Chapter 1 Solutions

1.1 Part (c) is false. The predicted value of Y when $X = 2$ is $\hat{Y} = 100 + 15(2) = 130$, not 110. Parts (a), (b), and (d) are true.

1.2 A residual plot does not help assess (c) the condition of independence of the residuals. It does help assess (a) linearity, (b) constant variance, and (d) zero mean.

1.3 The slope is given in the output under the heading `Coef` for the predictor $WingLength$. The estimate is $\hat{\beta}_1 = 0.4674$.

1.4 The slope is given in the output under the heading `Coef` for the predictor $Year$. The estimate is $\hat{\beta}_1 = 0.01251$.

1.5 The intercept is given in the output under the heading `Coef` for the `Constant`. The estimate is $\hat{\beta}_0 = 1.3655$.

1.6 The intercept is given in the output under the heading `Coef` for the `Constant`. The estimate is $\hat{\beta}_0 = -16.47$.

1.7 As wing length increases by 1 mm, the weight increases by 0.4674 g, on average.

1.8 As year increases by 1, the length of the winning long jump increases by 0.01251 m, on average.

1.9 The regression standard error is given in the output as `S = 1.39959`. We can also compute this from the information given in the `Error` row of the Analysis of Variance:

$$\hat{\sigma_\epsilon} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{223.31}{116-2}} = \sqrt{1.959} = 1.39959$$

A typical deviation of a sparrow weight from the line predicted by its wing length might be about 1.4 grams.

1.10 The regression standard error is given in the output as `S = 0.259522`. We can also compute this from the information given in the `Error` row of the Analysis of Variance:

$$\hat{\sigma_\epsilon} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1.751}{28-2}} = \sqrt{0.06735} = 0.2595$$

A typical deviation of a winning Olympic long jump length from the line predicted by its year might be about 0.26 meters.

1.11 The degrees of freedom for the regression standard error are $n - 2 = 116 - 2 = 114$. The value also appears in the `DF` column of the Analysis of Variance section of the output.

1.12 The degrees of freedom for the regression standard error are $n - 2 = 28 - 2 = 26$. The value also appears in the `DF` column of the Analysis of Variance section of the output.

1.13 The predicted value is $\hat{y}_1 = 25 + 7(10) = 95$. The residual is $y_1 - \hat{y}_1 = 100 - 95 = 5$.

1.14 The predicted value is $\hat{y}_1 = 78 - 0.5(30) = 63$. The residual is $y_1 - \hat{y}_1 = 60 - 63 = -3$.

1.15 a. Computer output gives the fitted regression model as $\hat{Width} = 37.72 - 0.01756Year$

b. As *Year* increases by 1, *Width* decreases by 0.01756 mm, on average.

c. Plugging 1966 into the fitted regression equation, we get $37.72 - 0.01756(1966) = 3.197$ mm.

1.16 a. The computer output gives the fitted regression model as $\hat{Eggs} = -8.98 + 7.33Lantern$.

b. As lantern size increases by 1 mm, the predicted number of eggs laid increases by 7.3 on average.

c. Plugging 14 into the fitted regression equation, we get $-8.98 + 7.33(14) = 93.6$ eggs.

1.17 a. The computer output gives the fitted regression equation as $\hat{MaxGripStrength} = 36.16 + 4.705Attractive$.

b. As *Attractive* increases by 1, *MaxGripStrength* increases by 4.7 kg, on average.

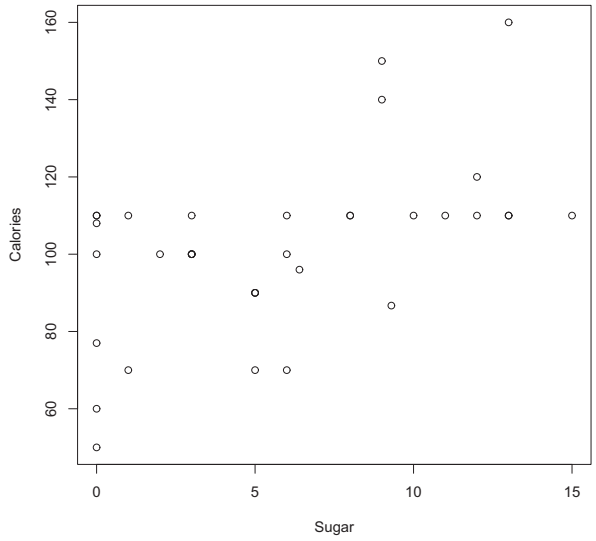
c. Plugging 3 into the fitted equation from part (a) we get a predicted $\hat{MaxGripStrength} = 36.16 + 4.705(3) = 50.3$ kg.

1.18 a. The computer output gives the fitted regression equation as $\hat{MaxGripStrength} = 9.3 + 29.0SHR$.

b. As *SHR* increases by 1, *MaxGripStrength* increases by 29 kg, on average.

c. Plugging 1.5 into the fitted equation from part (a) we get a predicted $\hat{MaxGripStrength} = 9.3 + 29.0(1.5) = 52.8$ kg.

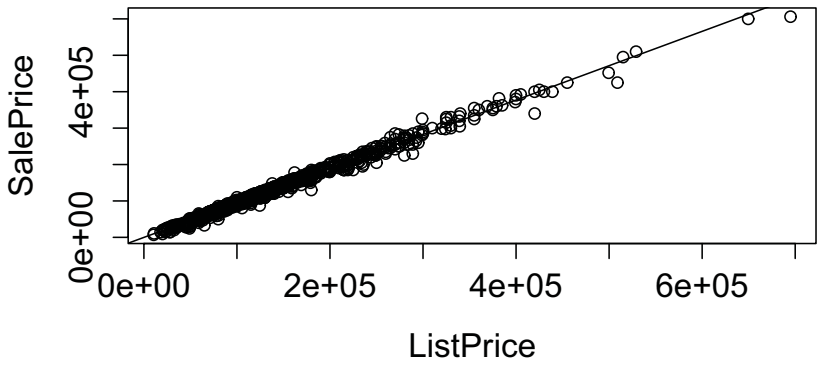
1.19 a. The scatterplot shows a moderate positive association between *Calories* and *Sugar*.



b. Based on regression output, the prediction equation is $\widehat{Calories} = 87.43 + 2.48 Sugar$.

c. For every additional gram of sugar in a serving of cereal, the expected calories increase by 2.48 calories.

1.20 a. There is a clear, linear, and strong relationship between list price and sale price, as the plot indicates.



b. The regression summary is given below.

```
   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 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8019 on 927 degrees of freedom
Multiple R-squared:  0.9894,Adjusted R-squared:  0.9894
```

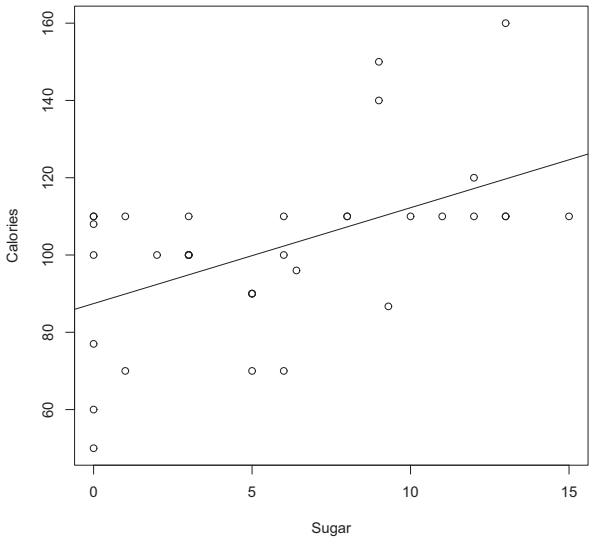
This shows us that the regression equation is $\widehat{SalePrice} = -144.8 + 0.943 ListPrice$.

c. Each increase of a dollar to the list price corresponds to a \$0.94 increase in sales price.

1.21 a. The prediction equation is $\widehat{Calories} = 87.43 + 2.48 Sugar$, so when $Sugar = 10$, the predicted $Calories$ is $\widehat{Calories} = 87.43 + 2.48(10) = 112.23$ calories.

b. For Cheerios, $\widehat{Calories} = 87.43 + 2.48(1) = 89.91$, so the residual is $110 - 89.91 = 20.09$ calories.

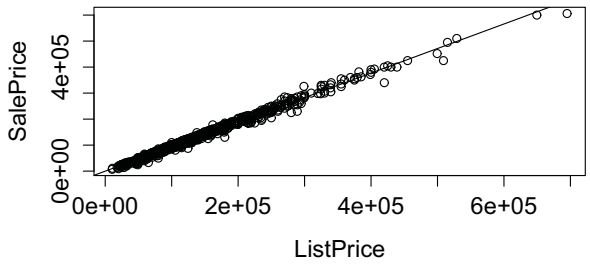
c. Although there is a somewhat positive association, there is still quite a bit of scatter away from the line.



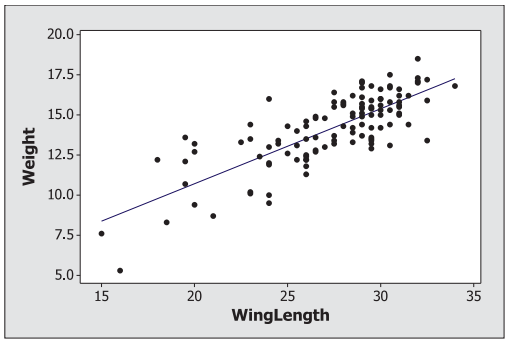
1.22 a. The prediction equation is $\widehat{SalePrice} = -144.8 + 0.943 ListPrice$, so when $ListPrice = 99,500$, the predicted $SalePrice$ is $\widehat{SalePrice} = -144.8 + 0.943(99,500) = 93,683.7$ dollars.

b. For the house at 1317 Prince St, $\widehat{SalePrice} = 93,683.7$, so the residual is $95,000 - 93,683.7 = 1316.3$ dollars.

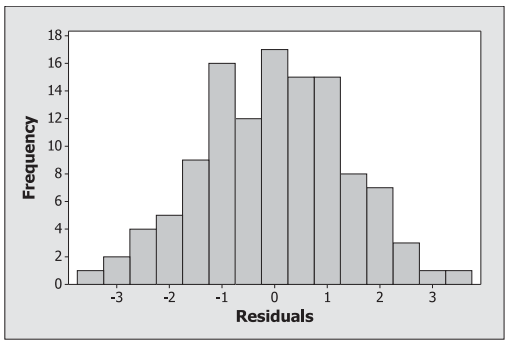
c. The relationship between list price and sales price is very strong and linear for the sample of houses from Grinnell, Iowa.



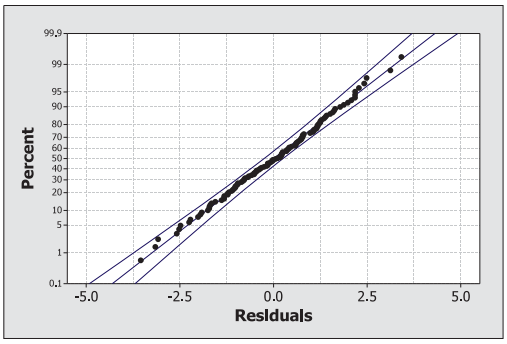
1.23 a. The scatterplot with the least squares line illustrates a very good fit and does not suggest any outliers or influential points.



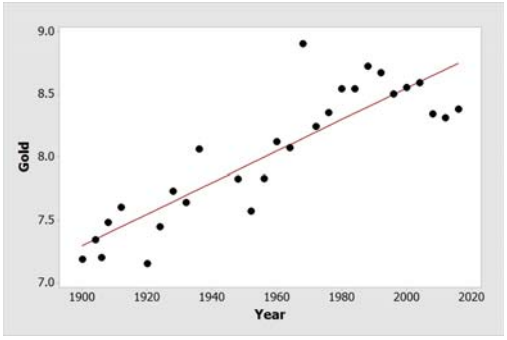
b. A histogram of the residuals shows a nice bell-shaped pattern centered at zero. Thus, the histogram does not reveal any problems with the conditions for this simple linear model.



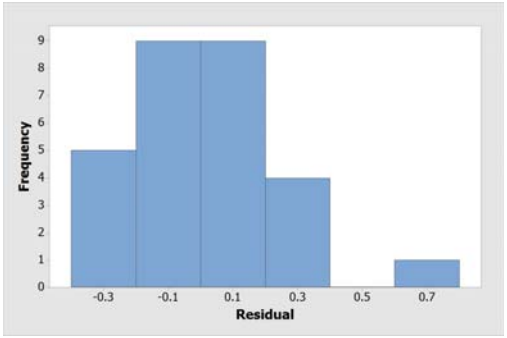
c. A normal probability plot shows a clear linear pattern. Thus, the residuals appear to follow a normal distribution.



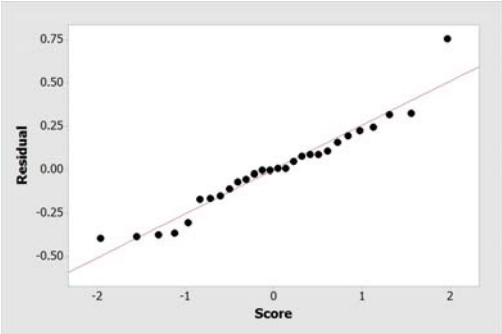
1.24 a. The scatterplot with the least squares line illustrates a good fit. There is one point that is higher than expected in 1968.



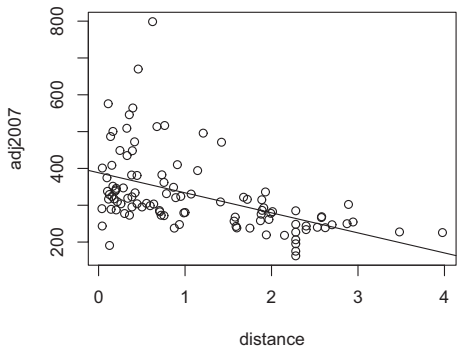
b. A histogram of the residuals shows a mostly nice bell-shaped pattern centered at zero. There is one larger residual that might cause one to worry.



c. A normal probability plot shows a clear linear pattern. Once again there is one residual that is larger and does not fit the pattern.



1.25 a. The scatterplot shows a weak downward trend; homes farther away from the bike trail tend to sell for less. The scatter about the trend line is great for homes near the trail and much smaller for homes far away from the trail.



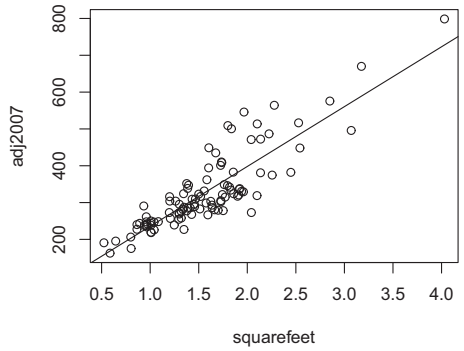
b. The equation of the best-fit line is $\widehat{Adj2007} = 388.204 - 54.427 Distance$. Each mile farther from a trail reduces, on average, the selling price by about 54,000 dollars.

```
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
```

c. The regression standard error is 92.13. If model conditions are met, then the average deviation from the line is about 92,000 dollars. Such a simple interpretation is compromised here because of the lack of consistent scatter about the line.

d. The model conditions are violated here because of the lack of consistent scatter about the line, as mentioned in part (a).

1.26 a. The scatterplot shows a fairly strong, positive, linear trend between *SquareFeet* and *Adj2007*.



b. The equation of the simple linear regression line is:

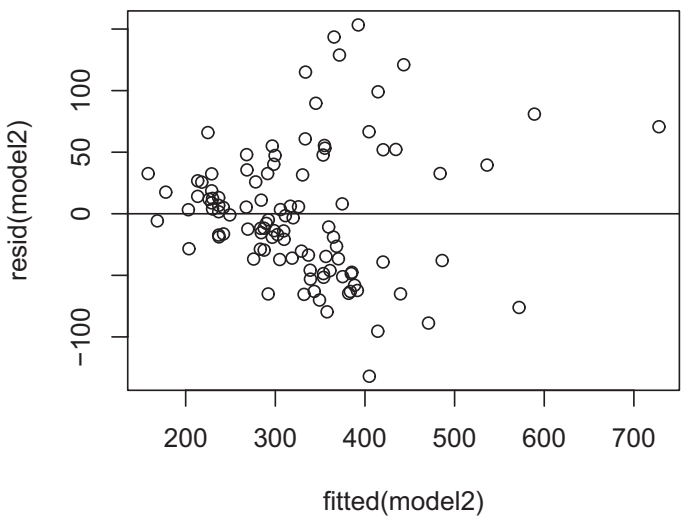
$\hat{Adj2007} = 72.973 + 162.526 SquareFeet$.

Each additional thousand squarefeet of floorspace is associated with an approximate added \$162,000 in selling price.

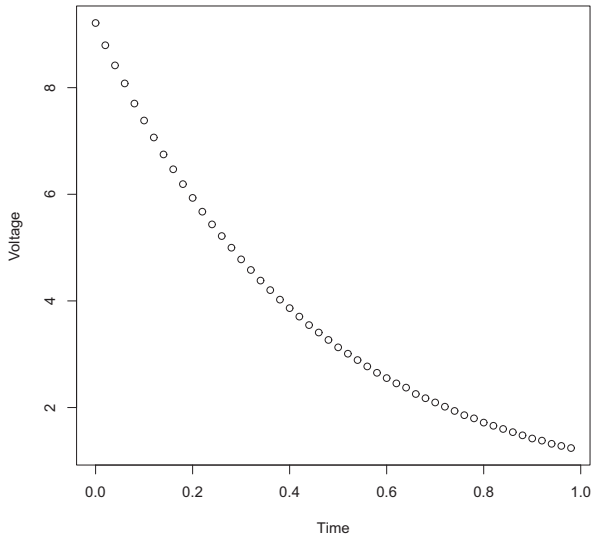
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   72.973     15.541   4.695 8.32e-06 ***
squarefeet   162.526      9.351  17.381  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

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
```

c. The regression standard error for this model is 53,000; on average, the line predicts selling price to within about 53,000 dollars of reality.

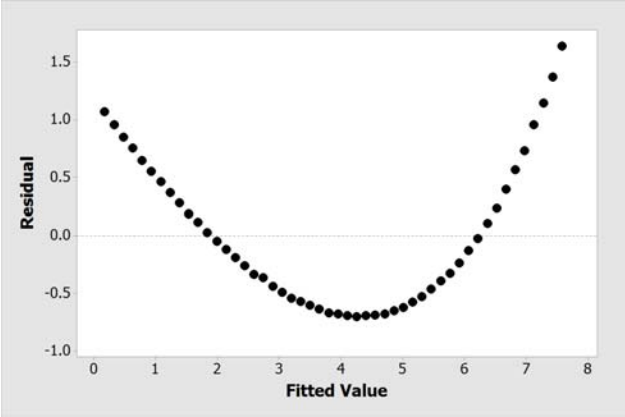
d. There is a slight nonconstancy of variance, as evidenced by the residual-versus-fit plot; larger homes are associated with larger residuals from the line.



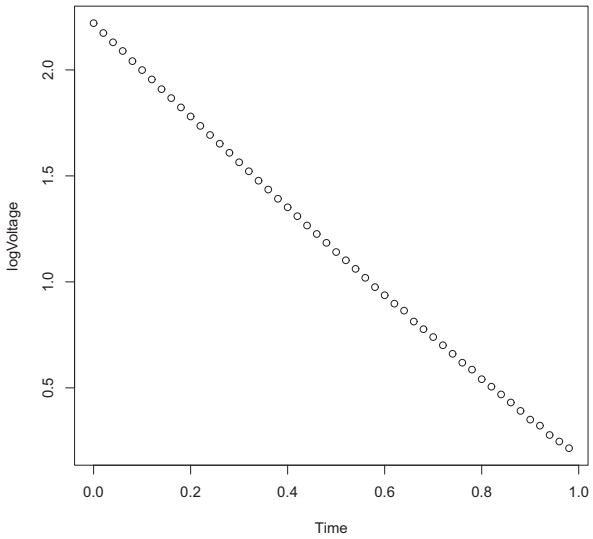
1.27 a. The scatterplot shows that as $Time$ increases, $Voltage$ goes down sharply. However, the decrease shows a nonlinear (curved) pattern.



b. The residual versus fits plot shows a clear curved pattern.



c. After creating a new variable, $logVoltage$, the scatterplot with $Time$ (below) is much more linear.

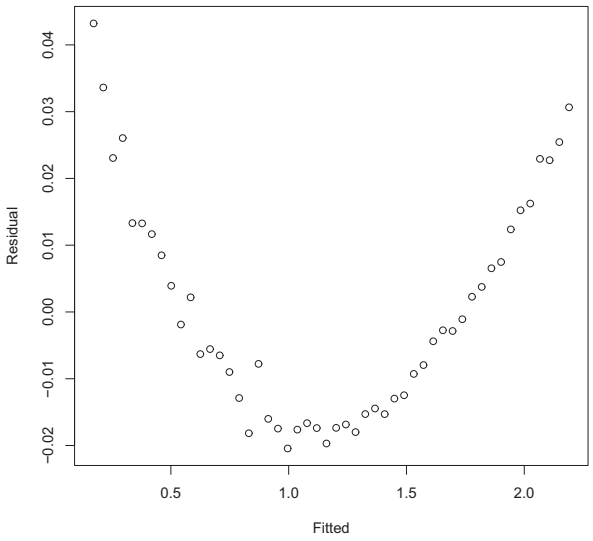


d. Fitting the regression line with technology gives the output

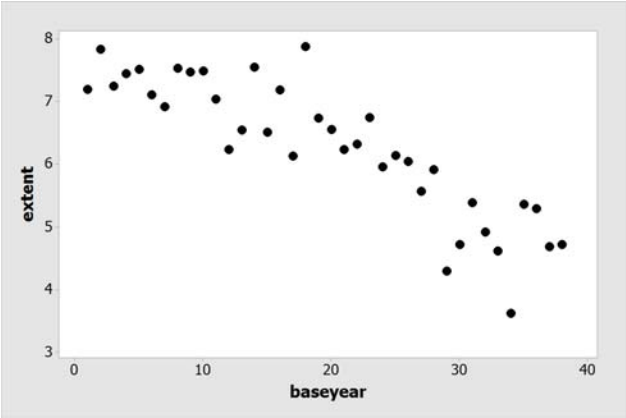
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.189945   0.004637   472.3   <2e-16 ***
Time        -2.059065   0.008154  -252.5   <2e-16 ***
```

This yields the prediction equation $\widehat{logVoltage} = 2.19 - 2.059Time$.

e. The plot of residuals versus fitted values for the model to predict $logVoltage$ shows a striking curved pattern in the residuals. The original (transformed) data have a mostly linear relationship, but some curvature remains after the dominant linear trend is removed. Using the regression model will give predictions that are too high in the middle and too low at the extremes of the $Time$ range.

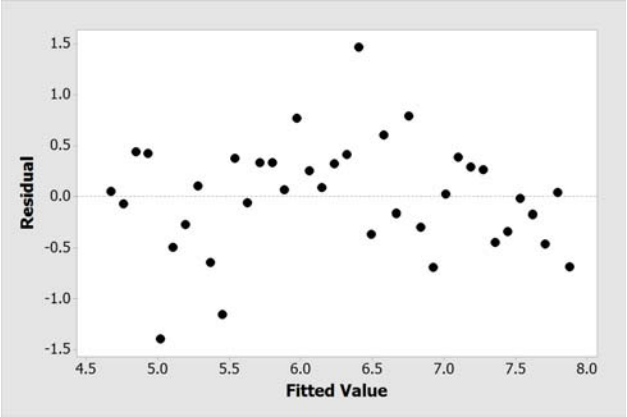


1.28 a. The scatterplot is below. There is clearly a reasonably strong, negative trend to this data. As the years go on, in general, the *Extent* of the sea ice is decreasing. This trend is not, however, linear. There is curvature to it that suggests that as time goes on the amount of decrease is increasing.

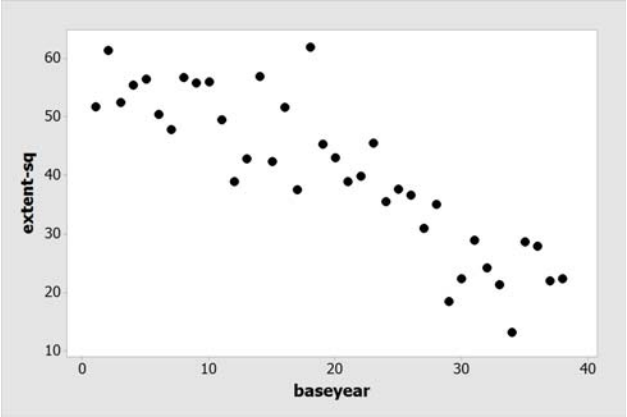


b. The residuals versus fits graph which follows also shows the curvature. In fact, it is somewhat

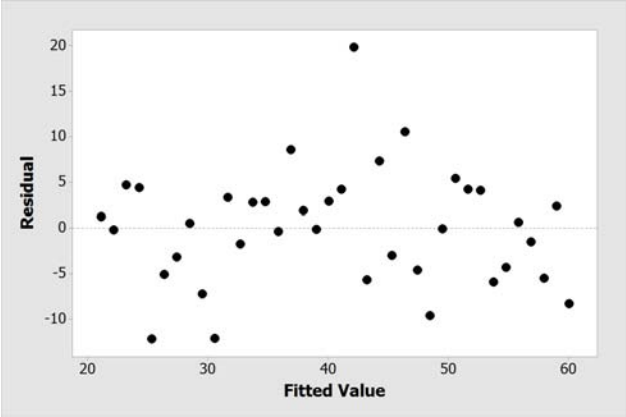
easier to see in this plot. Given this amount of curvature, we should not fit a linear model to this data.



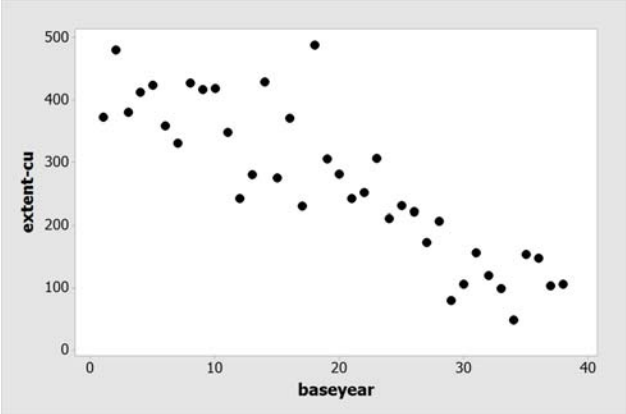
c. The scatterplot is below. While there is still some curvature, it is much less than in the scatterplot from part (a).

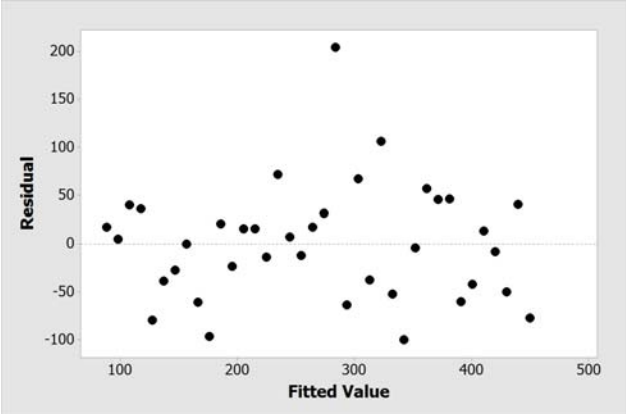


d. The residuals versus fits graph (given below) also shows that there is less curvature in this relationship.



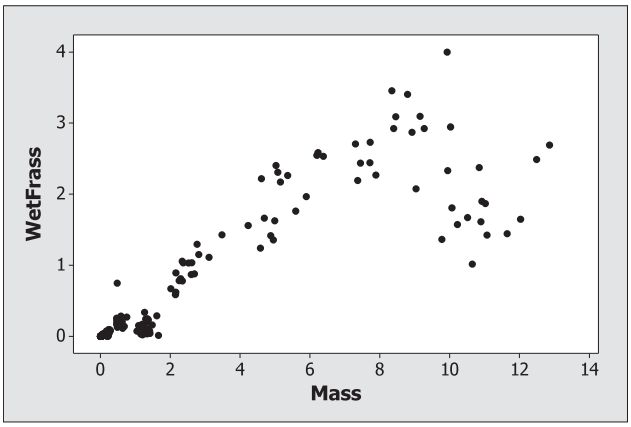
e. The scatterplot and residual plot are given below. In this case there seems to be a decent linear relationship. Very little curvature is evident in either plot.



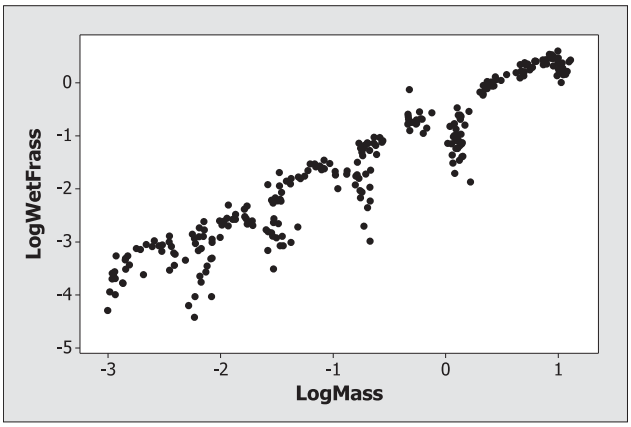


f. The most linear model is the one with the cube of *Extent*. The data is sufficiently linear that this model would be appropriate.

1.29 a. The scatterplot of $WetFrass$ versus $Mass$ shows clear curvature with more variability in the amount of wet frass for the larger caterpillars.



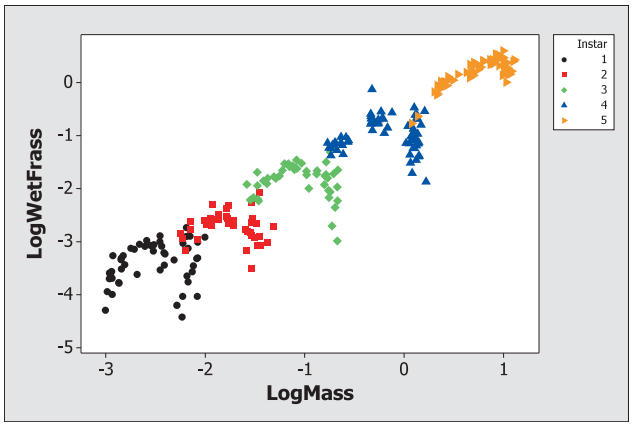
b. The scatterplot of $LogWetFrass$ versus $LogMass$ shows a strong positive association between the transformed variables, with intermittent periods of increased variability.



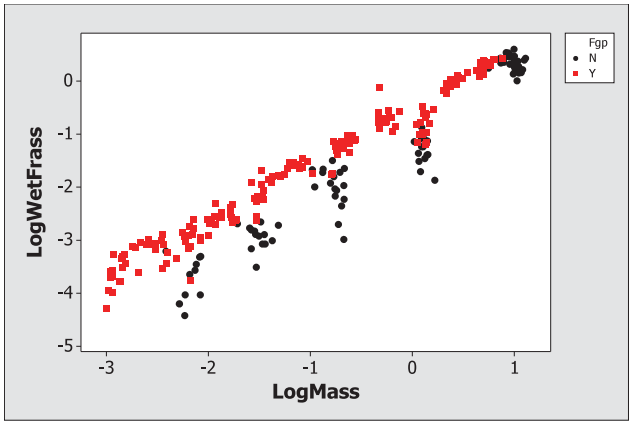
c. The log transformed variables show a more linear pattern. The fitted regression line for these variables is

$$\widehat{LogWetFrass} = -0.739 + 1.054 LogMass$$

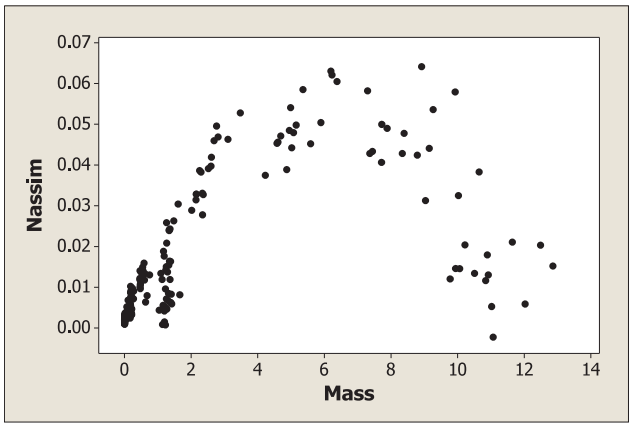
d. Here is a plot for the relationship with different symbols/colors for the five $Instar$ groups. There is curvature within the $Instars$, especially for the larger caterpillars in each group, but the linear model provides a good summary of the overall pattern for each $Instar$.



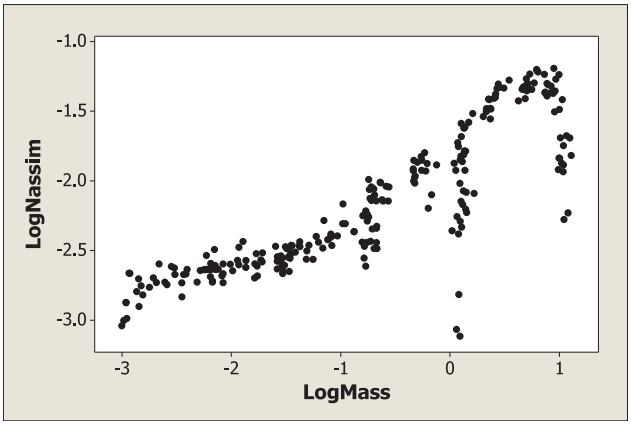
e. Here is a plot for the relationship with different symbols/colors for the free-growth and no-free-growth periods. Yes, the overall pattern is definitely more linear when the caterpillars are in a free-growth period.



1.30 a. The scatterplot of $Nassim$ versus $Mass$ shows clear curvature (perhaps quadratic) with more variability in nitrogen assimilation for the larger caterpillars.



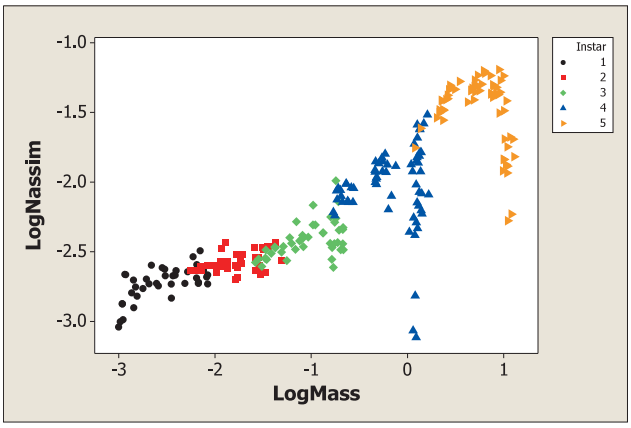
b. The scatterplot of $LogNassim$ versus $LogMass$ shows a strong positive association between the transformed variables, slightly curved but much more linear than the untransformed variables. There are a couple of intermittent periods of increased variability.



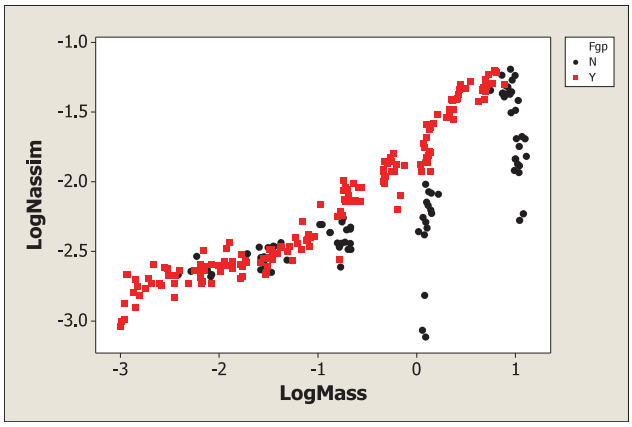
c. The log transformed variables show a more linear pattern. The fitted regression line for these variables is

$$\widehat{LogNassim} = -1.89 + 0.371 LogMass$$

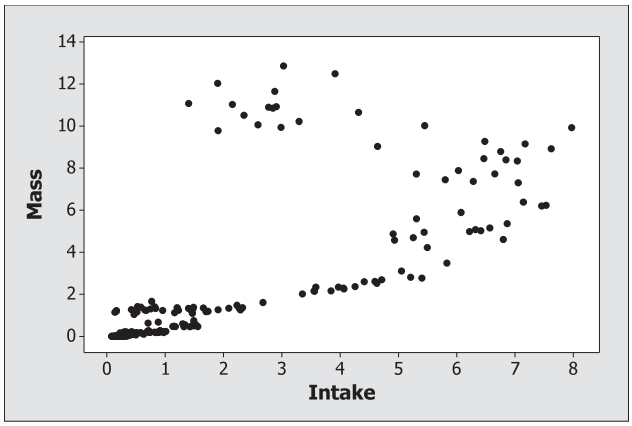
d. Here is a plot for the relationship with different symbols/colors for the five $Instar$ groups. There appears to be some curvature within some of the $Instars$, especially for $Instars$ 3, 4, and 5. However, the linear model provides a good summary of the overall pattern for the first two or three $Instars$.



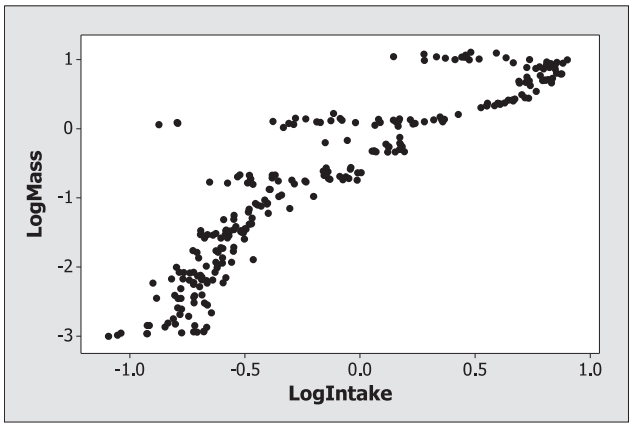
e. Here is a plot for the relationship with different symbols/colors for the free-growth and no-free-growth periods. Yes, the overall pattern is definitely more linear when the caterpillar is in a free-growth period. The curvature for $Instars$ 3, 4, and 5 is coming from the points when the caterpillars are NOT in a free-growth period.



1.31 a. The scatterplot of *Mass* versus *Intake* shows a nonlinear pattern—perhaps even two very different lines.

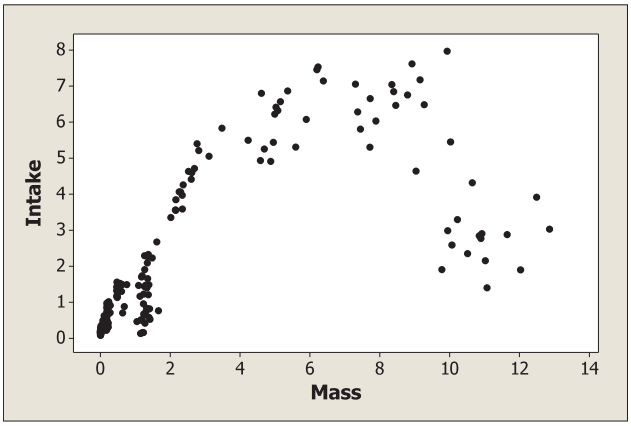


b. The scatterplot of *LogMass* versus *LogIntake* shows a more consistent positive association between the transformed variables with a slightly curved pattern that increases less steeply for larger values of *LogIntake*.

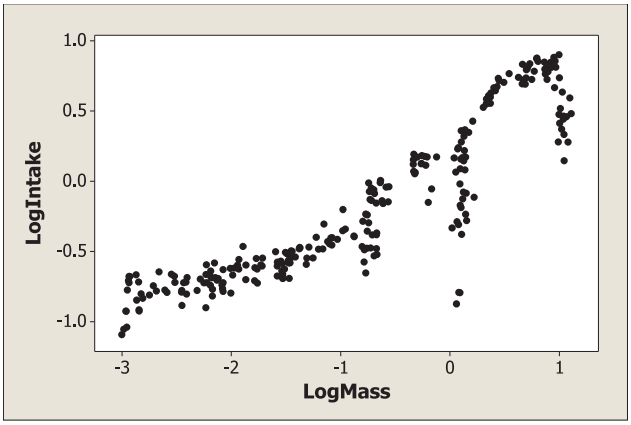


c. No, the association for the transformed variables is somewhat more linear, but the linear model does not appear like it would provide a good fit in either situation.

1.32 a. The scatterplot of *Intake* versus *Mass* shows substantial curvature with increasing variability in *Intake* as *Mass* increases.



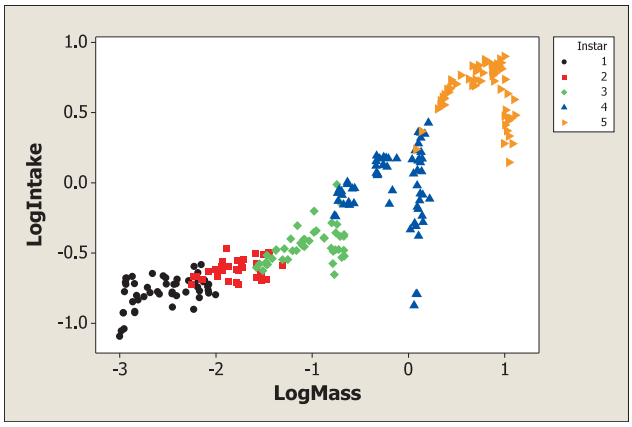
b. The scatterplot of *LogIntake* versus *LogMass* shows a more consistent positive association between the transformed variables, although there a several places that show increased variability and decreased values in *LogIntake* for relatively specific larger values of *LogMass*.



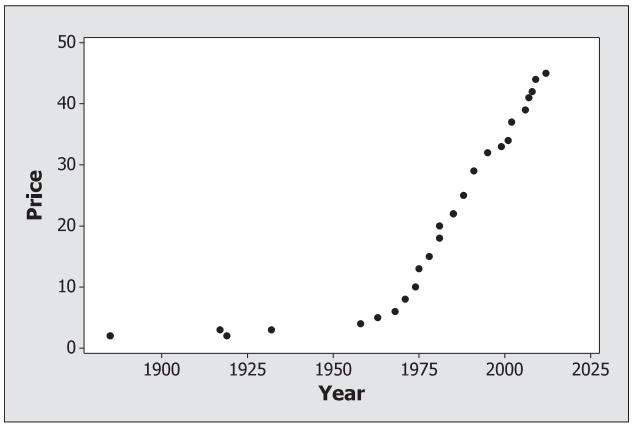
c. The log transformed variables show a more linear pattern. The fitted regression line for these variables is

$$\widehat{LogIntake} = 0.169 + 0.417 LogMass$$

d. Here is a plot for the relationship with different symbols/colors for the five *Instar* groups. There appears to be some curvature within some of the *Instars*, especially for *Instars* 3, 4, and 5. However, the linear model provides a good summary of the overall pattern for the first two or three *Instars*.



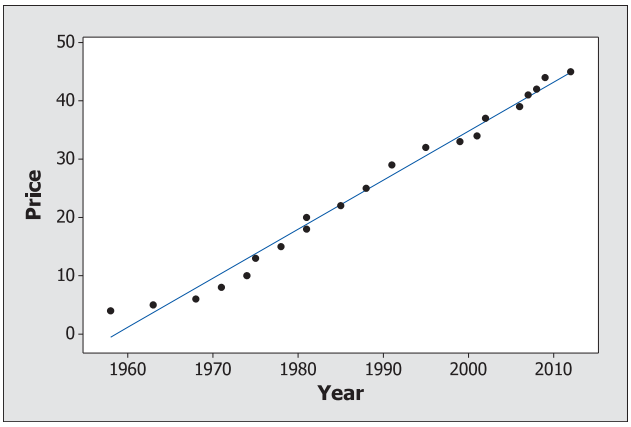
1.33 a. The scatterplot shows a strong positive linear association between $Price$ and $Year$. The first four points do not fit the overall linear pattern well, but the cost of mailing a letter must be greater than 0 cents!



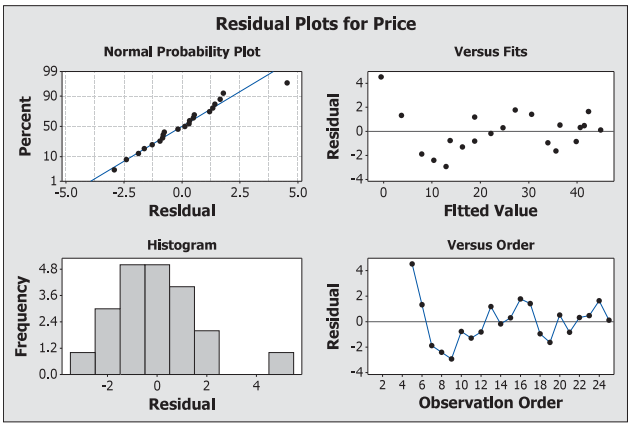
b. Here is some output for fitting the model after eliminating the first four observations. This shows the least squares line is $\widehat{Price} = -1647.17 + 0.841 Year$.

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

c. A plot of $Price$ versus $Year$ with the regression line after the first few points are omitted follows. The regression line appears to provide a very good fit. The first two prices are above the regression line and then the next five prices are below the regression line, but this regular pattern is not present for the rest of the points. The overall trend is clearly linear.

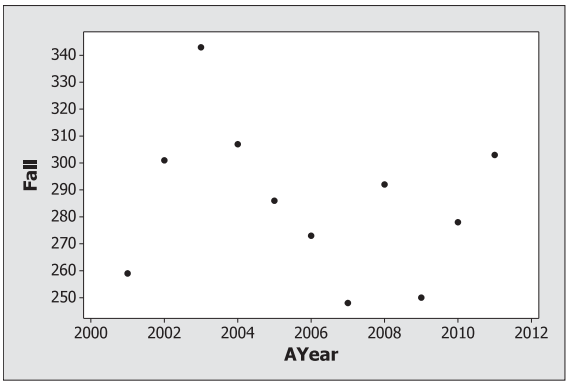


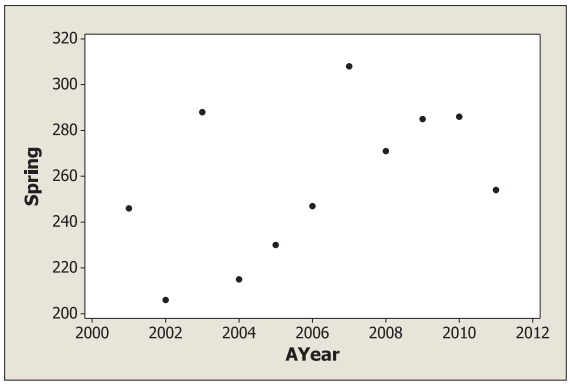
d. Several plots of residuals are shown as follows. The normal probability plot is linear with only one unusually large point in the top right corner, but the normality condition appears to be reasonable. The histogram of the residuals is roughly symmetric and centered around zero, with the exception of the one unusually large residual. The plot of the residuals against the fitted values illustrates the regular pattern for the first seven points, but then shows the unstructured pattern. The conditions appear to be reasonably well met for these data.



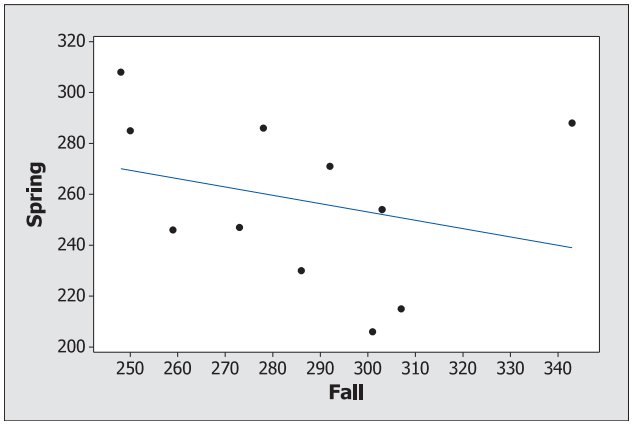
e. The largest residual is for the first year, 1958, where the stamp price was 4 cents and the predicted price based on the fitted model is $\widehat{Price} = -1647.17 + 0.841(1958) = -0.49$ to give a residual of $4-(-0.49) = 4.49$ (or a residual of 4.53 using software and more decimal places). Also using software, the standardized residual for 1958 is 2.95, which is somewhat unusually large.

1.34 a. Scatterplots for the relationship between *Enrollment* and *Year* are shown below for the spring and fall semesters. The overall trend for mathematics enrollments in the fall is very weak and slightly decreases over time. In the spring, the association is positive and moderate.



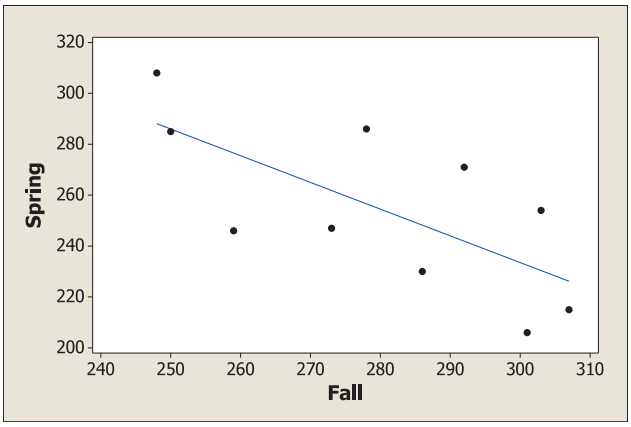


b. No, the overall association is negative, but weak. The following scatterplot shows the least squares line of $\widehat{Spring} = 351.1 - 0.3266Fall$, with an unusual point in the upper right.

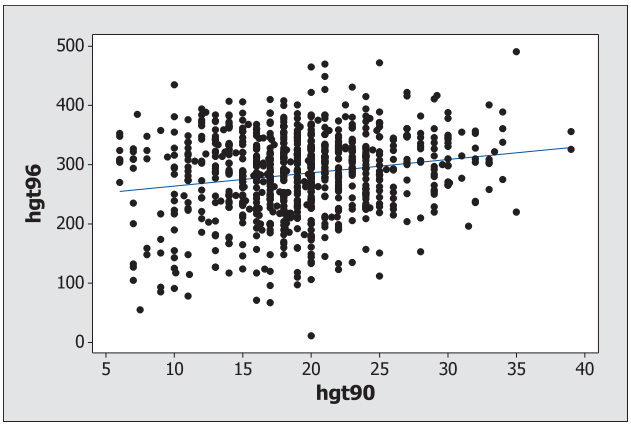


c. The third observation from 2003 has an unusually high fall enrollment of 343 and a spring enrollment of 288.

d. After removing the 2003 data, the association between fall and spring math enrollments looks much stronger. The least squares line without AYear = 2003 (shown as follows) is $Spring = 548 - 1.0483Fall$. The substantial changes in both the intercept and slope of the least squares line indicate that the enrollments in 2003 should be tagged as influential.



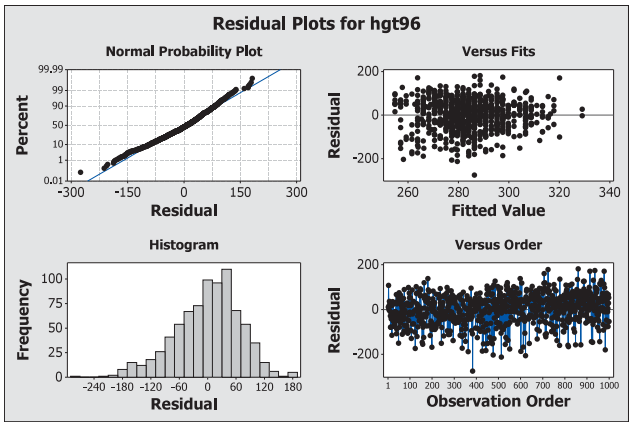
1.35 a. The scatterplot shown below displays a weak positive linear relationship between the initial seedling height and the height in 1996.



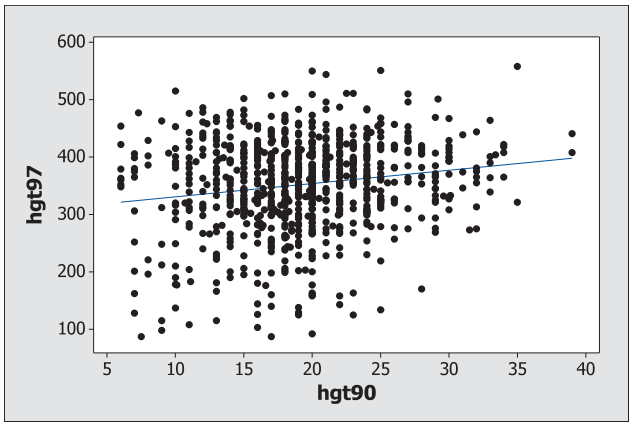
b. Here is some output for fitting the model for height in 1996 based on height in 1990. This shows that the least squares line is $\widehat{Hgt}96 = 241.3 + 2.250Hgt90$.

```
The regression equation is Hgt96 = 241 + 2.25 Hgt90

Predictor      Coef  SE Coef      T      P
Constant    241.285    8.621  27.99  0.000
Hgt90        2.2504   0.4311   5.22  0.000
```

c. The preceding scatterplot and some residual plots that follow show that there is considerable variation around the least squares line, with a regression standard error of 69.0173. The normal probability plot is roughly linear, with one unusually small residual, but otherwise the normality condition is met. Overall, the conditions for the linear model are met, and the linear model provides a reasonable fit.



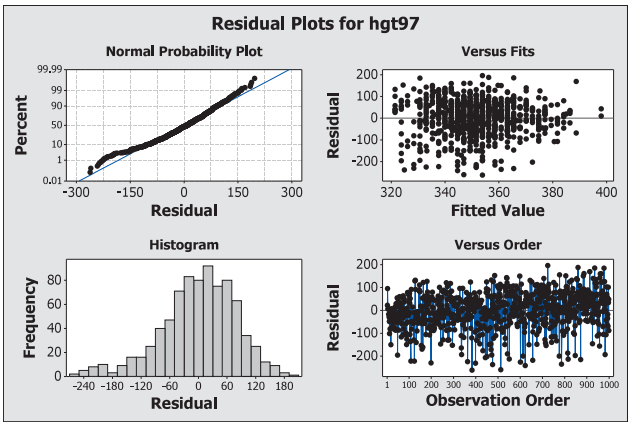
1.36 a. The scatterplot shown below displays a weak positive linear relationship between the initial seedling height and the height in 1997.



b. Here is some output for fitting the model for height in 1997 based on height in 1990. This shows that the least squares line is $\widehat{Hgt}97 = 307.44 + 2.3224Hgt90$.

```
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
```

c. The preceding scatterplot and some residual plots that follow show that there is considerable variation around the least squares line, with a regression standard error of 78.79. The normal probability plot is roughly linear, with very slight curvature in the tails. Overall, the normality and constant variance conditions for the linear model are met, and the linear model provides a reasonable fit.

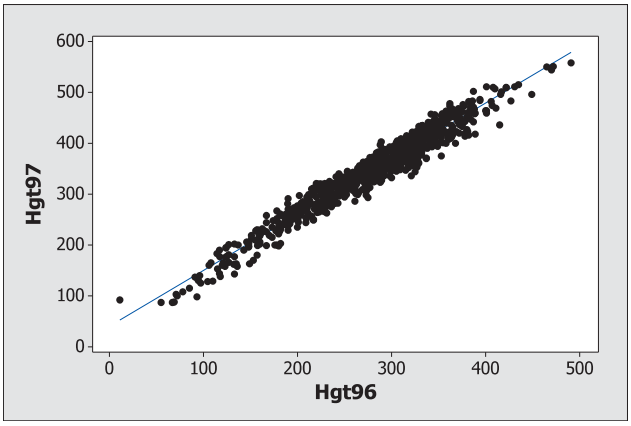


1.37 a. Yes, there is only one year of growth between the heights in 1996 and 1997, so the linear relationship should be much stronger than the relationship between the initial seedling height and the height in 1997.

b. Here is some output for fitting the model for height in 1997 based on height in 1996. This shows that the least squares line is $\widehat{Hgt}97 = 40.6 + 1.10Hgt96$.

```
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
```

c. Yes, there is a strong, positive, linear relationship between the heights in 1996 and 1997. The regression standard error is 18.4653, and the heights are tightly clustered around the least squares line.



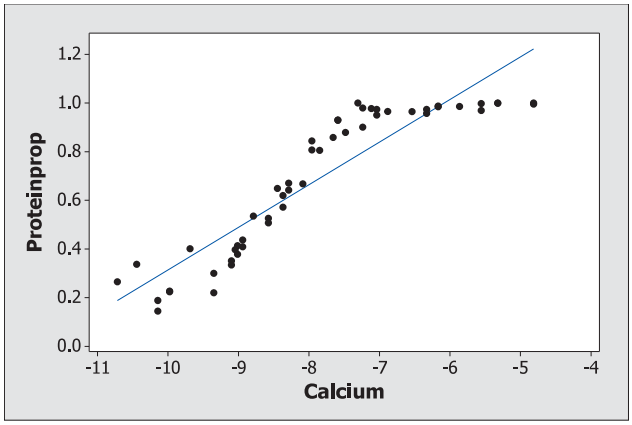
1.38 Following is some output for fitting the model for *ProteinProp* based on *Calcium*.

```
The regression equation is Proteinp = 2.07 + 0.175 Calcium

Predictor     Coef  SE Coef      T      P
Constant   2.06586  0.08876  23.28  0.000
Calcium    0.17514  0.01107  15.82  0.000

S = 0.119866   R-Sq = 83.6%   R-Sq(adj) = 83.3%
```

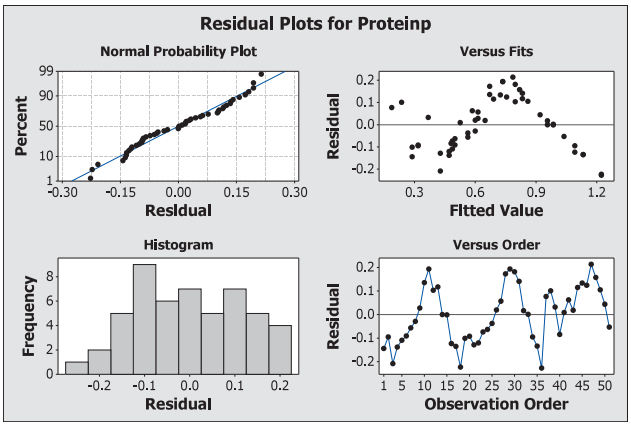
a. In the output, we see that the least squares line is $\widehat{ProteinProp} = 2.0659 + 0.1751 Calcium$.

b. In the output, we see that the regression standard error is $\widehat{\sigma}_\epsilon = 0.119866$.

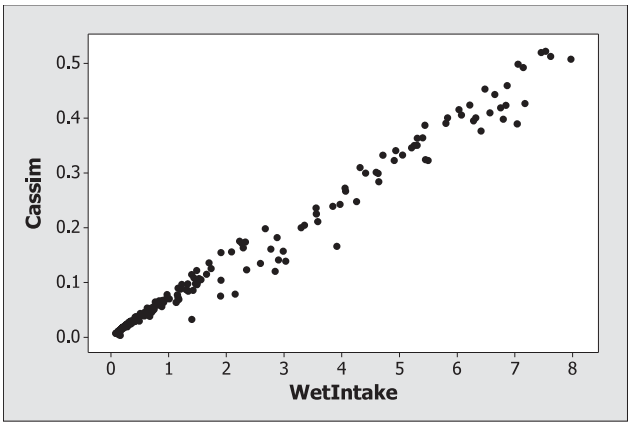
c. A scatterplot with the regression line is shown as follows. The regression line does not provide a good fit. The overall pattern shows some curvature and a more complex model would probably work better.



d. The linearity condition is not met. The plot of the residuals against the fitted value shows a very clear pattern, which indicates that a more complicated model might be needed. The normal probability plot shows some slight departures from linear trend in the tails, but the overall pattern is linear, so the normality condition does seem reasonable. The histogram of the residuals is very roughly symmetric and centered at zero. The plot of residual against order shows a very clear pattern, which indicates that the residuals are not independent of time order.



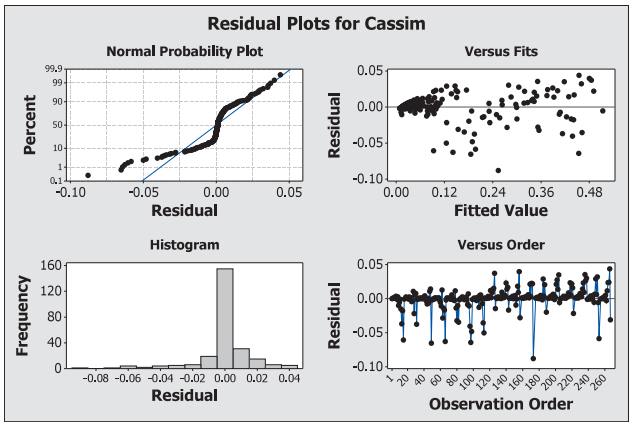
1.39 a. The scatterplot below shows a strong positive association between *Cassim* and *Intake*.



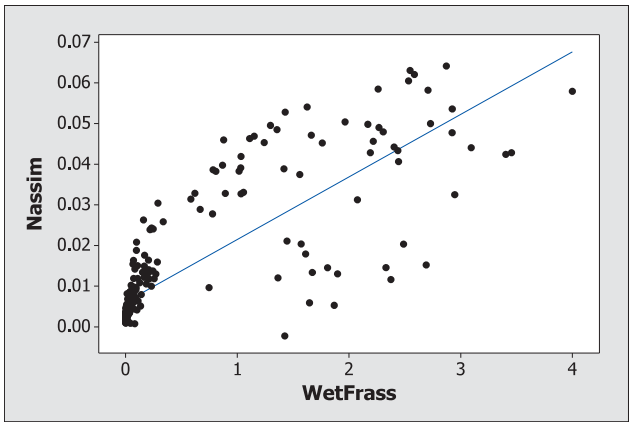
b. Here is some output for fitting the model to predict *Cassim* based on *Intake*. This shows that the least squares line is $\widehat{Cassim} = 0.00379 + 0.0639 Intake$.

```
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
```

c. No, the conditions for inference are not met. The plot of residuals against fitted values shows that the variance is not constant, rather it increases for larger values of *Cassim*. The normal probability plot shows clear departures from a linear trend, indicating a lack of normality, which is also reflected in the histogram of the residuals that is skewed to the left.



1.40 a. The scatterplot that follows shows a positive association between $Nassim$ and $WetFrass$ that is strong for small values of $WetFrass$, but less strong with much more variability for larger values of $WetFrass$.

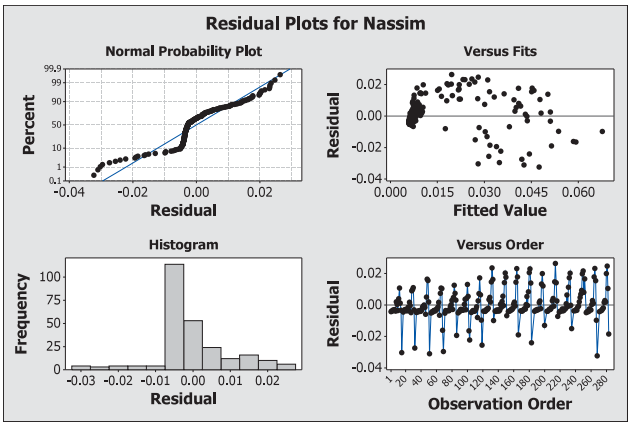


b. Here is some output for fitting the model for $Nassim$ based on $WetFrass$. This shows that the least squares line is $\widehat{Nassim} = 0.00606 + 0.0154 WetFrass$.

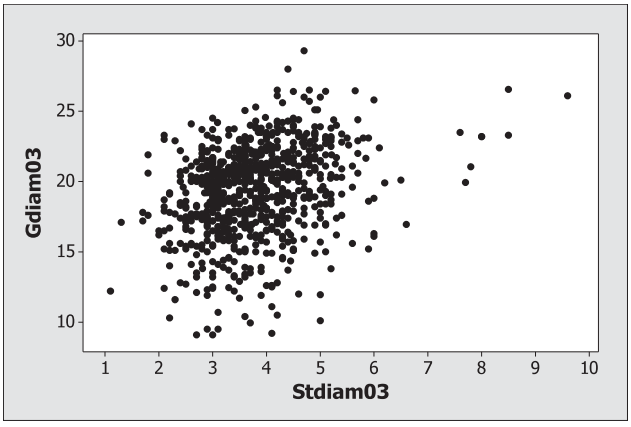
```
The regression equation is Nassim = 0.00606 + 0.0154 WetFrass

Predictor       Coef     SE Coef       T      P
Constant   0.0060618   0.0006913    8.77  0.000
WetFrass   0.0153991   0.0006830   22.55  0.000
```

c. No, the conditions for inference are not met. The plot of residuals against fitted values shows that the variance is not constant, rather it increases for larger predicted values of $Nassim$. The normal probability plot shows clear departures from a linear trend, indicating a lack of normality. This is also reflected in the histogram of the residuals, which is not bell-shaped.

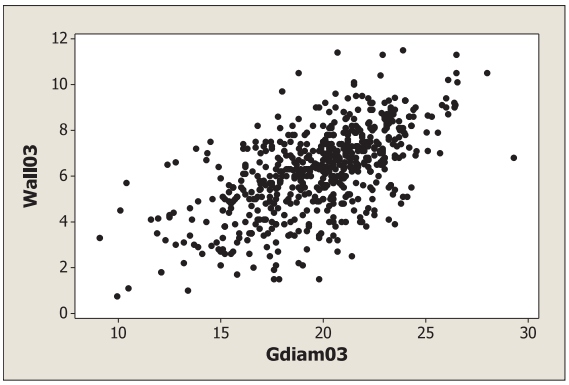
Also, the plot of residuals versus data order shows a regular, repeating pattern of increasing values followed by one big decrease.

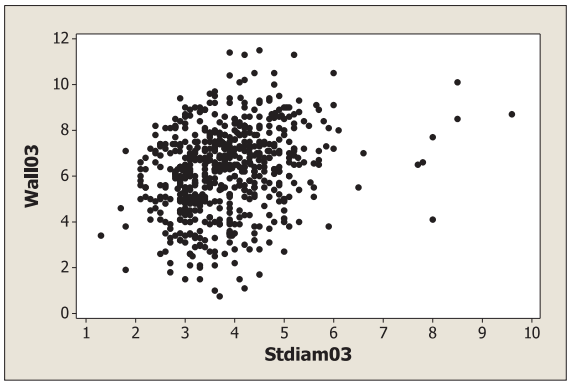


1.41 a. The scatterplot below shows that stem diameter and gall diameter in 2003 are positively associated, but the association is weak.



b. The scatterplots below show that wall thickness in 2003 has a stronger linear relationship with gall diameter than with stem diameter.





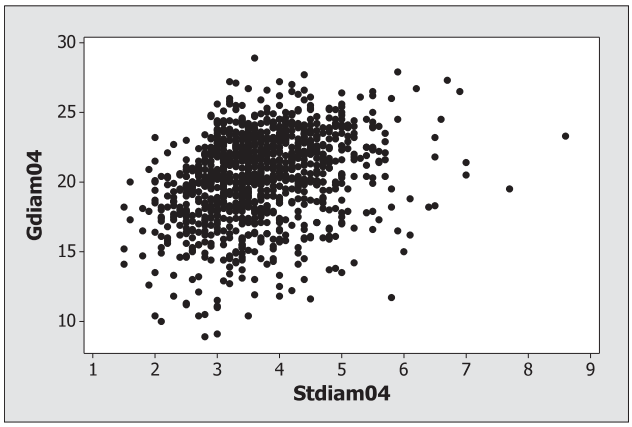
c. Here is some output for fitting the model for predicting $Wall03$ based on $Gdiam03$. It shows that the least squares line is $\widehat{Wall}03 = -1.0521 + 0.36821 Gdiam03$.

```
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%
```

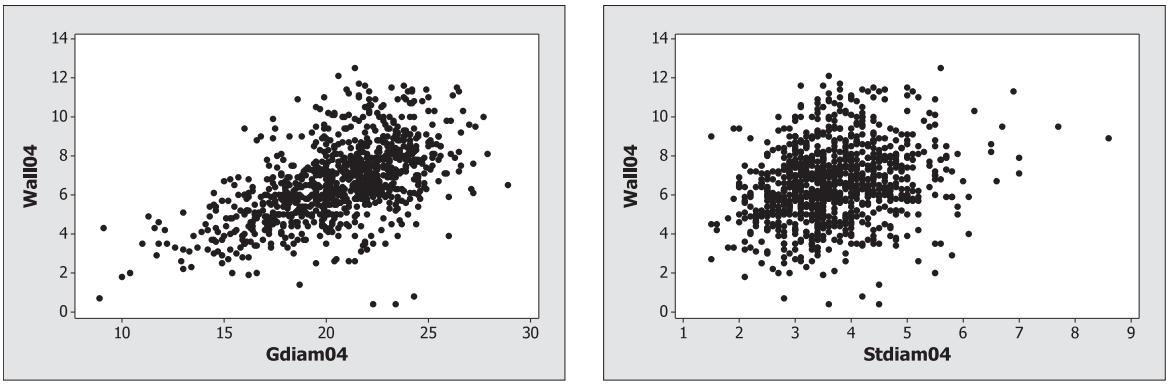
d. The fitted value when $Gdiam03 = 20.7$ is $\widehat{Wall}03 = -1.0521 + 0.36821(20.7) = 6.57$, and the residual is $6 - 6.57 = -0.57$.

e. We see in the output of part (c) that the regression standard error (that estimates the magnitude of a typical error) is $\widehat{\sigma}_\epsilon = 1.50$.

1.42 a. The scatterplot below shows that stem diameter and gall diameter in 2004 are positively associated, but the association is weak.



b. The scatterplots that follow show that wall thickness in 2004 has a stronger linear relationship with gall diameter than with stem diameter.



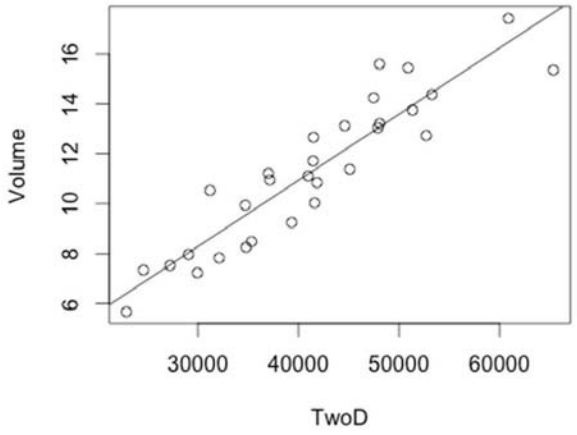
c. Here is some output for fitting the model for predicting $Wall04$ based on $Gdiam04$. It shows that the least squares line is $\widehat{Wall}04 = -0.845 + 0.3632 Gdiam04$.

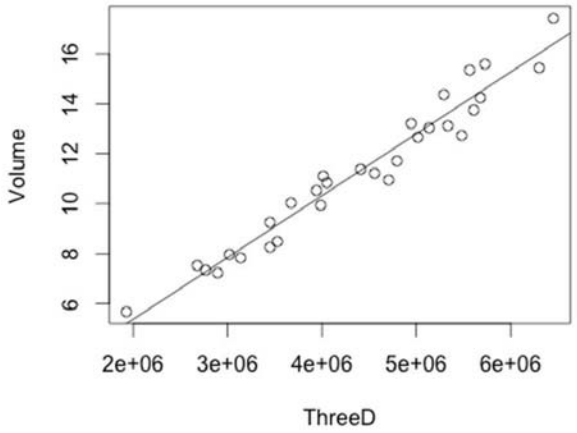
```
The regression equation is Wall04 = - 0.845 + 0.363 Gdiam04

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%
```

d. The first measurement for the wall thickness is missing in 2004, so we need to use the first observation with measurements on both variables. The fitted value when $Gdiam04 = 23.1$ is $\widehat{Wall}04 = -0.845 + 0.3632(23.1) = 7.54$, and the residual is $9.4 - 7.54 = 1.86$.

e. We see in the output of part (c) that the regression standard error (that estimates the magnitude of a typical error) is $\widehat{\sigma}_\epsilon = 1.61$.

1.43 a. Below you will find fitted line plots for both the 2-D and 3-D models. From these we see that both linear fits are tight, but that 3-D is clearly a bit tighter fit. Both correlations are high and both relationships look to be linear.



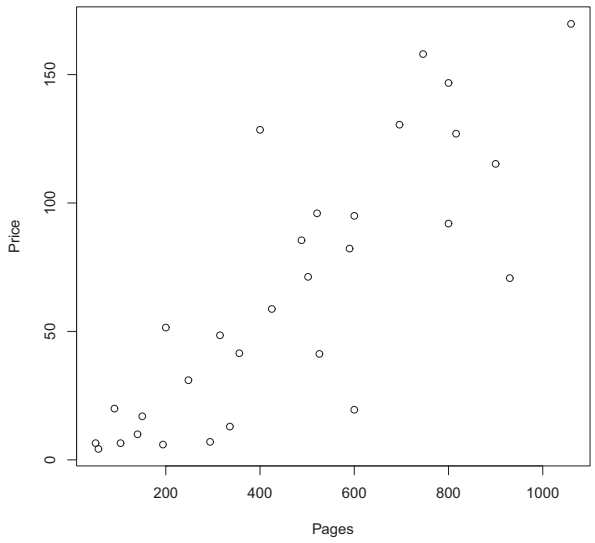


b. The regression tables are given below. The typical size of an error when predicting with 2-D—the standard error for regression (below called the residual standard error)—is 1.183. For 3D this value is 0.6488. So 3-D makes more precise predictions. Also the 3-D model wins the R-squared contest: 95.37% versus 84.61%.

```
2D summary:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.367e-01  9.076e-01   0.371    0.713
TwoD        2.649e-04  2.135e-05  12.406 6.77e-13 ***
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.183 on 28 degrees of freedom
Multiple R-squared:  0.8461,Adjusted R-squared:  0.8406
F-statistic: 153.9 on 1 and 28 DF,  p-value: 6.77e-13

3D summary:

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.196e-01  4.671e-01   0.898    0.377
ThreeD      2.475e-06  1.031e-07  24.019   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6488 on 28 degrees of freedom
Multiple R-squared:  0.9537,Adjusted R-squared:  0.9521
F-statistic: 576.9 on 1 and 28 DF,  p-value: < 2.2e-16
```

1.44 a. As pages go up, price goes up. There is a linear trend evident here, although the points do not cluster tightly.

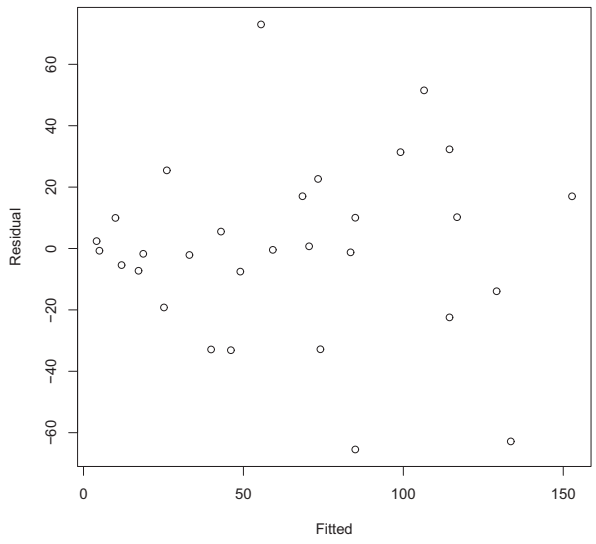


b. Fitting the regression line with technology gives the output

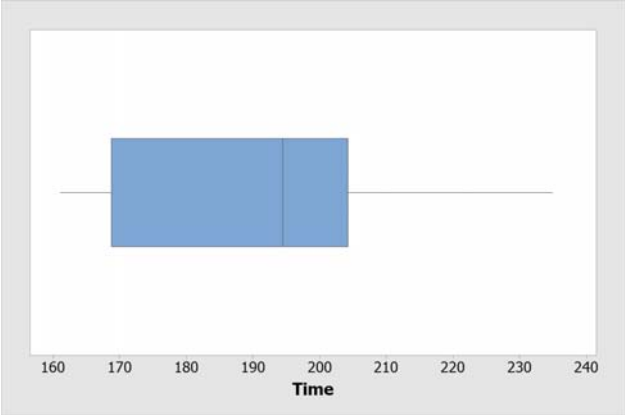
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.42231   10.46374  -0.327    0.746
Pages        0.14733    0.01925   7.653 2.45e-08 ***
```

This yields the prediction equation $\widehat{Price} = -3.42 + 0.1473 Pages$.

c. A plot of residuals versus fitted values for the regression of *Price* on *Pages* is shown on the next page. The linearity condition is met, as there is no trend in the residuals. However, there is something of a megaphone pattern here, with larger variability for large predictions (i.e., high page and price values) than for small predictions (low page and price values). Thus, the homoscedasticity condition is somewhat in doubt—although things don't look too bad, as the spread in the residuals is fairly constant when *Pages* is above 60.



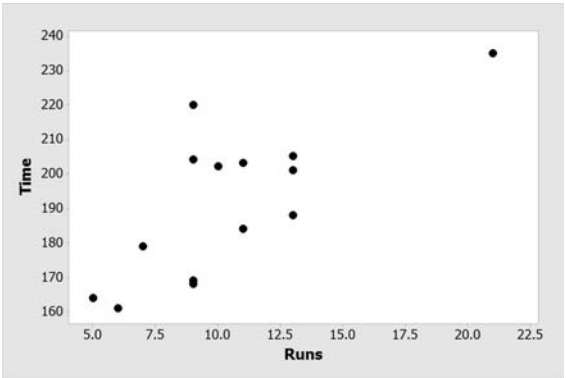
1.45 a. A boxplot of the *Time* variable shows a reasonably symmetric distribution.

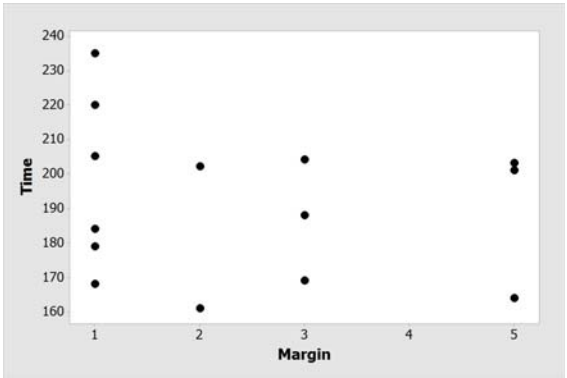


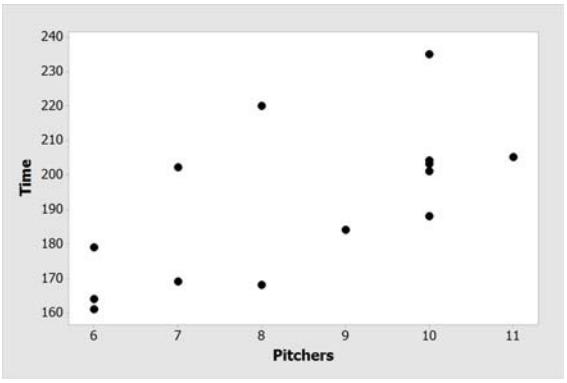
Some summary statistics for the sample of 15 game times are given in the table that follows.

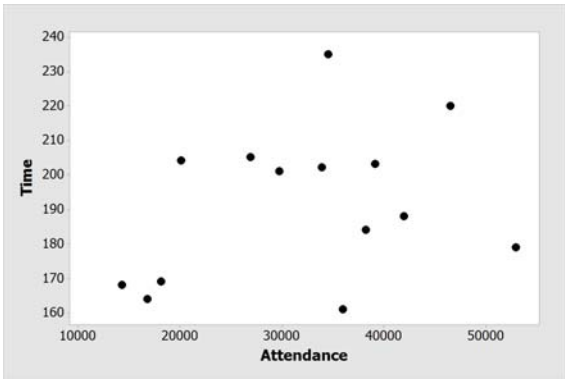
min	Q_1	median	mean	Q_3	max	std dev
161.0	168.8	194.5	191.6	204.3	235.0	22.1

b. Scatterplots for each of the potential predictors with *Time* are shown below. The strongest linear pattern among these plots is between *Time* and number of *Runs*. The next best predictor of *Time* would be *Pitchers*. Neither *Margin* or *Attendance* show much of a linear relationship with *Time* in these scatterplots.







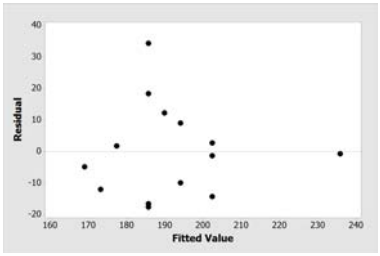


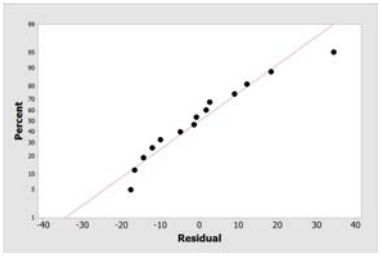
c. Fitting the regression line to predict $Time$ based on $Runs$ with technology gives the output

```
Coefficients
Term      Coef  SE Coef  T-Value   P-Value
Constant 148.0     12.0    12.34     0.000
Runs      4.18     1.08     3.87     0.002
```

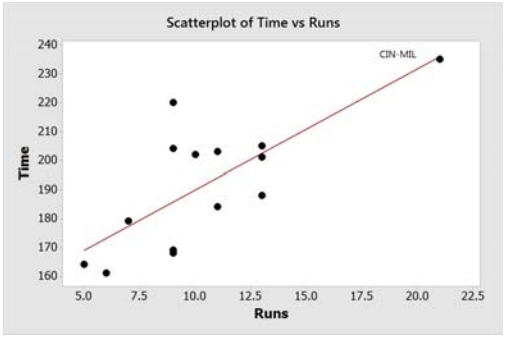
This yields the prediction equation $\widehat{Time} = 148.0 + 4.18Runs$. The slope indicates that for every extra run in a game we expect the average game time to increase by about 4.2 minutes.

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.



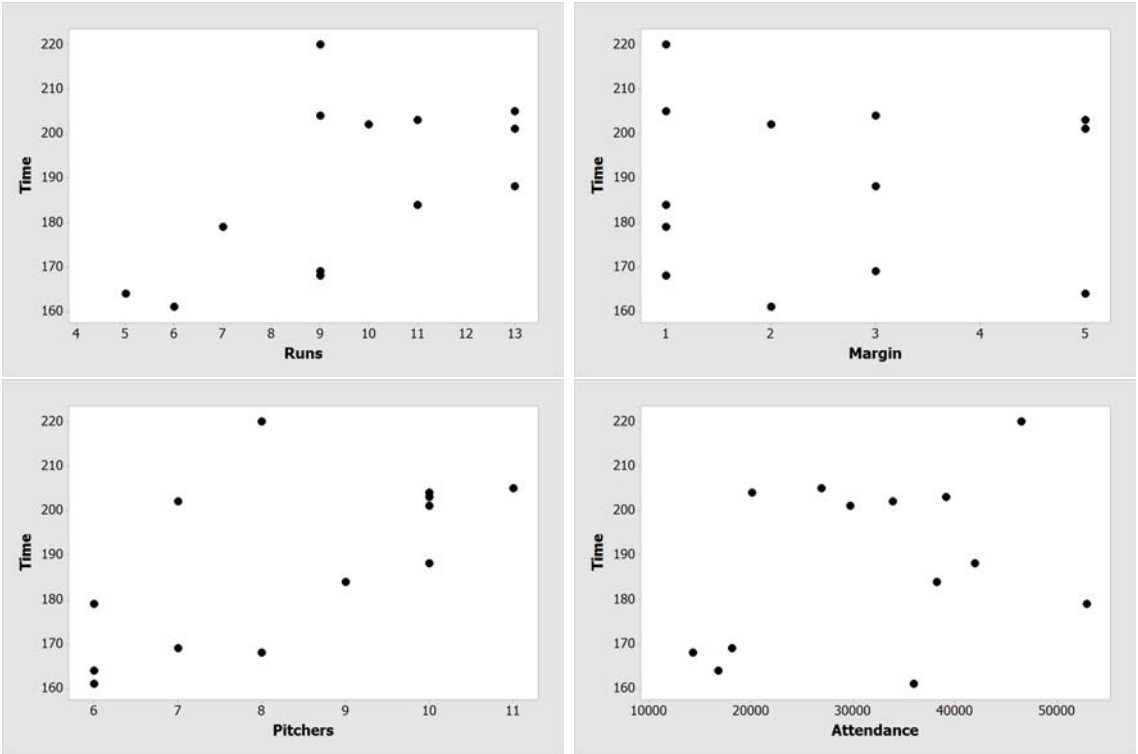


1.46 a. Following is the scatterplot with the CIN-MIL point highlighted. This, while quite far away from the bulk of the data, seems to follow in the same pattern, so may not be very influential with respect to this linear relationship.



b. Without the potential influential point, the least squares regression line becomes $\widehat{Time} = 147.0 + 4.3Runs$. The previous equation (using the CIN-MIL data point) was $\widehat{Time} = 148.0 + 4.18Runs$. There is very little change to either the slope or the y-intercept so we find that this particular point does not appear to be very influential.

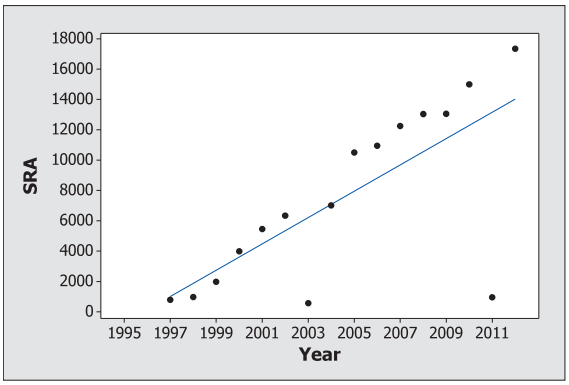
c. Scatterplots for each of the potential predictors (excluding CIN-MIL) with $Time$ are shown below. Now there is a tossup for which variable has the strongest linear relationship with $Time$. It could be $Runs$ or $Pitchers$. So, while the CIN-MIL game did not have much influence on the least squares regression equation for predicting $Time$ from $Runs$, its presence did suggest a stronger relationship between those two variables than otherwise would have been there.

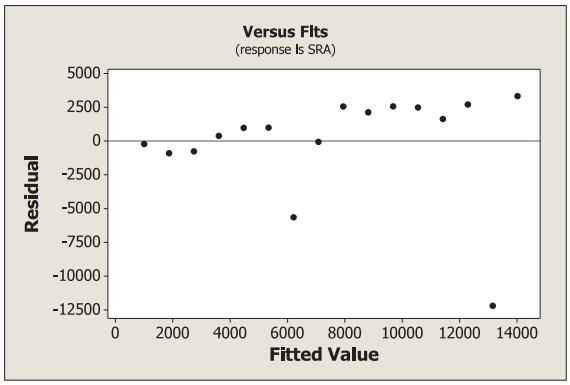


1.47 a. Here is some output for fitting the model for predicting SRA based on $Year$. It shows that the least squares line is $\widehat{SRA} = -1,732,400 + 868Year$.

```
The regression equation is SRA = - 1732400 + 868 Year

Predictor      Coef  SE Coef      T      P
Constant   -1732400   439865  -3.94  0.001
Year           868.0    219.4   3.96  0.001

S = 4046.24   R-Sq = 52.8%   R-Sq(adj) = 49.4%
```

Using technology, we find the residuals for the two sabbatical years are -5642.7 in 2003 and $-12,201$ in 2011. The scatterplot and residual plots that follow both show that these two points are unusual. These two points lie way below the overall linear pattern for the other points.





To standardize each of these residuals, we can divide by the regression standard error ($\hat{\sigma}_\epsilon = 4046.24$ in the output).

$$2003: \quad \frac{-5642.7}{4046.24} = -1.39 \qquad 2011: \quad \frac{-12201}{4046.24} = -3.02$$

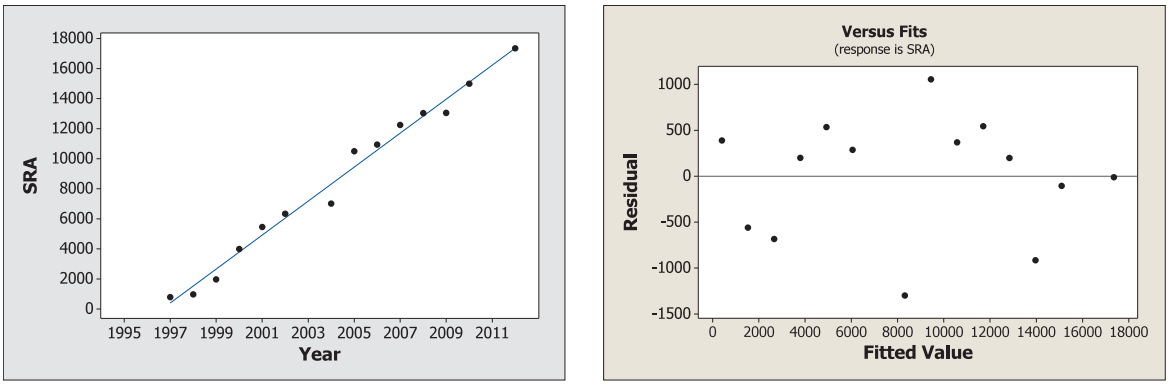
or we can use the slightly different standardized residuals provided by software (-1.45 in 2003 and -3.34 in 2011). These show that the 2003 residual is not so unusual (not beyond -2), but the 2011 residual should be considered an outlier (beyond -3).

b. Here is some output for fitting the model for predicting SRA based on $Year$ after removing the data for the sabbatical years of 2003 and 2011. It shows that the least squares line changes to $\widehat{SRA} = -2,257,997 + 1131Year$.

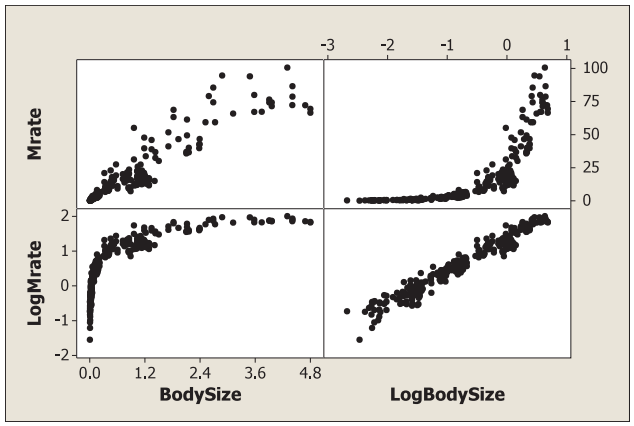
```
The regression equation is SRA = - 2257997 + 1131 Year
```

```
Predictor       Coef  SE Coef       T      P
Constant   -2257997    78905  -28.62  0.000
Year        1130.89    39.37   28.72  0.000

S = 674.742   R-Sq = 98.6%   R-Sq(adj) = 98.4%
```

This model provides a much better fit for the annual SRA contributions. The substantial changes to both the slope and intercept of the regression line indicate that the two sabbatical years are influential. The scatterplot and the residual plots (shown as follows) indicate a much stronger linear pattern, with much less variation from the regression line. The regression standard error has dropped from 4046.24 to 674.74.



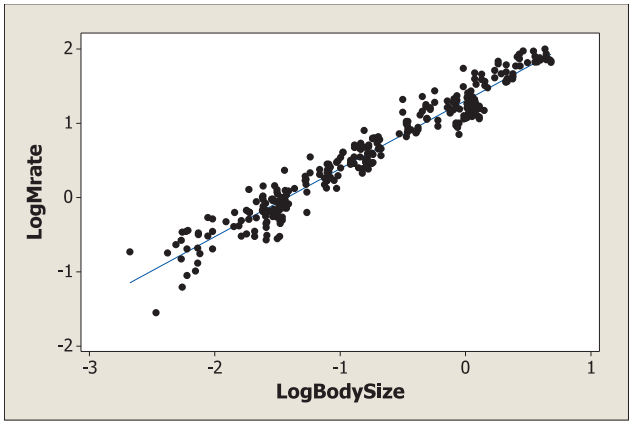
1.48 a. The following plot shows scatterplots for each of the possible response variables ($MRate$ and $LogMrate$) with $BodySize$ and $LogBodySize$. Note: With some software, you might need to produce these scatterplots individually.



The most appropriate relationship among these for a linear model is $Y = LogMrate$ versus $X = LogBodySize$. Here is some output for fitting this model and a scatterplot with the

regression line, $\widehat{LogMrate} = 1.31 + 0.916 LogBodySize$. This appears to be a good model for summarizing the relationship between these variables

```
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%
```



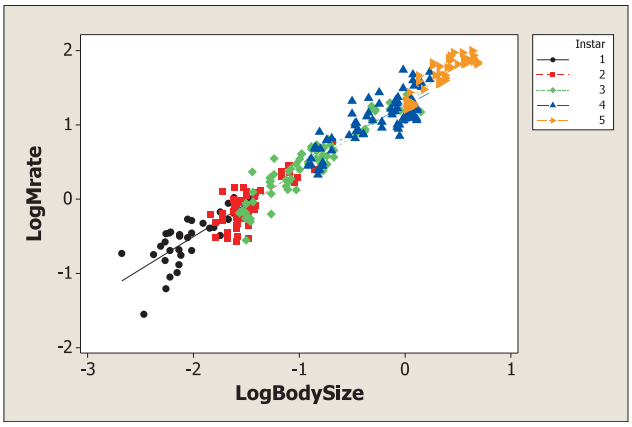
b. To predict the metabolic rate for a caterpillar with a body size of 1 gram, we first find $LogBodysize = log_{10}(1) = 0$, so the predicted log of the metabolic rate is

$$\widehat{LogMrate} = 1.30655 + 0.91641(0) = 1.30655$$

Since the logs in this situation are base 10, we find the predicted metabolic rate with

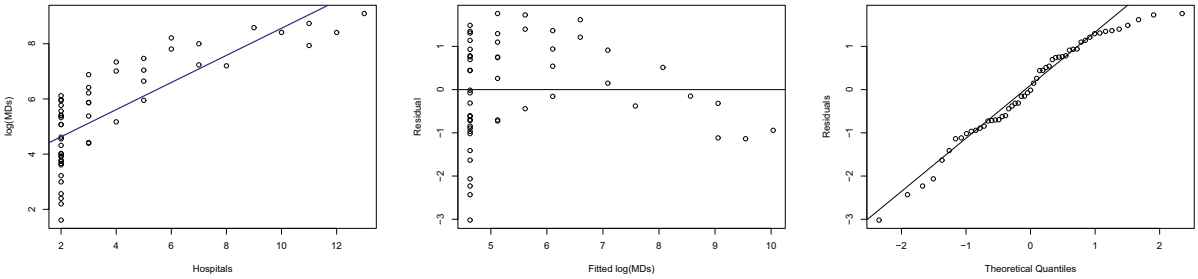
$$\widehat{Mrate} = 10^{1.30655} = 20.3$$

1.49 Here is a plot of $LogMrate$ versus $LogBodySize$ with different symbols/colors for the five levels of $Instar$. The linear trend appears to be quite consistent across the different stages of a caterpillar's life.



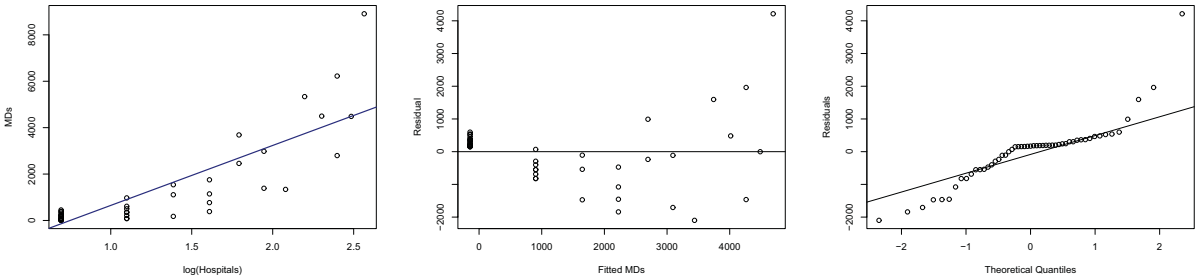
1.50 Here are three plots (scatterplot with fitted line, residuals versus fits plot, and normal quantile plot of the residuals) for each of the combinations of log transformations.

$log(MDs)$ versus $Hospitals$:



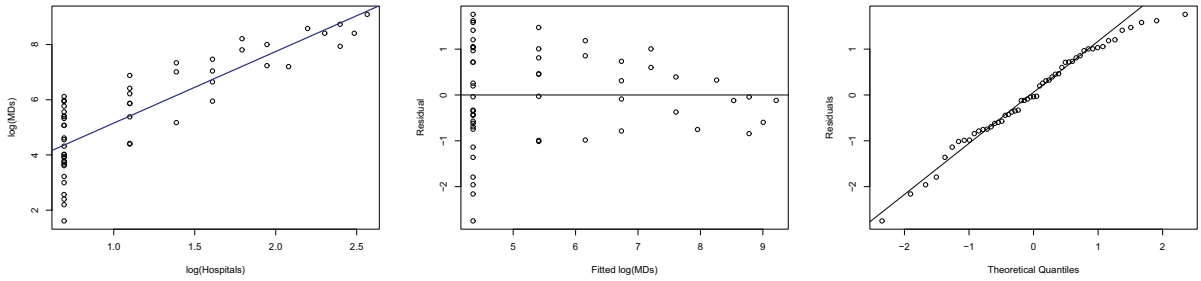
In the first two plots above, we see curvature in the opposite direction from the original data (here the scatterplot and residual versus fits plot show concave down). Also the variability in $log(MDs)$ looks larger for small numbers of hospitals and smaller for counties with more hospitals. So there are problems with both linearity and equal variance. The normality plot looks pretty good, except for some straying from the line for the few largest values.

MDs versus $log(Hospitals)$:



In the first two plots above, we see clear curvature (even more extreme than in the original scales) and concave up patterns in both the scatterplot and residual versus fits plot. In this case the variability in *MDs* gets larger as $log(Hospitals)$ increases. There are also big problems in both tails of the normal quantile plot. This reexpression appears to make the conditions look even worse than in the original scale.

$log(MDs)$ versus $log(Hospitals)$:



This is the best of these three options. The scatterplot and residuals versus fits plot show no obvious curvature, although we still see a problem with decreasing variance in both plots. The normal quantile plot is similar to the first case where we only transformed the response ($log(MDs)$).

Although the $log(MDs)$ versus $log(Hospitals)$ reexpressions together look like the best option among these three, the transformation with $sqrt(MDs)$ presented in the original text example is probably better, because it helps stabilize the variance as well as dealing with the curvature.

1.51 Start with any small dataset, such as the one shown below.

x	1	2	3	4	5	$\overline{x} = 3$
y	10	5	12	8	20	$\overline{y} = 11$

Pick any slope, say $\hat{\beta}_1 = -3$, and compute the intercept with

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 11 - (-3)(3) = 20$$

The following table shows the predicted values using $\hat{y} = 20 - 3x$ and the residuals from the actual y values.

$\hat{y}$	17	14	11	8	5	
$y - \hat{y}$	-7	-9	1	0	15	mean residual = 0

When the intercept is chosen as $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, the mean of the residuals will always be zero, even when the line doesn't follow the trend of the data.