CHAPTER 6

6.1 (a) The Minitab output of the three regressions is shown below.
In the model involving x_1 alone, the hypothesis $\beta_1 = 0$ can not be rejected. This indicates that x_1 by itself is not important.
Similarly, in the model involving x_2 alone, x_2 by itself is not significant ($\beta_2 = 0$ can not be rejected).
The model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ leads to a large R^2 = 0.794, and the partial t-tests for $\beta_1 = 0$ and $\beta_2 = 0$ are significant. This indicates that x_1 helps explain y at fixed levels of x_2; and x_2 helps explain y at fixed levels of x_1.
This example is instructive as it shows that regressors may be insignificant when studied alone, but taken jointly they may help explain a large part of the variability. It provides an example where stepwise procedures lead to different solutions. Forward selection and stepwise regression would not include any variables, whereas backward elimination would select the model with both regressors. This shows that it is preferable to look at all possible regressions. Note that x_1 and x_2 are correlated (r = 0.734).

```
The regression equation is
Y = 889 - 6.52 X1

Predictor        Coef     SE Coef          T        P
Constant        889.3       268.9       3.31    0.011
X1             -6.519       8.289      -0.79    0.454

S = 123.2       R-Sq = 7.2%      R-Sq(adj) = 0.0%


The regression equation is
Y = 387 + 1.55 X2

Predictor        Coef     SE Coef          T        P
Constant        387.4       287.4       1.35    0.214
X2              1.550       1.509       1.03    0.334

S = 120.2       R-Sq = 11.7%     R-Sq(adj) = 0.6%


The regression equation is
Y = 547 - 31.1 X1 + 6.00 X2

Predictor        Coef     SE Coef          T        P
Constant        547.1       152.0       3.60    0.009
X1            -31.147       6.491      -4.80    0.002
X2              6.003       1.212       4.95    0.002

S = 62.04       R-Sq = 79.4%     R-Sq(adj) = 73.5%
```

```
Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      103859       51930     13.49    0.004
Residual Error     7       26941        3849
Total              9      130800
```

(b) Observation #2 (with $x_1 = 43$, $x_2 = 223$ and y = 480) is unusual and somewhat different than the rest. We remove this observation and refit the three models. The results are similar, with the model with both x_1 and x_2 leading to the best representation.

```
The regression equation is
Y = 287 - 17.6 X1 + 5.18 X2

Predictor        Coef     SE Coef          T        P
Constant        286.8       155.1       1.85    0.114
X1            -17.557       7.323      -2.40    0.053
X2             5.1801      0.9733       5.32    0.002

S = 46.90       R-Sq = 84.7%     R-Sq(adj) = 79.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2       73159       36579     16.63    0.004
Residual Error     6       13197        2199
Total              8       86356
```

6.2

(a) Linear model: $\hat{\mu} = 23.35 + 1.045x$; R^2 = 0.955; s = 0.737;
F(lack of fit) = 10.01; p-value = 0.002; lack of fit.

Source	d.f	S.S	M.S	F	Prob$\geq$F
Model	1	195.2428	195.2428	359.3	0.0001
Error	17	9.2382	0.5434		
Lack of Fit	9	8.4849	0.9427	10.01	<0.01
Pure Error	8	0.7533	0.0942		

(b) Quadratic model: $\hat{\mu} = 22.56 + 1.67x - 0.068x^2$; R^2 = 0.988; s = 0.394;
$t(\hat{\beta}_2) = -0.06796 / 0.01031 = -6.59$; reject $\beta_2 = 0$;
F(lack-of-fit) = 2.30; p-value = 0.13; no lack of fit.

Source	d.f	S.S	M.S	F	Prob>F
Model	2	201.9944	100.9972	649.86	0.0001
Error	16	2.4866	0.1554		
Lack of Fit	8	1.7333	0.2166	2.3	>.10
Pure Error	8	0.7533	0.0947		

6.3 Vector of fitted values and residuals: $\hat{\boldsymbol{\mu}} = \mathrm{H}\boldsymbol{y}$; $\boldsymbol{e} = (\mathrm{I} - \mathrm{H})\boldsymbol{y} = (\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}')\boldsymbol{y}$,
where $\mathrm{X} = [\mathbf{1}, \boldsymbol{x}]$ is the n x 2 matrix, and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$.
True model : $\boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \boldsymbol{\varepsilon}$ where $\boldsymbol{x}_2' = (\mathrm{x}_1^2, ..., \mathrm{x}_n^2)$

$$\mathrm{E}(\boldsymbol{e}) = (\mathrm{I} - \mathrm{X}'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}')\mathrm{E}(\boldsymbol{y}) = (\mathrm{I} - \mathrm{H})[\mathrm{X}\boldsymbol{\beta} + \beta_2 \boldsymbol{x}_2 + \mathrm{E}(\boldsymbol{\varepsilon})] = (\mathrm{I} - \mathrm{H})\mathrm{X}\boldsymbol{\beta} + \beta_2 (\mathrm{I} - \mathrm{H})\boldsymbol{x}_2$$
$$= \beta_2 (\mathrm{I} - \mathrm{H})\boldsymbol{x}_2 \quad \text{since} \quad (\mathrm{I} - \mathrm{H})\mathrm{X} = \mathrm{O}$$

6.4

(a) $\mathrm{E}(\hat{\boldsymbol{\mu}}) = \mathrm{E}(\mathrm{X}\hat{\boldsymbol{\beta}}) = \mathrm{X}\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathrm{X}\boldsymbol{\beta}$

$\mathrm{V}(\hat{\boldsymbol{\mu}}) = \mathrm{V}(\mathrm{X}\hat{\boldsymbol{\beta}}) = \mathrm{X}\mathrm{V}(\hat{\boldsymbol{\beta}})\mathrm{X}' = \mathrm{X}(\sigma^2 (\mathrm{X}'\mathrm{X})^{-1})\mathrm{X}' = \sigma^2 \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$

(b) $\sum_{i=1}^{n} \mathrm{V}(\hat{\mu}_i) = \sigma^2 \, \mathrm{tr}\,(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}') = \sigma^2 \, \mathrm{tr}\,((\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}) = \sigma^2 \, \mathrm{tr}\,(\mathrm{I}) = \sigma^2 (\mathrm{p}+1)$

Hence $\frac{1}{n}\sum_{i=1}^{n} \mathrm{V}(\hat{\mu}_i) = \frac{(\mathrm{p}+1)}{n}\sigma^2$

(c) $\boldsymbol{a}_i'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{a}_i = \boldsymbol{a}_i'\mathrm{H}\boldsymbol{a}_i \geq 0$ because $(\mathrm{X}'\mathrm{X})^{-1}$ is a positive semidefinite matrix.
Select $\boldsymbol{a}_i$ as the vector with all components 0 except for a “1” in the ith element.
Thus $\mathrm{h}_{ii} \geq 0$.

H is symmetric and idempotent. $\mathrm{H} = \mathrm{HH}$ implies $\mathrm{h}_{ii} = \mathrm{h}_{ii}^2 + \sum_{j \neq i}^{n} \mathrm{h}_{ij}^2 \geq 0$ and

$\sum_{j \neq i}^{n} \mathrm{h}_{ij}^2 = \mathrm{h}_{ii}(1 - \mathrm{h}_{ii}) \geq 0$. Since $\mathrm{h}_{ii} \geq 0$, we find that $(1 - \mathrm{h}_{ii}) \geq 0$ and $\mathrm{h}_{ii} \leq 1$.

(d) We can parameterize the model as $\boldsymbol{y} = \mathbf{1}\alpha + \mathrm{V}\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}$ where
$\alpha = \beta_0 + \beta_1 \bar{\mathrm{x}}_1 + ... + \beta_p \bar{\mathrm{x}}_p$, $\mathrm{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ...\boldsymbol{v}_p]$ contains the mean corrected
regressors $\boldsymbol{v}_j = \boldsymbol{x}_j - \bar{\mathrm{x}}_j \mathbf{1}$, $\bar{\mathrm{x}}_j$ is the average of the elements of the vector $\boldsymbol{x}_j$,
and $\boldsymbol{\beta}_*$ is the vector $\boldsymbol{\beta}$ without the element β_0.
Note that $\mathrm{X} = [\mathbf{1}, \mathrm{V}]$ and $\mathbf{1}'\boldsymbol{v}_j = 0$, for j = 1, 2, …, p. Hence

$$X'X = \begin{bmatrix} n & 0 \\ 0 & V'V \end{bmatrix}, \quad (X'X)^{-1} = \begin{bmatrix} n^{-1} & 0 \\ 0 & (V'V)^{-1} \end{bmatrix}, \text{ and}$$

$$H = [\mathbf{1} \quad V] \begin{bmatrix} n^{-1} & 0 \\ 0 & (V'V)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ V' \end{bmatrix} = [n^{-1}\mathbf{1}\mathbf{1}' + V(V'V)^{-1}V'].$$

The matrix $H^* = V(V'V)^{-1}V'$ is symmetric and idempotent; we have shown in 6.4(c) that its diagonal elements h^*_{ii} are between 0 and 1. Hence the ith diagonal element of H, $h_{ii} = n^{-1} + h^*_{ii} \ge n^{-1}$.

(e) Both $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are solutions of $(X'X)\boldsymbol{\beta} = X'\boldsymbol{y}$. Hence $(X'X)\hat{\boldsymbol{\beta}} = X'\boldsymbol{y}$ and $(X'X)\tilde{\boldsymbol{\beta}} = X'\boldsymbol{y}$, and $(X'X)(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = \mathbf{0}$.

Let $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\mu}} = X\tilde{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\mu}} - \tilde{\mu} = X(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$.

$$\sum_{i=1}^{n} (\hat{\mu}_i - \tilde{\mu}_i)^2 = (\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}})'(\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}) = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})' X'X(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\mathbf{0} = 0$$

The sum of squares is zero if and only if $(\hat{\mu}_i - \tilde{\mu}_i) = 0$ for all i. Hence $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}$.

6.5

(a) We need to show: $(I + \alpha\boldsymbol{v}\boldsymbol{w}')\left[I - \left(\frac{\alpha}{1 + \alpha\boldsymbol{v}'\boldsymbol{w}}\right)\boldsymbol{v}\boldsymbol{w}'\right] = I$

The left hand side is given by

$$\text{LHS} = I + \alpha\boldsymbol{v}\boldsymbol{w}' - \left[\frac{\alpha[\boldsymbol{v}\boldsymbol{w}' + \alpha\boldsymbol{v}\boldsymbol{w}'\boldsymbol{v}\boldsymbol{w}']}{1 + \alpha\boldsymbol{v}'\boldsymbol{w}}\right]$$

$$= I + \alpha\boldsymbol{v}\boldsymbol{w}' - \left[\frac{\alpha}{1 + \alpha\boldsymbol{v}'\boldsymbol{w}}\right][1 + \alpha\boldsymbol{v}'\boldsymbol{w}]\boldsymbol{v}\boldsymbol{w}' = I + \alpha\boldsymbol{v}\boldsymbol{w}' - \alpha\boldsymbol{v}\boldsymbol{w}' = I$$

(b) For full rank matrices with the same dimension: $(CD)^{-1} = D^{-1}C^{-1}$. Hence $(A + \boldsymbol{w}\boldsymbol{w}')^{-1} = [A(I + A^{-1}\boldsymbol{w}\boldsymbol{w}')]^{-1} = (I + A^{-1}\boldsymbol{w}\boldsymbol{w}')^{-1}A^{-1}$.

Let $A^{-1}\boldsymbol{w} = \boldsymbol{v}$ and $\alpha = 1$. Then

$$(A + \boldsymbol{w}\boldsymbol{w}')^{-1} = (I + \boldsymbol{v}\boldsymbol{w}')^{-1}A^{-1} = \left[I - \left(\frac{1}{1 + \boldsymbol{v}'\boldsymbol{w}}\right)\boldsymbol{v}\boldsymbol{w}'\right]A^{-1} = A^{-1} - \frac{A^{-1}\boldsymbol{w}\boldsymbol{w}'A^{-1}}{1 + \boldsymbol{w}'A^{-1}\boldsymbol{w}}.$$

(c) (i) Note that $X_1 = \begin{bmatrix} X \\ \boldsymbol{w}' \end{bmatrix}$; $(X_1'X_1)^{-1} = (X'X + \boldsymbol{w}\boldsymbol{w}')^{-1}$

Let $X'X = A$. Then

$$(X_1'X_1)^{-1} = A^{-1} - \frac{A^{-1}\boldsymbol{ww}'A^{-1}}{1-\boldsymbol{w}'A^{-1}\boldsymbol{w}} = (X'X)^{-1} - \frac{(X'X)^{-1}\boldsymbol{ww}'(X'X)^{-1}}{1-\boldsymbol{w}'(X'X)^{-1}\boldsymbol{w}}$$

(ii) $\hat{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'\boldsymbol{y}_1 = (X'X+\boldsymbol{ww}')^{-1}(X\boldsymbol{y}+\boldsymbol{w}y_{n+1}) =$

$$= \hat{\boldsymbol{\beta}} - \frac{(X'X)^{-1}\boldsymbol{ww}'\hat{\boldsymbol{\beta}}}{1-\boldsymbol{w}'(X'X)^{-1}\boldsymbol{w}} + (X'X)^{-1}\boldsymbol{w}y_{n+1} - \frac{(X'X)^{-1}\boldsymbol{ww}'(X'X)^{-1}\boldsymbol{w}y_{n+1}}{1-\boldsymbol{w}'(X'X)^{-1}\boldsymbol{w}}$$

Define the scalar $h = \boldsymbol{w}'(X'X)^{-1}\boldsymbol{w}$. Then

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}} - \frac{(X'X)^{-1}\boldsymbol{ww}'}{1-h}\hat{\boldsymbol{\beta}} + \frac{(X'X)^{-1}\boldsymbol{w}(1-h)y_{n+1}}{1-h}$$

$$= \hat{\boldsymbol{\beta}} + (X'X)^{-1}\boldsymbol{w}(y_{n+1} - \frac{1}{1-h}\boldsymbol{w}'\hat{\boldsymbol{\beta}})$$

6.6 The estimate of $\boldsymbol{\beta}$ in the model with all the x's, $\boldsymbol{y} = X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, is

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_{(K)} \\ \hat{\beta}_K \end{bmatrix} = \begin{bmatrix} \tilde{X}'\tilde{X} & \tilde{X}\boldsymbol{x}_K \\ \boldsymbol{x}_K'\tilde{X} & \boldsymbol{x}_K'\boldsymbol{x}_K \end{bmatrix}^{-1} \begin{bmatrix} \tilde{X}'\boldsymbol{y} \\ \boldsymbol{x}_K'\boldsymbol{y} \end{bmatrix}$$

where the n x (k-1) matrix $\tilde{X}$ is as defined in the hint and where $\hat{\boldsymbol{\beta}}_{(K)}$ denotes the vector of estimates $\hat{\boldsymbol{\beta}}$ without the element $\hat{\beta}_K$.

Using the results on the inverse of a partitioned matrix given in the appendix of Chapter 6, we obtain

$\hat{\beta}_K = \frac{\boldsymbol{x}_K'(I-\tilde{H})\boldsymbol{y}}{\boldsymbol{x}_K'(I-\tilde{H})\boldsymbol{x}_K}$ where $\tilde{H} = I - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'$ is an idempotent matrix; $\tilde{H}\tilde{H} = \tilde{H}$.

In step 1, when we regress $\boldsymbol{y}$ on $\tilde{X}$ we obtain the vector of residuals $\boldsymbol{r} = (I-\tilde{H})\boldsymbol{y}$.

In step 2, when we regress $\boldsymbol{x}_K'$ on $\tilde{X}$ we obtain the vector of residuals $\boldsymbol{u} = (I-\tilde{H})\boldsymbol{x}_K'$

Note that the means of the residual vectors $\boldsymbol{r}$ and $\boldsymbol{u}$ are zero. Hence the slope of the regression of $\boldsymbol{r}$ on $\boldsymbol{u}$ in step 3 is

$$\tilde{\beta}_K = \boldsymbol{u}'\boldsymbol{r}/\boldsymbol{u}'\boldsymbol{u} = \frac{\boldsymbol{x}_K'(I-\tilde{H})(I-\tilde{H})\boldsymbol{y}}{\boldsymbol{x}_K'(I-\tilde{H})(I-\tilde{H})\boldsymbol{x}_K} = \boldsymbol{x}_K'(I-\tilde{H})\boldsymbol{y}\big/\boldsymbol{x}_K'(I-\tilde{H})\boldsymbol{x}_K = \hat{\beta}_K$$

6.7
(a) True. For a correct model, $\text{Cov}(\boldsymbol{e}, \hat{\boldsymbol{\mu}}) = \text{O}$, and a plot of the residuals e_i against the fitted values $\hat{\mu}_i$ should show no association. However, $\text{Cov}(\boldsymbol{e}, \boldsymbol{y}) = \sigma^2(\text{I} - \text{H})$; the correlation makes the interpretation of the plot of e_i against y_i difficult.
(b) Not true. Outliers should be scrutinized, but not necessarily rejected.
(c) True

6.8 (a) 5; (b) 2; (c) 4; (d) 1

6.9 (a) True; (b) True; (c) False; (d) False; (e) False

6.10 (d) True. Linear regression of $\ln(y)$ on $\ln(x_1)$ and $\ln(x_2)$ to estimate β_1 and β_2 .

6.11 (a) No; (b) No; (c) No; (d) No; (e) True

6.12 A (Palm Beach); B (Broward); C (Dade); D (Pasco)

6.13 Consider the stock price data **lenzing** and refer to Exercise 10.9

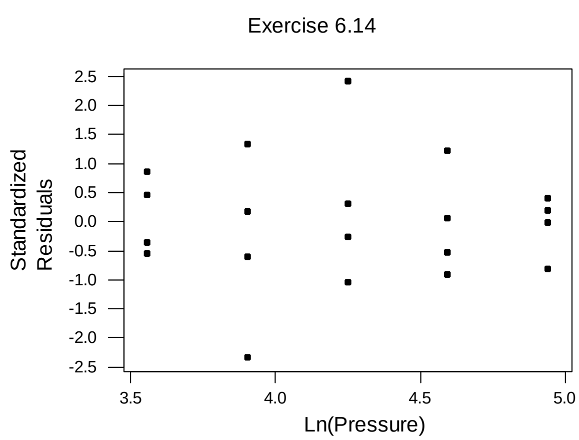
6.14 Note that the pressures are equally spaced on the logarithmic scale, suggesting that the investigator expected equal changes in the ratio of pressures to produce equal changes in the tearing factor. This suggests that a logarithmic transformation of pressure (x) may be appropriate.
Scatter plots of y against x, y against ln(x), ln(y) against x, and ln(y) against ln(x) were constructed. For a data set of such small size, the choice among the various transformations is difficult. Here we consider a model of y on ln(x).

R-output

```
              Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)    152.451       10.493    14.529    2.19e-11
lnx            -10.604        2.453    -4.322    0.000411

Residual standard error: 5.378 on 18 degrees of freedom
Multiple R-Squared: 0.5093,     Adjusted R-squared: 0.482
F-statistic: 18.68 on 1 and 18 DF,  p-value: 0.0004105
```

Because of the replications it is possible to calculate a test for lack of fit. The F-statistic is small and no lack of fit is indicated. The residual plot suggests that the variability in the response may not be the same at all settings of pressure. However, this fact is difficult to assess with a small data set such as this.

```
Minitab output and test for lack of fit
The regression equation is
Y=Tear = 152 - 10.6 LnX

Predictor        Coef     SE Coef          T        P
Constant       152.45       10.49      14.53    0.000
LnX           -10.604       2.453      -4.32    0.000

S = 5.378       R-Sq = 50.9%     R-Sq(adj) = 48.2%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      540.23      540.23     18.68    0.000
Residual Error    18      520.57       28.92
  Lack of Fit      3       28.57        9.52      0.29    0.832
  Pure Error      15      492.00       32.80
Total             19     1060.80
```



6.15 Scatter plots of y , ln(y) and 1/y against x point to a log transformation. The estimate of the transformation parameter in Box-Cox family is $\hat{\lambda} \approx 0$, indicating a logarithmic transformation of the response y.
Regression of ln(y) on x: $\hat{\mu} = 2.436 + 0.000567x$; $R^2 = 0.986$; s = 0.0845.
The first case is quite influential (x = 574; y = 21.9; Cook = 0.585).

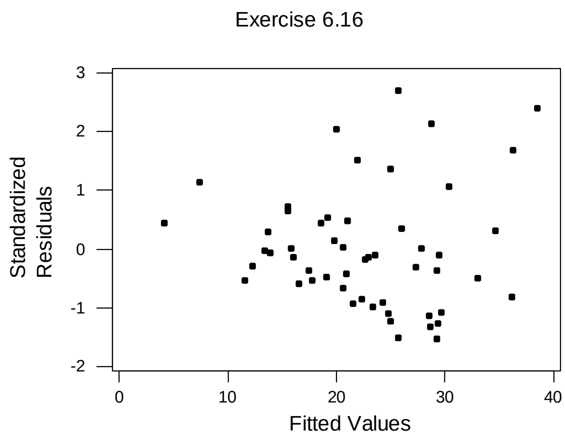
Box -Cox transformation

λ	s(λ)	R^2
-1.00	11.270	0.922
-0.75	8.569	0.948
-0.50	6.331	0.969
-0.25	4.690	0.982
-0.10	4.165	0.985
0.001 (ln)	4.082	0.986
0.10	4.232	0.985
0.25	4.849	0.980
0.50	6.629	0.965
0.75	9.033	0.942
1.00	11.960	0.912

s(λ) is the residual standard error and R^2 is the coefficient of determination in the regression of $\frac{y^{\lambda}-1}{\lambda(\bar{y}_g)^{\lambda-1}}$ on x.

6.16 The regression shows that neither of the two variables can be omitted from the model. The residual plot indicates no major model violations. Also the scatter plots of the residuals against the two explanatory variables are unremarkable. The case with the largest Cook's distance is case # 48 with $x_1 = 2.35$, $x_2 = 56$ and $y = 72$ (Cook = 0.27)

```
The regression equation is
Y = 23.0 + 23.6 X1 - 0.715 X2

Predictor        Coef     SE Coef          T        P
Constant        23.01       18.28       1.26    0.214
X1             23.639       6.848       3.45    0.001
X2            -0.7147      0.3014      -2.37    0.022

S = 14.84       R-Sq = 20.2%     R-Sq(adj) = 17.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      2783.2      1391.6      6.32    0.004
Residual Error    50     11007.9       220.2
Total             52     13791.2
```



6.17 Scatter plots indicate that a linear regression of rigidity on elasticity and density is appropriate. Partial output from R is given below:

```
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  -1.8300    121.1577   -0.015      0.988
x1            3.4179      0.7925    4.313   8.21e-05
x2           19.5830      3.2851    5.961   3.08e-07

Residual standard error: 185.9 on 47 degrees of freedom
Multiple R-Squared: 0.8119, Adjusted R-squared: 0.8039
F-statistic: 101.4 on 2 and 47 DF, p-value: < 2.2e-16
```

Residual diagnostics indicate that observation # 40 has large influence (Cook = 0.572). This observation should be scrutinized.
We remove this observation and refit the model on the reduced data set. The Minitab results are shown below. The residual plot is unremarkable, except perhaps for a large positive and a large negative residual. However, the Cook influence from the case with the large positive residual (original case # 46) is not particularly worrisome (Cook = 0.215).

```
The regression equation is
Y = - 9.2 + 4.21 X1 + 15.9 X2

Predictor        Coef     SE Coef          T        P
Constant        -9.17       94.51      -0.10    0.923
X1             4.2146      0.6344       6.64    0.000
X2             15.949       2.644       6.03    0.000

S = 145.0       R-Sq = 87.6%     R-Sq(adj) = 87.1%
```

```
Analysis of Variance

Source            DF        SS        MS        F        P
Regression         2   6843941   3421971   162.76    0.000
Residual Error    46    967129     21025
Total             48   7811070
```

6.18

(a) The correlation between liver weight (LW) and body weight (BW) is 0.5. This is also confirmed by the plot of LW versus BW.

(b) Pair-wise scatter plots of y against the three regressors show very little association. We regress y (dose in liver) on BW = body weight, LW = liver weight and DL = dose. The regression results indicate that BW and DL are significant, which is somewhat surprising as we have not seen strong associations in the pair-wise scatter plots.

Case # 3 (with BW = 190, LW = 9.0, Dose = 1.00, and y = 0.56) is a very influential observation (Cook = 0.930). This case should be scrutinized. Dropping this case from the data set, leads to the regression results shown below. Neither one of the three regressors is significant (F-statistic = 0.10), which supports the conclusion from the earlier scatter plots.

R output (all observations)

```
Coefficients:
              Estimate  Std. Error   t value    Pr(>|t|)
(Intercept)  0.265922    0.194585     1.367      0.1919
BW          -0.021246    0.007974    -2.664      0.0177
LW           0.014298    0.017217     0.830      0.4193
D            4.178111    1.522625     2.744      0.0151

Residual standard error: 0.07729 on 15 degrees of freedom
```

```
Multiple R-Squared: 0.3639,  Adjusted R-squared: 0.2367
F-statistic:  2.86 on 3 and 15 DF, p-value: 0.07197
```

Minitab output (case # 3 removed)

```
The regression equation is
Y = 0.311 - 0.0078 BW + 0.0090 LW + 1.48 Dose

Predictor        Coef     SE Coef          T        P
Constant       0.3114      0.2051       1.52    0.151
BW           -0.00778     0.01872      -0.42    0.684
LW            0.00899     0.01866       0.48    0.637
Dose            1.485       3.713       0.40    0.695

S = 0.07825     R-Sq = 2.1%      R-Sq(adj) = 0.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3    0.001844    0.000615      0.10    0.958
Residual Error    14    0.085717    0.006123
Total             17    0.087561
```

6.19

Pair-wise scatter plots of y against the two regressors show moderate association and an outlying case (case #17 with x_1 = 26.8, x_2 = 58 and y =168). The regression results shown below indicate a significant regressor x_1 and R^2 = 0.482. The influence of case #17 is large (Cook = 0.838). Removing this case from the data set leads to the revised estimates. Variable x_2 can be dropped from the model. Inorganic phosphorus explains about half of the variation in plant phosphorus (R^2 = 0.519).

Minitab output

```
The regression equation is
Y = 56.3 + 1.79 X1 + 0.087 X2

Predictor        Coef     SE Coef          T        P
Constant        56.25       16.31       3.45    0.004
X1             1.7898      0.5567       3.21    0.006
X2             0.0866      0.4149       0.21    0.837

S = 20.68       R-Sq = 48.2%     R-Sq(adj) = 41.3%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      5975.7      2987.8      6.99    0.007
Residual Error    15      6413.9       427.6
Total             17     12389.6
```

Minitab output (case #17 omitted)

```
The regression equation is
Y = 66.5 + 1.29 X1 - 0.111 X2

Predictor        Coef     SE Coef          T        P
Constant       66.465       9.850       6.75    0.000
X1             1.2902      0.3428       3.76    0.002
X2            -0.1110      0.2486      -0.45    0.662

S = 12.25       R-Sq = 52.5%     R-Sq(adj) = 45.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      2325.2      1162.6      7.75    0.005
Residual Error    14      2101.3       150.1
Total             16      4426.5
```

Minitab output (x1 only; case #17 omitted)

```
The regression equation is
Y = 62.6 + 1.23 X1

Predictor        Coef     SE Coef          T        P
Constant       62.569       4.452      14.05    0.000
X1             1.2291      0.3058       4.02    0.001

S = 11.92       R-Sq = 51.9%     R-Sq(adj) = 48.6%
Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      2295.2      2295.2     16.15    0.001
Residual Error    15      2131.2       142.1
Total             16      4426.5
```

6.20

The scatter plot of vocabulary (y) against age (x) indicates an approximate linear relationship, with the exception of case #1 (Age = 1; Vocabulary = 3). Fitting the linear regression on age leads to the results shown below. The first case exerts large influence (Cook = 1.126). Omitting this observation leads to the revised estimates. The fit improves; the standard deviation of the residuals decreases from 116.7 to 81.45. Also the residual plots improve.

R output (all observations)

```
Coefficients:
               Estimate  Std. Error   t value    Pr(>|t|)
(Intercept)     -763.86       88.25    -8.656    2.47e-05
Age              561.93       24.29    23.134    1.29e-08

Residual standard error: 116.7 on 8 degrees of freedom
Multiple R-Squared: 0.9853,     Adjusted R-squared: 0.9834
F-statistic: 535.2 on 1 and 8 DF,  p-value: 1.294e-08
```

```
R output (after dropping case #1)
Coefficients:
               Estimate  Std. Error   t value    Pr(>|t|)
(Intercept)     -894.75       74.88    -11.95    6.54e-06
Age              592.34       19.63     30.18    1.13e-08

Residual standard error: 81.45 on 7 degrees of freedom
Multiple R-Squared: 0.9924,     Adjusted R-squared: 0.9913
F-statistic: 910.7 on 1 and 7 DF,  p-value: 1.131e-08
```

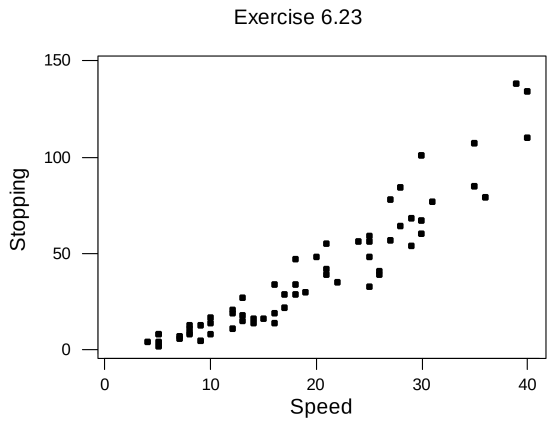
6.21
Scatter plot of ln(y) against ln(x) shows a linear association with three outlying observations (brachiosaurus, diplodocus, and triceratops). Omitting these three cases and fitting the linear model to the reduced data set leads to an adequate fit. Estimated equation: $\hat{\mu} = 2.15 + 0.752\ln(x)$; $R^2 = 0.922$; s = 0.726. The two observations with the largest positive residuals and the largest Cook influence are human (stand. residual = 2.72; Cook = 0.174) and Rhesus monkey (stand. residual = 2.25; Cook =0.119).

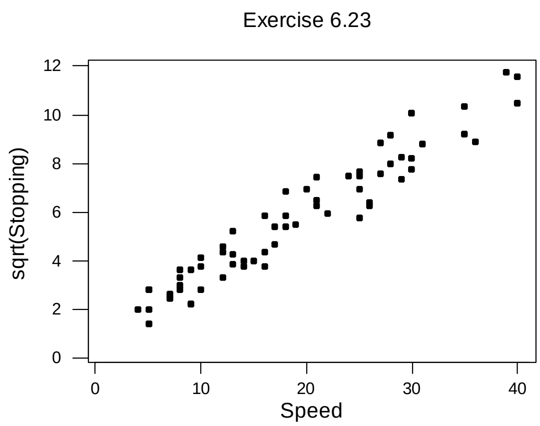
6.22
Estimated equation: $\hat{\mu} = 74.319 - 2.089\text{Conc} + 0.430\text{Ratio} - 0.372\text{Temp}$;
$R^2 = 0.939$; s = 0.74; F(lack of fit) = 7.44; p-value = 0.036; indication of lack of fit.

```
Analysis of Variance

Source            DF          SS          MS         F         P
Regression         3      92.304      30.768     56.17     0.000
Residual Error    11       6.026       0.548
  Lack of Fit      7       5.596       0.799      7.44     0.036
  Pure Error       4       0.430       0.108
Total             14      98.329
```

Run #2 (Conc = 1, Ratio = -1,Temp = -1; Yield = 73.9) influential, with large Cook's distance. This run should be investigated. Without this run, no lack of fit.

6.23 Scatter plots of y, ln(y), $\sqrt{y}$, 1/y against x indicate that the square root transformation works best to (i) achieve a linear relationship, and (ii) stabilize the variance.

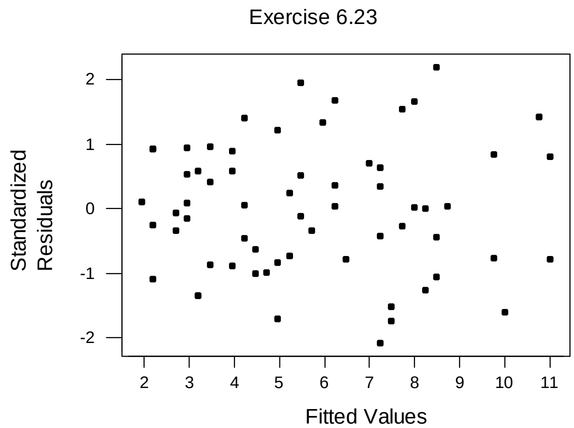


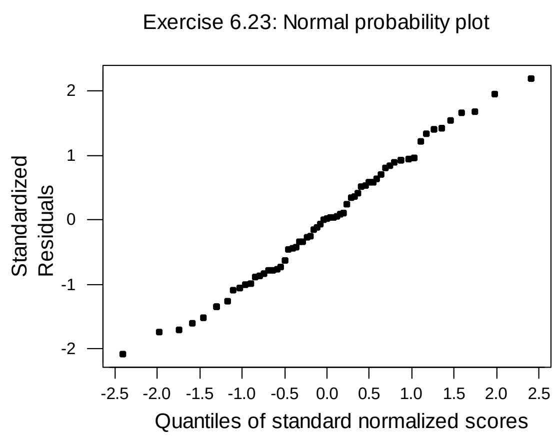


The regression results for the square root transformation of the response are shown below. The residual plot shows no remaining patterns. The normal probability plot of the residuals is adequate.

```
The regression equation is
sqrt(Stopping) = 0.918 + 0.253 Speed

Predictor         Coef     SE Coef         T        P
Constant        0.9183      0.1974      4.65    0.000
Speed         0.252568    0.009246     27.32    0.000

S = 0.7193      R-Sq = 92.4%     R-Sq(adj) = 92.3%
```

```
Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      386.06      386.06    746.22    0.000
Residual Error    61       31.56        0.52
Total             62      417.62
```



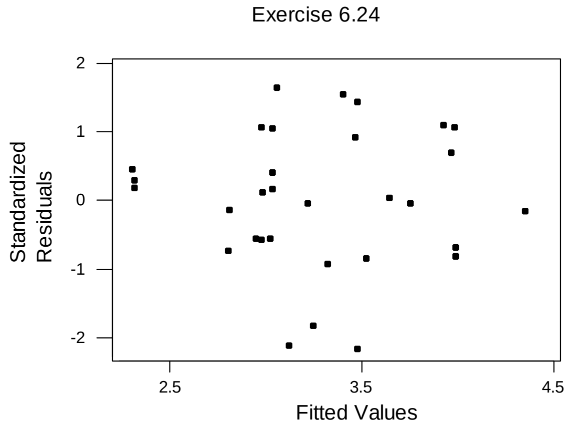


The transformation parameter of the Box-Cox family is estimated by regressing the transformed response $\frac{y^{\lambda}-1}{\lambda(\bar{y}_g)^{\lambda-1}}$ on x, and finding the λ that minimizes the error sum of squares or the residual standard error $s(\lambda)$. The results show that the square root transformation is the appropriate transformation to use.

λ	$s(\lambda)$
-1.00	40.90
-0.75	27.11
-0.50	18.49
-0.25	12.99
0.00 ln	9.49
0.25	7.61
0.50 sqrt	7.34
0.75	8.77
1.00	11.80

6.24 From the equation for the volume of a cylinder, one can expect a model of the form $V = \alpha(x_1)^2 x_2$, or after taking the logarithm, $\ln(V) = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2)$. The fit of this model is quite good; R^2 = 0.626. The residual plot is adequate, and even the largest Cook's influence (0.224 for case #18) is not particularly worrisome.

```
The regression equation is
lny = - 6.63 + 1.98 lnx1 + 1.12 lnx2

Predictor        Coef     SE Coef         T        P
Constant      -6.6316      0.7998     -8.29    0.000
lnx1          1.98265     0.07501     26.43    0.000
lnx2           1.1171      0.2044      5.46    0.000

S = 0.08139     R-Sq = 97.8%     R-Sq(adj) = 97.6%

Analysis of Variance

Source            DF          SS          MS        F        P
Regression         2      8.1232      4.0616   613.19    0.000
Residual Error    28      0.1855      0.0066
Total             30      8.3087
```



6.25 The linear model is capable of approximating the relationship; $R^2 = 0.626$. Cases #6 and #10 have the largest influence on the results (Cook = 0.327 and 0.414). Models that include the squares and the product of x1 and x2 (which could be expected from the formula for the volume of an ellipsoid) do not fare better.

```
The regression equation is
Volume = - 8.63 + 1.90 Diameter + 5.45 CrossSection

Predictor        Coef     SE Coef          T        P
Constant       -8.634       3.694      -2.34    0.044
Diameter       1.9037      0.6867       2.77    0.022
CrossSec        5.446       1.624       3.35    0.008

S = 0.07831     R-Sq = 62.6%     R-Sq(adj) = 54.3%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2    0.092505    0.046253      7.54    0.012
Residual Error     9    0.055187    0.006132
Total             11    0.147692
```

6.26

Linear model: $\hat{\mu} = 0.131 + 0.241x$, with $R^2 = 0.874$, is not appropriate.

Quadratic model: $\hat{\mu} = -1.16 + 0.723x - 0.0381x^2$, with $R^2 = 0.968$, is a possibility.

90% confidence interval: (1.972, 2.102).

Reciprocal transformation on x: $\hat{\mu} = 2.98 - 6.93(1/x)$, with $R^2 = 0.980$, is better.

90% confidence interval: (1.951, 2.026).