    |
S  0 +-----------------------------------------------------
     |  *
     |
     |
     |                                       *           *
     |
     |                          *
  -1 +
     |                          *
     |
     |
     |
     |                                       *
     |
  -2 +
     |
     ---+-----------+-----------+-----------+-----------+--
     27.760      31.825      35.890      39.955      44.020

                        PREDICTED VALUE
```

The residual plot indicates that higher order terms in x may be needed in the model.

11.32 a. A test of lack of fit will be conducted:

$SSP_{exp} = \sum_{ij}(y_{ij} - \bar{y}_{i.})^2 = (28.1-27.85)^2 + (27.6-27.85)^2 + (32.3-32.75)^2 + (33.2-32.75)^2 + (34.8-34.90)^2 + (35.0-34.90)^2 + (38.2-38.80)^2 + (39.4-38.80)^2 + (43.5-45.15)^2 + (46.8-45.15)^2 = 6.715$

From the output of exercise 11.31, SS(Residuals) = 15.6245. Thus,

$SS_{Lack} = SS(Residuals) - SS_{exp} = 15.6245 - 6.715 = 8.9095$

$df_{Lack} = n - 2 - \sum_i(n_i - 1) = 10 - 2 - 5(2-1) = 3$

$df_{exp} = \sum_i(n_i - 1) = 5(2-1) = 5$

$F = \frac{8.9095/3}{6.715/5} = 2.21 < 5.41 = F_{.05,3,5}$

There is not sufficient evidence that the linear model is inadequate.

b. $S_{xx} = \sum_i(x_i - \bar{x}) = 20.$

95% Prediction interval for y at x:

$\hat{y} \pm t_{.025,8} s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} = 3.37 + 4.065x \pm (2.306)(1.398)\sqrt{1 + \frac{1}{10} + \frac{(x-8)^2}{20}}$

For x=6,7,8,9,10 we have

x	$\hat{y}$	95% P.I.
6	27.760	(24.086, 31.434)
7	31.825	(28.369, 35.281)
8	35.890	(32.510, 39.270)
9	39.955	(36.499, 43.411)
10	44.020	(40.346, 47.694)

11.33 $\hat{y} = 1.47 + 0.797x$

$H_o : \beta_1 \leq 0$ versus $H_a : \beta_1 > 0$

$t = 7.53 \Rightarrow$ p-value $= Pr(t_8 \geq 7.53) < 0.0005 \Rightarrow$

There is sufficient evidence in the data that the slope is positive.

11.34 a. $\hat{x} = (13 - 1.47)/0.797 = 14.47$

b. $s_\epsilon = 1.017, S_{xx} = 92.4, \bar{x} = 15.6, t_{.025,8} = 2.306 \Rightarrow c^2 = \frac{(2.306)^2(1.017)^2}{(.797)^2(92.4)} = 0.0937$

$d = \frac{(2.306)(1.017)}{.797}\sqrt{\frac{10+1}{10}(1 - .0937) + \frac{(14.47-15.6)^2}{92.4}} = 2.958$

$\hat{x}_L = 15.6 + \frac{1}{(1-.0937)}(14.47 - 15.6 - 2.958) = 11.09$

$\hat{x}_U = 15.6 + \frac{1}{(1-.0937)}(14.47 - 15.6 + 2.958) = 17.42 \Rightarrow$

$(11.09, 17.42)$ is a 95% prediction interval on the actual volume when the estimated volume is 13.

11.35 a. $\hat{y} = 0.11277 + 0.11847x \approx 0.113 + 0.118x$

Dependent variable is Transformed CUMVOL and Independent variable is Log(Dose).

b. $\hat{x} = (y - 0.11277)/0.11847$

$s_\epsilon = 0.04597, S_{xx} = 4.9321, \bar{x} = 2.40, t_{.005,8} = 2.819 \Rightarrow$

$c^2 = \frac{(2.819)^2(.04597)^2}{(.11847)^2(4.9321)} = 0.2426$

$d = \frac{(2.819)(.04597)}{.11847}\sqrt{\frac{24+1}{24}(1 - .242467) + \frac{(\hat{x}-2.40)^2}{4.9321}}$

$\hat{x}_L = 2.40 + \frac{1}{(1-.2426)}(\hat{x} - 2.40 - d)$

$\hat{x}_U = 2.40 + \frac{1}{(1-.2426)}(\hat{x} - 2.40 + d) \Rightarrow$

y	TRANS(y)	$\widehat{LOG(x)}$	$\hat{x}$	d	$\widehat{LOG(x)}_L$	$\widehat{LOG(x)}_U$	$\hat{x}_L$	$\hat{x}_U$
10	.322	1.764	5.84	.242289	1.24038	1.88018	3.46	6.55
14	.383	2.285	9.83	.198634	1.98616	2.51068	7.29	12.31
19	.451	2.855	17.38	.221359	2.70876	3.29328	15.01	26.93

11.36 The four values of CUMVOL are 10, 20, 30, 12 yielding $\bar{y} = 0.42969$ and $s_{\bar{y}} = .05849$, where y is the arcsine of the square root of CUMVOL.

For 50%: $\hat{x} = 2.367$ and 95% confidence limits are (0.79, 4.45)

For 75%: $\hat{x} = 5.861$ and 95% confidence limits are (2.20, 12.88)

11.37 The output yields R-square=0.9452. The estimated slope of the regression line is 0.0111049 which is positive, indicating an increasing relation between Branches and Business. Thus, the correlation is the positive square root of 0.9452, i.e., $r = 0.9722$.

11.38 a. Test $H_o : \rho_{yx} \leq 0$ versus $H_a : \rho_{yx} > 0$

Test Statistic: $t = \frac{.9722\sqrt{12-2}}{\sqrt{1-(.9722)^2}} = 13.13$

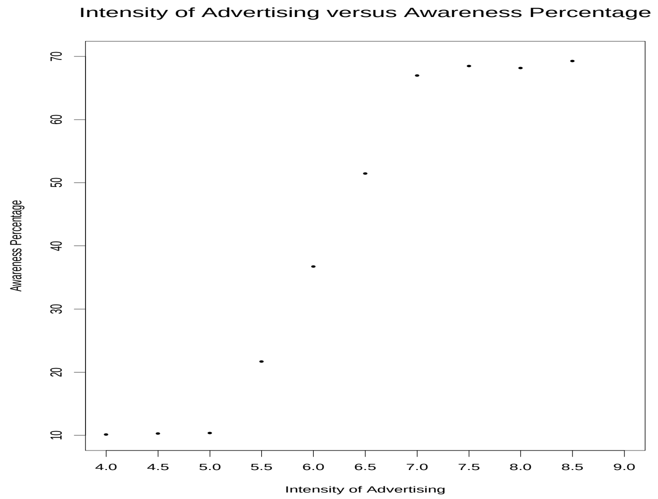
p-value $= Pr(t_{10} \geq 13.13) < 0.0005$

Conclusion: There is significant evidence that the correlation is positive.

b. The t-test for testing whether the slope is 0 is exactly the same (except for rounding error) as the t-test for whether the correlation is 0. From the output we have $t =$ 13.138, which is approximately the same as we computed in part a. In fact, they would be identical if we would have used a greater number of decimal places in our calculations in part a.

11.39 a. The correlation is given as 0.956. This value indicates a strong positive trend between intensity of the advertising and the awareness percentage. That is, as intensity of the advertising increases we would generally observe an increase in awareness percentage.

b. Scatterplot of the data is given here:



The relation between intensity and awareness does not appear to follow a straight line. It is generally an increasing relation with a threshold value of approximately x=5 before awareness increases beyond a value of 10%. A leveling off of the sharp increase in awareness occurs at an intensity of x=7.

11.40 a. There appears to be a general increase in salary as the level of experience increases. However, there is considerable variability in salary for persons having similar levels of experience.

b. Note that Case 11 has a person with 14 years of experience and a salary of 37.9. Salaries of less than 40 are generally associated with persons having less than 5 years experience.

11.41 For the points with the smaller symbols, there is a general upward trend in SALARY as EXPER increases in value. The point denoted with a large box, SALARY=37.9 and EXPER 14 years, clearly does not fall within the general pattern of the other points. This point is an outlier; it has high leverage because the experience for this graduate is considerably larger than the average, and the point has high influence because the starting salary value is much smaller than all the other points having experience value close to 14 years.

11.42 a. The high-influence point was below the line and at the right edge of the data. Such points twist the line down and hence decrease the value of the slope. Omitting the point will cause the line to rise and thus the slope would increase, as was observed, 1.470 to 1.863.

b. In general, outliers tend to increase the standard deviation. Therefore, removing an outlier should decrease the standard deviation, as was observed, 5.40 to 4.07. This is approximately a 25% decrease, a substantial change.

c. High-influence points can either increase or decrease the correlation depending on structure in the remaining data values. In this data set, because the high-influence point resided well outside the pattern of the remaining data values, it should decrease the correlation. Thus, removing the high-influence point should result in an increase in the correlation, as was observed, 0.703 to 0.842.

11.43 a. $\hat{y} = 5.890659 + 0.0148652x$

b. $H_o : \beta_1 \leq 0$ versus $H_a : \beta_1 > 0$, from the output $t = 5.812$.

The p-value on the output is for a two-sided test. Thus, p-value $= Pr(t_5 \geq 5.812) =$ 0.00106 which indicates that there is significant evidence that the slope is greater than 0.

11.44 a. $\hat{y} = 1.007445 + 2.307015ln(x)$

b. $H_o : \beta_1 \leq 0$ versus $H_a : \beta_1 > 0$, from the output $t = 7.014$.

The p-value on the output is for a two-sided test. Thus, p-value $= Pr(t_5 \geq 7.014) =$ 0.00045 which indicates that there is significant evidence that the slope is greater than 0.

11.45 The regression line using ln(x) appears to provide the better fit. The scatterplots indicate that the line using ln(x) more closely matches the data points. Also, the residual standard deviation using ln(x) is smaller than the standard deviation using x (2.0135 vs 2.3801).

11.46 The preferred model is $y = \beta_0 + \beta_1 ln(x) + \epsilon$ with estimates $\hat{y} = 1.007445 + 2.307015ln(x)$. With x=75, $\hat{y} = 1.007445 + 2.307015ln(75) = 10.97$. A 95% prediction interval for y when x=75 is

$10.97 \pm (2.571)(2.0135)\sqrt{1+\frac{1}{7}+\frac{(4.31749-3.6502)^2}{37.4731}} \Rightarrow 10.97 \pm 5.56 \Rightarrow (5.41, 16.53)$. This is a fairly wide interval and thus the regression line is not providing a very accurate prediction.

11.47 a. The prediction equation is $\hat{y} = 140.074 + 0.61896x$.

b. The coefficient of determination is $R^2 = 0.9420$ which implies that 94.20% of the variation in fuel usage is accounted for by its linear relationship with flight miles. Because the estimated slope is positive, the correlation coefficient is the positive square root of R^2, i.e., $r = \sqrt{0.9402} = 0.97$.

c. The only point in testing $H_o : \beta_1 \leq 0$ versus $H_a : \beta_1 > 0$ would be in the situation where the flights were of essentially the same length and there is an attempt to determine if there are other important factors that may affect fuel usage. Otherwise, it would be obvious that longer flights would be associated with greater fuel usage.

11.48 a. With $x = 1000, \hat{\mu}_{y|x=1000} = 140.074 + (0.61896)(1000) = 759.03$. From the output, a 95% C.I. on the mean fuel usage of 1000 mile flights is (733.68, 784.38).

b. With $x = 1000, \hat{y} = 140.074 + (0.61896)(1000) = 759.03$. From the output, a 95% P.I. on the fuel usage for a particular 1000 mile flight is (678.33, 839.73). A value of y equal to 628 does not fall in the 95% P.I. It would be regarded as unusually low and should be investigated to see why the flight performed so well on this particular flight.

11.49 a. The group felt that towns with small populations should be associated with large amounts per person, whereas towns with larger populations would have small amounts per person. Thus, the group claimed that as townsize increased the amount spent per person decreased. If the data supported their claim, then the slope would be negative.

b. The estimated slope was $\hat{\beta}_1 = 0.0005324$ which is positive. Thus, there claim is not supported by the data.

11.50 There is a single point in the upper right hand corner of the plot that is a considerable distance from the rest of the data points. A point located in this region will cause the fitted line to be drawn away from the line which would be fitted to the data values with this point excluded. Thus the regression line given in the plot seems to be somewhat misfitted. A regression line fitted to the data set excluding this point would most likely yield a line with a negative slope.

11.51 a. The point is a very high-influence outlier which has distorted the slope considerably.

b. The regression line with the one point eliminated has a negative slope, $\hat{\beta}_1 = -0.0015766$. This confirms the opinion of the group, which had argued that the smallest towns would have the highest per capita expenditures with decreasing expenditures as the size of the towns increased.

11.52 The slope with the unusual town included was $\hat{\beta}_1 = 0.0005324$. The output with the unusual town excluded is shown to be $\hat{\beta}_1 = -0.0015766$. The slope has changed sign and increased in magnitude.

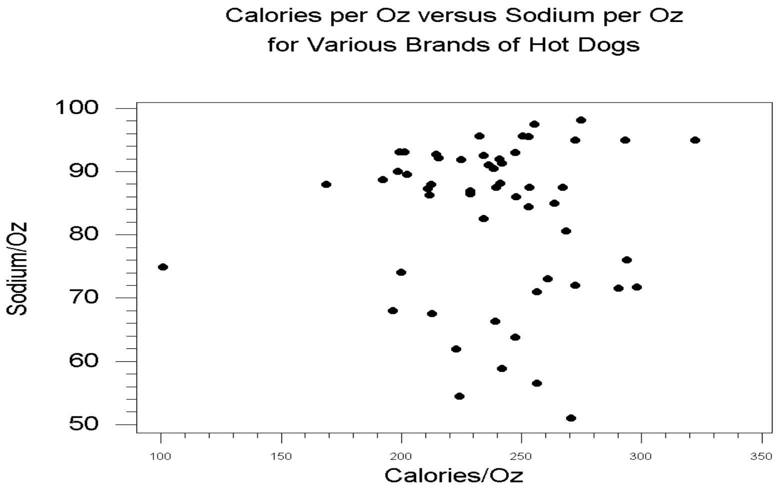
11.53 a. From the scatterplot, it would appear that a straight line model relating Homogenate to Pellet would provide an adequate fit.

b. No, the plot does not indicate any outliers and the variance appears to be constant.

c. We could use the calibration techniques and predict Pellet response using the observe Homogenate reading.

d. Use the calibration prediction interval.

11.54 a. The scatterplot is given here. The Pearson correlation of CalPerOz and SodPerOz equals -0.019 with a p-value = 0.892. From the scatterplot, the values of CalPerOz and SodPerOz appear to be randomly scattered with no discernable relationship. With such a very large p-value, there is not significant evidence that the Pearson correlation is different from 0.


Calories per Oz versus Sodium per Oz
for Various Brands of Hot Dogs
Sodium/Oz
Calories/Oz