Solutions Manual to Accompany

INTRODUCTION to LINEAR REGRESSION ANALYSIS

Fifth Edition

DOUGLAS C. MONTGOMERY
ELIZABETH A. PECK
G. GEOFFREY VINING
Prepared by ANNE G. RYAN

WILEY

Solutions Manual to Accompany

Introduction to Linear Regression Analysis

Fifth Edition

Solutions Manual to Accompany

Introduction to Linear Regression Analysis

Fifth Edition

Douglas C. Montgomery
Arizona State University
School of Computing, Informatics, and Decisions Systems Engineering
Tempe, AZ

Elizabeth A. Peck
The Coca-Cola Company (retired)
Atlanta, GA

G. Geoffrey Vining
Virginia Tech
Department of Statistics
Blacksburg, VA

Prepared by

Anne G. Ryan
Virginia Tech
Department of Statistics
Blacksburg, VA

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey. All rights reserved.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data is available.

ISBN 978-1-118-47146-3

10 9 8 7 6 5 4 3 2 1

PREFACE

This book contains the complete solutions to the first eight chapters and the odd-numbered problems for chapters nine through fifteen in *Introduction to Linear Regression Analysis, Fifth Edition.* The solutions were obtained using Minitab® , JMP® , and SAS®.

The purpose of the solutions manual is to provide students with a reference to check their answers and to show the complete solution. Students are advised to try to work out the problems on their own before appealing to the solutions manual.

Anne G. Ryan
Virginia Tech

Dana C. Krueger
Arizona State University

Scott M. Kowalski
Minitab, Inc.

Chapter 2: Simple Linear Regression

2.1 a. $\widehat{y} = 21.8 - .007x_8$

b.

Source	d.f.	SS	MS
Regression	1	178.09	178.09
Error	26	148.87	5.73
Total	27	326.96	

c. A 95% confidence interval for the slope parameter is $-0.007025 \pm 2.056(0.00126) = (-0.0096, -0.0044)$.

d. $R^2 = 54.5\%$

e. A 95% confidence interval on the mean number of games won if opponents' yards rushing is limited to 2000 yards is $7.738 \pm 2.056(.473) = (6.766, 8.711)$.

2.2 The fitted value is 9.14 and a 90% prediction interval on the number of games won if opponents' yards rushing is limited to 1800 yards is $(4.935, 13.351)$.

2.3 a. $\widehat{y} = 607 - 21.4x_4$

b.

Source	d.f.	SS	MS
Regression	1	10579	10579
Error	27	4103	152
Total	28	14682	

c. A 99% confidence interval for the slope parameter is $-21.402 \pm 2.771(2.565) = (-28.51, -14.29)$.

d. $R^2 = 72.1\%$

e. A 95% confidence interval on the mean heat flux when the radial deflection is 16.5 milliradians is $253.96 \pm 2.145(2.35) = (249.15, 258.78)$.

2.4 a. $\hat{y} = 33.7 - .047x_1$

b.

Source	d.f.	SS	MS
Regression	1	955.34	955.34
Error	30	282.20	9.41
Total	31	1237.54	

c. $R^2 = 77.2\%$

d. A 95% confidence interval on the mean gasoline mileage if the engine displacement is 275 in^3 is $20.685 \pm 2.042(.544) = (19.573, 21.796)$.

e. A 95% prediction interval on the mean gasoline mileage if the engine displacement is 275 in^3 is $20.685 \pm 2.042(3.116) = (14.322, 27.048)$.

f. Part d. is an interval estimator on the mean response at 275 in^3 while part e. is an interval estimator on a future observation at 275 in^3. The prediction interval is wider than the confidence interval on the mean because it depends on the error from the fitted model and the future observation.

2.5 a. $\hat{y} = 40.9 - .00575x_{10}$

b.

Source	d.f.	SS	MS
Regression	1	921.53	921.53
Error	30	316.02	10.53
Total	31	1237.54	

c. $R^2 = 74.5\%$

The two variables seem to fit about the same. It does not appear that x_1 is a better regressor than x_{10}.

2.6 a. $\hat{y} = 13.3 - 3.32x_1$

b.

Source	d.f.	SS	MS
Regression	1	636.16	636.16
Error	22	192.89	8.77
Total	23	829.05	

c. $R^2 = 76.7\%$

d. A 95% confidence interval on the slope parameter is $3.3244 \pm 2.074(.3903) = (2.51, 4.13)$.

e. A 95% confidence interval on the mean selling price of a house for which the current taxes are \$750 is $15.813 \pm 2.074(2.288) = (11.07, 20.56)$.

2.7 a. $\hat{y} = 77.9 - 11.8x$

b. $t = \frac{11.8}{3.485} = 3.39$ with $p = 0.003$. The null hypothesis is rejected and we conclude there is a linear relationship between percent purity and percent of hydrocarbons.

c. $R^2 = 38.9\%$

d. A 95% confidence interval on the slope parameter is $11.801 \pm 2.101(3.485) = (4.48, 19.12)$.

e. A 95% confidence interval on the mean purity when the hydrocarbon percentage is 1.00 is $89.664 \pm 2.101(1.025) = (87.51, 91.82)$.

2.8 a. $r = +\sqrt{R^2} = .624$

b. This is the same as the test statistic for testing $\beta_1 = 0$, $t = 3.39$ with $p = 0.003$.

c. A 95% confidence interval for ρ is

$$\begin{aligned}(\tanh[\operatorname{arctanh}(.624) - 1.96/\sqrt{17}], \tanh[\operatorname{arctanh}(.624) + 1.96/\sqrt{17}]) &= \tanh(.267, 1.21) \\ &= (.261, .837)\end{aligned}$$

2.9 The no-intercept model is $\hat{y} = 2.414$ with MSE = 21.029. The MSE for the model containing the intercept is 17.484. Also, the test of $\beta_0 = 0$ is significant. Therefore, the model should not be forced through the origin.

2.10 a. $\hat{y} = 69.104 + .419x$

b. $r = .773$

c. $t = 5.979$ with $p = 0.000$, reject H_0 and claim there is evidence that the correlation is different from zero.

d. The test is

$$\begin{aligned}Z_0 &= [\operatorname{arctanh}(.773) - \operatorname{arctanh}(.6)]\sqrt{26 - 3} \\ &= (1.0277 - .6932)\sqrt{23} \\ &= 1.60.\end{aligned}$$

Since the rejection region is $|Z_0| > Z_{\alpha/2} = 1.96$, we fail to reject H_0.

e. A 95% confidence interval for ρ is

$$\tanh(1.0277 - (1.96)/\sqrt{23}) \leq \rho \leq \tanh(1.0277 + (1.96)/\sqrt{23}) = (.55, .89)$$

2.11 $\widehat{y} = .792x$ with MSE = 158.707. The model with the intercept has MSE = 75.357 and the test on β_0 is significant. The model with the intercept is superior.

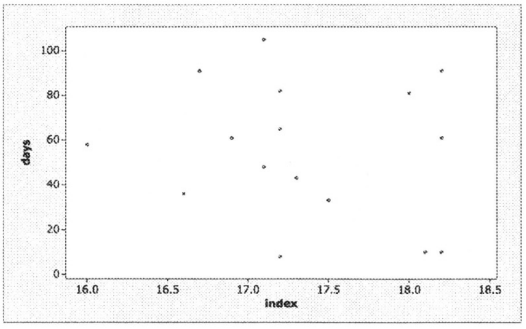
2.12 a. $\widehat{y} = -6.33 + 9.21x$

b. $F = 280590/4 = 74,122.73$, it is significant.

c. $H_0 : \beta_1 = 10000$ vs $H_1 : \beta_1 \neq 10000$ gives $t = (9.208 - 10)/.03382 = -23.4$ with $p = 0.000$. Reject H_0 and claim that the usage increase is less than 10,000.

d. A 99% prediction interval on steam usage in a month with average ambient temperature of 58° is $527.759 \pm 3.169(2.063) = (521.22, 534.29)$.

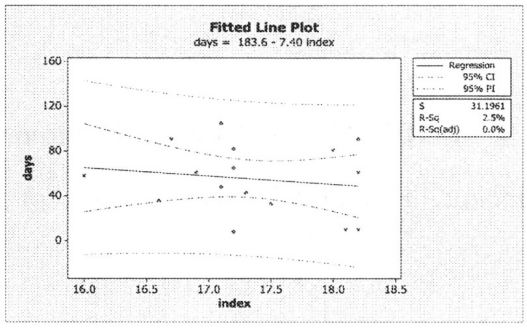
2.13 a.



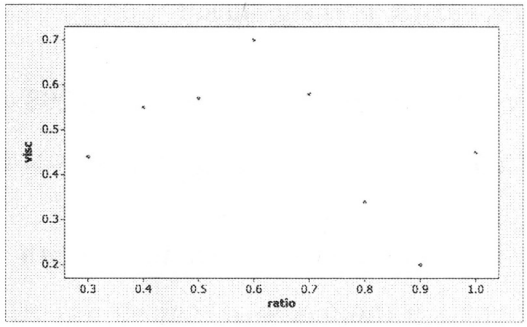
b. $\widehat{y} = 183.596 - 7.404x$

c. $F = 349.688/973.196 = .359$ with $p = 0.558$. The data suggests no linear association.

d.



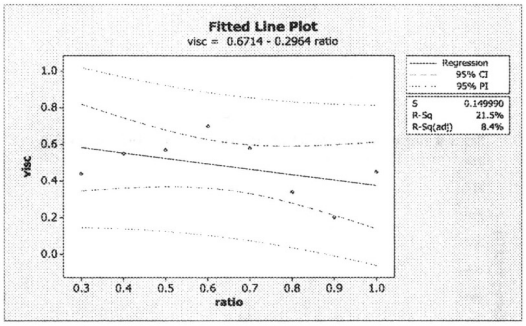
2.14 a.



b. $\widehat{y} = .671 - .296x$

c. $F = .0369/.0225 = 1.64$ with $p = 0.248$. $R^2 = 21.5\%$. A linear association is not present.

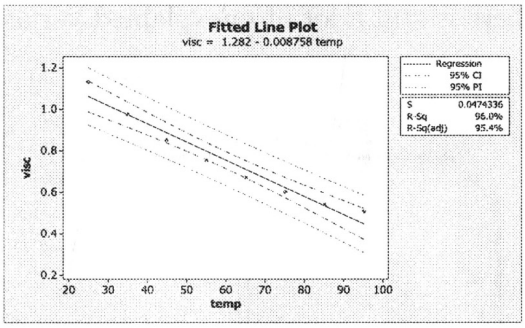
d.



2.15 a. $\widehat{y} = 1.28 - .00876x$

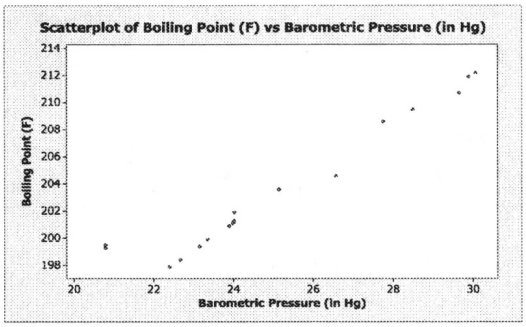
b. $F = .32529..00225 = 144.58$ with $p = 0.000$. $R^2 = 96\%$. There is a linear association between viscosity and temperature.

c.



2.16 $\widehat{y} = -290.707 + 2.346x$, $F = 34286009$ with $p = 0.000$, $R^2 = 100\%$. There is almost a perfect linear fit of the data.

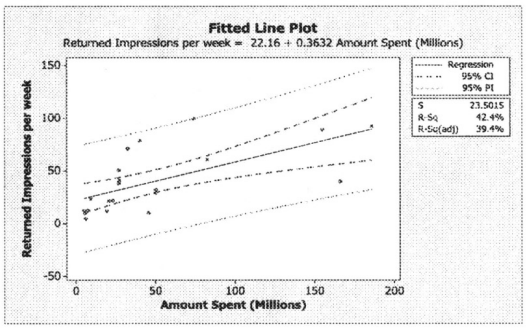
2.17 $\hat{y} = 163.931 + 1.5796x$, $F = 226.4$ with $p = 0.000$, $R^2 = 93.8\%$. The model is a good fit of the data.



2.18 a. $\hat{y} = 22.163 + 0.36317x$

b. $F = 13.98$ with $p = 0.001$, so the relationship is statistically significant. However, the $R^2 = 42.4\%$, so there is still a lot of unexplained variation in this model.

c.



d. A 95% confidence interval on returned impressions for MCI (x=26.9) is

$31.93 \pm (2.093)\sqrt{(552.3)(\frac{1}{21} + \frac{(26.9-50.4)^2}{111899})} = (20.654, 43.206)$.

A 95% prediction interval is

$31.93 \pm (2.093)\sqrt{(552.32)(1 + \frac{1}{21} + \frac{(26.9-50.4)^2}{111899})} = (-18.535, 82.395)$.

2.19 a. $\widehat{y} = 130.2 - 1.249x$, $F = 72.09$ with $p = 0.000$,$R^2 = 75.8\%$. The model is a good fit of the data.

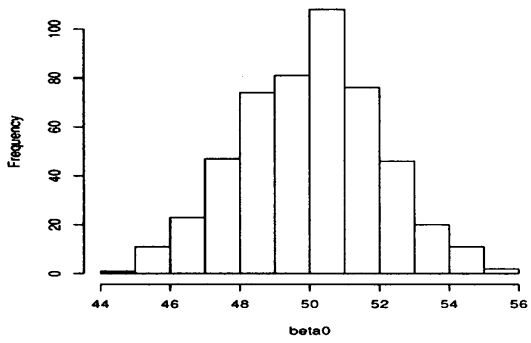
b. The fit for the SLR model relating satisfaction to age is much better compared to the fit for the SLR model relating satisfaction to severity in terms of R^2. For the SLR with satisfaction and age $R^2 = 75.8\%$ compared to $R^2 = 42.7\%$ for the model relating satisfaction and severity.

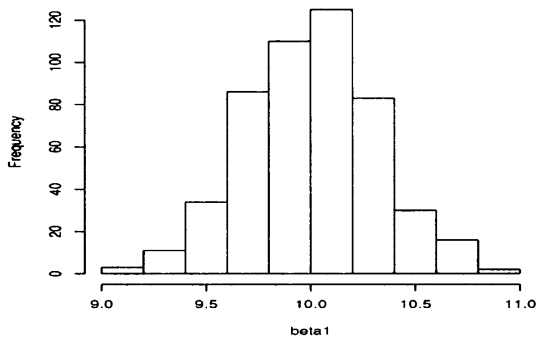
2.20 $\widehat{y} = 410.7 - 0.2638x$, $F = 7.51$ with $p = 0.016$,$R^2 = 34.9\%$. The engineer is correct that there is a relationship between initial boiling point of the fuel and fuel consumption. However, the $R^2 = 34.9\%$ indicating there is still a lot of unexplained variation in this model.

2.21 $\widehat{y} = 16.56 - 0.01276x$, $F = 4.94$ with $p = 0.034$,$R^2 = 14.1\%$. The winemaker is correct that sulfur content has a significant negative impact on taste with a $p - value = 0.034$. However, the $R^2 = 14.1\%$ indicating there is still a lot of unexplained variation in this model.

2.22 $\widehat{y} = 21.25 + 7.80x$, $F = 0.22$ with $p = 0.648$,$R^2 = 1.3\%$. The chemist's belief is incorrect. There is no relationship between the ratio of inlet oxygen to inlet methanol and percent conversion ($p - value = 0.648$). The $R^2 = 1.3\%$, which indicates that the ratio explains virtually none of the percent conversion.

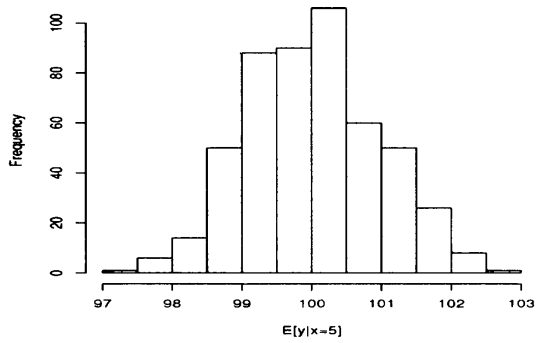
2.23 a.





Both histograms are bell-shaped. The one for β_0 is centered around 50 and the one for β_1 is centered around 10.

b. The histogram is bell-shaped with a center of 100.



c. 481 out of 500 which is 96.2% which is very close to the stated 95%.

d. 474 out of 500 which is 94.8% which is very close to the stated 95%.

2.24 Using a smaller value of n makes the estimates of the coefficients in the regression model less precise. It also increases the variability in the predicted value of y at $x = 5$. The lengths of the confidence intervals are wider for $n = 10$ and the histograms are more spread out.

2.25 a.

$$\begin{aligned} Cov(\widehat{\beta}_0, \widehat{\beta}_1) &= Cov(\bar{y} - \widehat{\beta}_1\bar{x}, \widehat{\beta}_1) \\ &= Cov(\bar{y}, \widehat{\beta}_1) - \bar{x}Cov(\widehat{\beta}_1, \widehat{\beta}_1) \\ &= 0 - \bar{x}\frac{\sigma^2}{S_{XX}} \qquad \text{(by part b)} \\ &= \frac{-\bar{x}\sigma^2}{S_{XX}} \end{aligned}$$

b.

$$\begin{aligned} Cov(\bar{y}, \widehat{\beta}_1) &= \frac{1}{nS_{XX}} Cov(\textstyle\sum y_i, \sum(x_i - \bar{x})y_i) \\ &= \frac{1}{nS_{XX}} \textstyle\sum(x_i - \bar{x})Cov(y_i, y_i) \\ &= \frac{\sigma^2}{nS_{XX}} \textstyle\sum(x_i - \bar{x}) \\ &= 0 \end{aligned}$$

2.26 a. Use the fact that $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-2}$. Then

$$\begin{aligned} E(\text{MSE}) &= E\left(\frac{\text{SSE}}{n-2}\right) \\ &= \frac{\sigma^2}{n-2} E\left(\chi^2_{n-2}\right) \\ &= \sigma^2 \end{aligned}$$

b. Use $\text{SSR} = \widehat{\beta}_1 S_{xy} = \widehat{\beta}_1^2 S_{xx}$.

$$\begin{aligned} E(\text{SSR}) &= S_{xx} E(\widehat{\beta}_1^2) \\ &= S_{xx}\left[Var(\widehat{\beta}_1 + (E(\widehat{\beta}_1))^2\right] \\ &= S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) \\ &= \sigma^2 + \beta_1^2 S_{xx} \end{aligned}$$

2.27 a. No,

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\frac{\sum(x_i - \bar{x})y_i}{S_{xx}}\right) \\ &= \frac{\sum(x_{i1} - \bar{x})}{S_{xx}} E(y_i) \\ &= \frac{\sum(x_{i1} - \bar{x})}{S_{xx}} (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) \\ &= \beta_1 + \frac{\sum(x_{i1} - \bar{x})x_{i2}}{S_{xx}} \end{aligned}$$

b. The bias is

$$\beta_1 - E(\hat{\beta}_1) = \frac{-\sum(x_{i1} - \bar{x})x_{i2}}{S_{xx}}$$

2.28 a. $\tilde{\sigma}^2 = \text{SSE}/n$. So, $E(\tilde{\sigma}^2) = \frac{n-2}{n}\sigma^2$ so the bias is $\left(1 - \frac{n-2}{n}\right)\sigma^2$.

b. As n gets large, the bias goes to zero.

2.29 If n is even, then half the points should be at $x = -1$ and the other half at $x = 1$. If n is odd, then one point should be at $x = 0$, then the rest of the points are evenly split between $x = -1$ and $x = 1$. There would be no way to test the adequacy of the model.

2.30 a. $r = +\sqrt{R^2} = 1.00$

b. The test of $\rho = 0$ is equivalent to the test of $\beta_1 = 0$. Therefore, $t = 272.25$ with $p = 0.000$.

c. For $H_0 : \rho = .5$, we get

$$\begin{aligned} Z_0 &= [\text{arctanh}(.99) - \text{arctanh}(.5)]\sqrt{9} \\ &= [2.647 - .549](3) \\ &= 6.29. \end{aligned}$$

We reject H_0.

d. $\left(\tanh[\text{arctanh}(.99) - 1.96/\sqrt{9}], \tanh[\text{arctanh}(.99) + 1.96/\sqrt{9}]\right) = (.963, .997)$

2.31 Since $R^2 = SS_R/S_{yy}$ and $S_{yy} = SS_R + SS_E$, then we need to show that in this case $SS_E > 0$. Now $SS_E = \sum(y_i - \widehat{y}_i)^2$, so for two different y_i's (say y_{1i} and y_{2i}) at the same value of x_i, both y_{1i} and y_{2i} cannot equal $\widehat{y}_i$ at x_i. Therefore at least one of $(y_{1i} - \widehat{y}_i)^2$ and $(y_{2i} - \widehat{y}_i)^2$ is > 0. Hence, $SS_E > 0$ and thus $R^2 < 1$.

2.32 a. $S(\beta_0, \beta_1) = \sum(y_i - \beta_0 - \beta_1 x_i)^2$ with β_0 known. We need to take the derivative of this with respect to β_1 and set it equal to zero. This gives

$$\begin{aligned} -2\sum_{i=1}^{n}(y_i - \beta_0 - \widehat{\beta}_1 x_i)x_i &= 0 \\ \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n}(y_i - \beta_0)x_i \\ \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(y_i - \beta_0)x_i}{\sum_{i=1}^{n} x_i^2} \end{aligned}$$

b.

$$\begin{aligned} Var(\widehat{\beta}_1) &= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} Var(\sum_{i=1}^{n} y_i x_i) \\ &= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} (\sum_{i=1}^{n} x_i^2)\sigma^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \end{aligned}$$

c. $\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{MS_E/\sum x_i^2}} \sim t_{n-2}$ so we get $\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{MS_E/\sum x_i^2}$ which is narrower than when both are unknown.

2.33

$$\begin{aligned} Var(e_i) &= Var(y_i - \widehat{y}_i) \\ &= Var(y_i) + Var(\widehat{y}_i) - 2Cov(y_i, \widehat{y}_i) \\ &= \sigma^2 + \left[\frac{\sigma^2}{n} + \frac{(x_i - \bar{x})^2\sigma^2}{S_{xx}}\right] - 2\left[\frac{\sigma^2}{n} + \frac{(x_i - \bar{x})^2\sigma^2}{S_{xx}}\right] \\ &= \sigma^2\left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right] \end{aligned}$$

which depends on the value of x_i and thus is not constant.

Chapter 3: Multiple Linear Regression

3.1 a. $\widehat{y} = -1.8 + .0036x_2 + .194x_7 - .0048x_8$

b. Regression is significant.

Source	d.f.	SS	MS	F	p-value
Regression	3	257.094	85.698	29.44	0.000
Error	24	69.87	2.911		
Total	27	326.964			

c. All three are significant.

Coefficient	test statistic	p-value
β_2	5.18	0.000
β_7	2.20	0.038
β_8	−3.77	0.001

d. $R^2 = 78.6\%$ and $R^2_{Adj} = 76.0\%$

e. $F_0 = (257.094 - 243.03)/2.911 = 4.84$ which is significant at $\alpha = 0.05$. The test statistic here is the square of the t-statistic in part c.

3.2 Correlation coefficient between y_i and $\widehat{y}_i$ is .887. So $(.887)^2 = .786$ which is R^2.

3.3 a. A 95% confidence interval on the slope parameter β_7 is $\widehat{\beta}_7 \pm 2.064(.08823) = (.012, .376)$

b. A 95% confidence interval on the mean number of games won by a team when $x_2 = 2300$, $x_7 = 56.0$ and $x_8 = 2100$ is

$$\widehat{y} \pm t_{\alpha/2,24}\sqrt{\widehat{\sigma}\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} = 7.216 \pm 2.064(.378)$$

$$= (6.44, 7.99)$$

3.4 a. $\widehat{y} = 17.9 + .048x_7 - .00654x_8$ with $F = 15.13$ and $p = 0.000$ which is significant.

b. $R^2 = 54.8\%$ and $R^2_{Adj} = 51.5\%$ which are much lower.

c. For β_7, a 95% confidence interval is $0.484 \pm 2.064(.1192) = (-.198, .294)$ and for the mean number of games won by a team when $x_7 = 56.0$ and $x_8 = 2100$, a 95% confidence interval is $6.926 \pm 2.064(.533) = (5.829, 8.024)$. Both lengths are greater than when x_2 was included in the model.

d. It can affect many things including the estimates and standard errors of the coefficients and the value of R^2.

3.5 a. $\widehat{y} = 32.9 - .053x_1 + .959x_6$

b. Regression is significant.

Source	d.f.	SS	MS	F	p-value
Regression	2	972.9	486.45	53.31	0.000
Error	29	264.65	9.13		
Total	31	1237.54			

c. $R^2 = 78.6\%$ and $R^2_{Adj} = 77.3\%$. For the simple linear regression with x_1, $R^2 = 77.2\%$.

d. A 95% confidence interval for the slope parameter β_1 is $-.053 \pm 2.045(.006145) = (-.0656, -.0405)$.

e. x_1 is significant while x_6 is not.

Coefficient	test statistic	p-value
β_1	-8.66	0.000
β_6	1.43	0.163

f. A 95% confidence interval on the mean gasoline mileage when $x_1 = 275$ in^3 and $x_6 = 2$ is $20.187 \pm 2.045(.643) = (18.872, 21.503)$.

g. A 95% prediction interval for a new observation on gasoline mileage when $x_1 = 275$ in^3 and $x_6 = 2$ is $20.187 \pm 2.045(3.089) = (13.887, 26.488)$

3.6 The lengths from problem 2.4 are 2.223 and 12.716, respectively. For problem 3.5, they are 2.631 and 12.634. The lengths are pretty much the same which indicates that adding x_6 does not help much.

3.7 a. $\hat{y} = 14.9 + 1.92x_1 + 7.00x_2 + .149x_3 + 2.72x_4 + 2.01x_5 - .41x_6 - 1.4x_7 - .0371x_8 + 1.56x_9$

b. $F = 9.04$ with $p = 0.000$ which is significant.

c. None of the t-tests are significant. There is a multicollinearity problem.

d. $F = \dfrac{(707.298 - 701.69)/2}{8.696} = .322$ which indicates their is no contribution of lot size and living space given that all the other regressors are in the model.

e. Yes, there is a multicollinearity problem.

3.8 a. $\hat{y} = 2.53 + .0185x_6 + 2.19x_7$

b. $F = 27.95$ with $p = 0.000$ which is significant. $R^2 = 70.0\%$ and $R^2_{Adj} = 67.5\%$.

c. Both are significant.

Coefficient	test statistic	p-value
β_6	6.74	0.000
β_7	2.25	0.034

d. For β_6, a 95% confidence interval is $.0185 \pm 2.064(.0027) = (.013, .024)$ and for β_7, a 95% confidence interval is $2.185 \pm 2.064(.9727) = (.177, 4.193)$.

e. $t = 6.62$ with $p = 0.000$ which is significant. $R^2 = 63.6\%$ and $R^2_{Adj} = 62.2\%$. These are basically the same as in part b.

f. A 95% confidence interval on the slope parameter β_6 is $.019 \pm 2.064(.0029) = (.013, .025)$. The length of this confidence interval is almost exactly the same as the one from the model including x_7.

g. As always, MS_{Res} is lower when x_6 and x_7 are in the model.

3.9 a. $\hat{y} = .00483 - .345x_1 - .00014x_4$

b. $F = 24.66$ with $p = 0.000$ which is significant.

c. $R^2 = 66.4\%$ and $R^2_{Adj} = 63.7\%$

d. x_1 is significant while x_4 is not.

Coefficient	test statistic	p-value
β_1	-5.12	0.000
β_7	$-.02$	0.986

e. It doesn't appear to be.

3.10 a. $\hat{y} = 4.00 + 2.34x_1 + .403x_2 + .273x_3 + 1.17x_4 - .684x_5$

b. $F = 16.51$ with $p = 0.000$ which is significant.

c. x_4 and x_5 appear to contribute to the model.

Coefficient	test statistic	p-value
β_1	1.35	0.187
β_2	1.77	0.086
β_3	0.82	0.418
β_4	3.84	0.001
β_5	-2.52	0.017

d. For the model in part a, $R^2 = 72.1\%$ and $R^2_{Adj} = 67.7\%$. For the model with only aroma and flavor, $R^2 = 65.9\%$ and $R^2_{Adj} = 63.9\%$. These are basically the same.

e. For the model in part a, the confidence interval is $1.1683 \pm 2.0369(.3045) = (.548, 1.789)$. For the model with only aroma and flavor, the confidence interval is $1.1702 \pm 2.0301(.2905) = (.581, 1.759)$. These two intervals are almost the same.

3.11 a. $\widehat{y} = 32.1 + .0556x_1 + .282x_2 + .125x_3 - .000x_4 - 16.1x_5$

b. $F = 29.86$ with $p = 0.000$ which is significant.

c. x_2 and x_5 appear to contribute to the model.

Coefficient	test statistic	p-value
β_1	1.86	0.093
β_2	4.90	0.001
β_3	0.31	0.763
β_4	-0.00	1.00
β_5	-11.03	0.000

d. For the model in part a, $R^2 = 93.7\%$ and $R^2_{Adj} = 90.6\%$. For the model with only temperature and particle size, $R^2 = 91.5\%$ and $R^2_{Adj} = 90.2\%$. These are basically the same.

e. For the model in part a, a 95% confidence interval is $.282 \pm 2.228(.05761) = (.154, .410)$. For the model with only aroma and flavor, a 95% confidence interval is $.282 \pm 2.16(.05883) = (.155, .409)$. These two intervals are almost the same.

3.12 a. $\widehat{y} = 11.1 + 350x_1 + .109x_2$

b. $F = 87.6$ with $p = 0.000$ which is significant.

c. Both contribute to the model.

Coefficient	test statistic	p-value
β_1	8.82	0.000
β_2	10.91	0.000

d. For the model in part a, $R^2 = 84.2\%$ and $R^2_{Adj} = 83.2\%$. For the model with only time, $R^2 = 46.8\%$ and $R^2_{Adj} = 45.2\%$. These are very different and suggest that amount of surfactant is needed in the model.

e. For the model in part a, a 95% confidence interval is $.1089 \pm 2.0345(.00998) = (.089, .129)$. For the model with only time, a 95% confidence interval is $.0977 \pm 2.0322(.01788) = (.061, .134)$. These second interval is wider.

3.13 a. $\widehat{y} = 5.89 - .498x_1 + .183x_2 + 35.4x_3 + 5.84x_4$

b. $F = 31.92$ with $p = 0.000$ which is significant.

c. x_2 and x_3 contribute to the model.

Coefficient	test statistic	p-value
β_1	1.41	0.165
β_2	10.63	0.000
β_3	3.19	0.002
β_4	2.01	.049

d. For the model in part a, $R^2 = 69.1\%$ and $R^2_{Adj} = 67.0\%$. For the model with only x_2 and x_3, $R^2 = 66.6\%$ and $R^2_{Adj} = 65.5\%$. These are basically the same.

e. For the model in part a, a 99% confidence interval is $.1827 \pm 2(.01718) = (.148, .217)$. For the model with only x_2 and x_3, a 99% confidence interval is $.1846 \pm 2(.01755) = (.149, .219)$. These intervals are basically the same.

3.14 a. $\hat{y} = .679 + 1.41x_1 - .0156x_2$

b. $F = 85.46$ with $p = 0.000$ which is significant.

c. Both contribute to the model.

Coefficient	test statistic	p-value
β_1	7.15	0.000
β_2	−10.95	0.000

d. For the model in part a, $R^2 = 82.2\%$ and $R^2_{Adj} = 81.2\%$. For the model with only temperature, $R^2 = 57.6\%$ and $R^2_{Adj} = 56.5\%$. These are very different and suggest that the ratio variable is needed in the model.

e. For the model in part a, a 99% confidence interval is $-.0156 \pm 2.7(.0014) = (-.019, -.012)$. For the model with only time, a 99% confidence interval is $-.0156 \pm 2.7(.0022) = (-.022, -.009)$. The second interval is wider.

3.15 a. $\hat{y} = 996 + 1.41x_1 - 14.8x_2 + 3.20x_3 - 0.108x_4 + 0.355x_5$

b. $F = 22.39$ with $p = 0.000$ which is significant.

c. PRECIP(x_1), EDUC(x_2), NONWHITE(x_3), and SO2(x_5) contribute to the model.

Coefficient	test statistic	p-value
β_1	2.04	0.046
β_2	−2.11	0.040
β_3	5.14	0.000
β_4	−0.80	0.427
β_5	3.90	0.000

d. $R^2 = 67.5\%$ and $R^2_{Adj} = 64.4\%$.

e. A 95% confidence interval on β_5 is $0.355 \pm (2.005)(0.09096) = (0.1726, 0.5374)$

3.16a. For LifeExp, $\widehat{y} = 70.2 - 0.0226x_1 - 0.000447x_2$.

For LifeExpMale, $\widehat{y} = 73.1 - 0.0257x_1 - 0.000479x_2$.

For LifeExpFemale, $\widehat{y} = 67.4 - 0.0199x_1 - 0.000409x_2$.

b. For LifeExp,$F = 13.46$ with $p = 0.000$ which is significant.

For LifeExpMale, $F = 12.53$ with $p = 0.000$ which is significant.

For LifeExpFemale, $F = 14.07$ with $p = 0.000$ which is significant.

c. Both predictors are significant in all three models.

Model	Coefficient	test statistic	p-value
LifeExp	β_1	-2.35	0.024
LifeExp	β_2	-2.22	0.033
LifeExpMale	β_1	-2.34	0.025
LifeExpMale	β_2	-2.07	0.046
LifeExpFemale	β_1	-2.36	0.024
LifeExpFemale	β_2	-2.31	0.027

d. For LifeExp, $R^2 = 43.5\%$, $R^2_{Adj} = 40.2\%$.

For LifeExpMale, $R^2 = 41.7\%$, $R^2_{Adj} = 38.4\%$.

For LifeExpFemale, $R^2 = 44.6\%$, $R^2_{Adj} = 41.4\%$.

e. For LifeExp, $-0.0004470 \pm (2.024)(0.0002016) = (-0.000855, -.00003896)$.

For LifeExpMale, $-0.0004785 \pm (2.024)(0.0002308) = (-0.0009456, -0.00001136)$.

For LifeExpFemale, $-0.0004086 \pm (2.024)(0.0001766) = (-0.000766, -0.00005116)$.

3.17 The multiple linear regression model that relates age, severity, and anxiety to patient satisfaction is significant with $F = 30.97$ and $p = 0.000$. It also appears that age and severity contribute significantly to the model, while anxiety is insignificant ($p = 0.417$). Compared to the simple linear regression in Section 2.7 that related only severity to patient satisfaction, the addition of age and anxiety has improved the model. The R^2 has increased from 0.43 to 0.82. The mean square error in the multiple linear regression is 95.1, considerably smaller than the MSE in the simple linear regression, which was 270.02. Compared to the multiple linear regression is Section 3.6, adding anxiety to the model does not seem to improve the model. The R^2_{Adj} decreases slightly from 0.792 to 0.789, the MSE increases from 93.7 to 95.1, and the regressor is insignificant with $p = 0.417$.

The regression equation is $\widehat{y} = 140 - 1.12x_{age} - 0.463x_{severity} + 1.21x_{anxiety}$.

Coefficient	test statistic	p-value
β_{age}	-6.11	0.000
$\beta_{severity}$	-2.53	0.019
$\beta_{anxiety}$	0.83	0.417

3.18 The multiple linear regression model for the fuel consumption data is insignificant with $F = 0.94$ and $p = 0.527$. The variance inflation factors (VIFs) indicate a severe multicollinearity problem with many VIFs much greater than 10. In addition none of the t-tests are significant. This model is not satisfactory.

The regression equation is $\widehat{y} = -315 + 0.159x_2 + 1.03x_3 - 8.6x_4 - 0.432x_5 - 0.14x_6 - 0.32x_7 - 0.52x_8$.

Coefficient	test statistic	p-value	VIF
β_2	0.17	0.871	1.901
β_3	0.36	0.729	168.467
β_4	-0.19	0.851	43.104
β_5	-0.47	0.648	60.791
β_6	-0.12	0.910	275.473
β_7	-0.10	0.924	185.707
β_8	-0.24	0.819	44.363

3.19 The multiple linear regression model for the wine quality of young red wines is significant with $F = 6.25$ and $p = 0.000$. However, x_7 the anthocyanin color and x_{10} the ionized anthocyanins (percent) are removed from the model due to linear dependencies. The anthocyanin color is equal to the wine color minus polymeric pigment color $(x_5 - x_6)$. The ionized anthocyanins is equal to $\frac{x_5 - x_6}{50}$.

The VIFs indicate an extreme problem with multicollinearity. Remedial methods will be discussed in Chapter 9. Due to multicollinearity caution is taken when making interpretations from this model.

The regression equation is $\widehat{y} = -5.2 + 6.15x_2 + 0.00455x_3 - 2.96x_4 + 6.58x_5 - 0.66x_6 - 14.5x_8 - 0.261x_9$.

Coefficient	test statistic	p-value	VIF
β_2	1.77	0.090	3.834
β_3	0.59	0.560	3.482
β_4	-1.37	0.183	543.612
β_5	2.15	0.042	444.590
β_6	-0.37	0.711	30.433
β_8	-1.87	0.074	7.356
β_9	-1.32	0.200	27.849

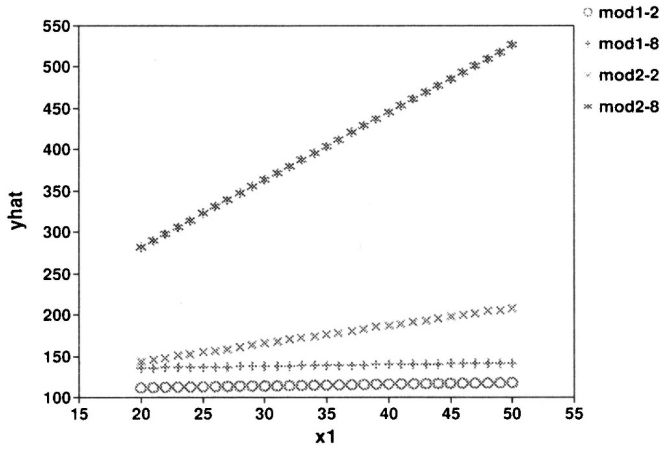
3.20 The multiple linear regression model for methanol oxidation data is significant with $F = 28.02$ and $p = 0.000$. The $R^2 = 92.1\%$ and $R^2_{Adj} = 88.8\%$. The variables x_1,

x_2 and x_3 seem to contribute to the model based on the t-tests, however there is a problem with multicollinearity as evident by the VIFs.Due to multicollinearity caution is taken when making interpretations from this model.

The regression equation is $\widehat{y} = -2669 + 22.3x_1 + 3.89x_2 + 102x_3 + 0.81x_4 - 1.63x_5$.

Coefficient	test statistic	p-value	VIF
β_1	3.09	0.009	1.519
β_2	5.70	0.000	26.284
β_3	3.91	0.002	26.447
β_4	0.21	0.840	2.202
β_5	-0.21	0.833	1.923

3.21 a. If $x_2 = 2$, then for model (1), $\widehat{y} = 108 + .2x_1$ and for model (2), $\widehat{y} = 101 + 2.15x_1$. If $x_2 = 8$, then for model (1), $\widehat{y} = 132 + .2x_1$ and for model (2), $\widehat{y} = 119 + 8.15x_1$. The interaction term in model 2 affects the slope of the line.



b. This is just the slope which is .2 regardless of the value of x_2.

c. The mean change here is $5 + .15$ which is $x_2 + .15$. Thus the result depends on the value of x_2.

3.22

$$\begin{aligned} F &= \frac{MS_R}{MS_E} \\ &= \frac{SS_R/k}{SS_E/(n-k-1)} \\ &= \frac{SS_R/(p-1)}{SS_E/(n-p)} \\ &= \frac{SS_R/(p-1)(S_{yy})}{SS_E/(n-p)(S_{yy})} \\ &= \frac{R^2(n-p)}{(p-1)(1-R^2)} \\ &= F_0 \end{aligned}$$

which then has an F distribution with $p-1$ and $n-p$ degrees of freedom.

3.23 a. $F_0 = \frac{(.9)(25-3)}{(3-1)(1-.9)} = 99$ which exceeds the critical value of $F_{.05,2,22} = 3.44$ so H_0 is rejected.

b. The value of R^2 should be surprisingly low.

$$\begin{aligned} \frac{R^2(n-p)}{(p-1)(1-R^2)} &> 3.44 \\ \frac{R^2(22)}{(2)(1-R^2)} &> 3.44 \\ \frac{R^2}{1-R^2} &> .312727 \\ R^2 &> .312727 - .312727R^2 \\ R^2 &> .238 \end{aligned}$$

3.24 $SS_R = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum y_i\right)^2}{n} = \mathbf{y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} - \frac{n^2\bar{y}^2}{n} = \sum_{i=1}^{n} \widehat{y}_i^2 - n\bar{y}^2$

3.25 a. Use $\mathbf{T}\boldsymbol{\beta} = \mathbf{c}$.

$$\mathbf{T} = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} \quad \mathbf{c} = \begin{pmatrix} 0 \\ \beta \\ \beta \\ \beta \\ \beta \end{pmatrix}$$

b. Use $\boldsymbol{\beta}$ from part a, $\mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and

$$\mathbf{T} = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

c. Use $\boldsymbol{\beta}$ from part a, $\mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and

$$\mathbf{T} = \begin{pmatrix} 0 & 1 & -2 & -4 & 0 \\ 0 & 1 & 2 & 0 & 0 \end{pmatrix}$$

3.26 a. Consider a new variable $z = \left\{ \begin{matrix} 0 & \text{if sample 1} \\ 1 & \text{if sample 2} \end{matrix} \right\}$. Then write the model as $y_i = \beta_0 + \beta_1 x_i + (\gamma_0 - \beta_0)z + (\gamma_1 - \beta_1)x_i z + \varepsilon_i$.

b. Call $\gamma_0 - \beta_0 = \nu_1$ and $\gamma_1 - \beta_1 = \nu_2$. Then we want to test $H_0 : \nu_2 = 0$. Then use

$$\mathbf{T} = \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \nu_1 \\ \nu_2 \end{pmatrix} \quad c = 0$$

c. This is test of $\nu_1 = 0$ and $\nu_2 = 0$.

$$\mathbf{T} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \nu_1 \\ \nu_2 \end{pmatrix} \quad \mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

d. Use $\beta_1 = c$ and $\nu_2 = 0$.

$$\mathbf{T} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \nu_1 \\ \nu_2 \end{pmatrix} \quad \mathbf{c} = \begin{pmatrix} c \\ 0 \end{pmatrix}$$

3.27

$$\begin{aligned} Var(\hat{\mathbf{y}}) &= Var(\mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{X}[Var(\hat{\boldsymbol{\beta}})]\mathbf{X}' \\ &= [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\,\sigma^2 \\ &= \sigma^2\mathbf{H} \end{aligned}$$

3.28

$$\begin{aligned} \mathbf{HH} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{H} \end{aligned}$$

and

$$\begin{aligned} (\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) &= \mathbf{I}-\mathbf{H}-\mathbf{H}+\mathbf{HH} \\ &= \mathbf{I}-\mathbf{H}-\mathbf{H}+\mathbf{H} \\ &= \mathbf{I}-\mathbf{H} \end{aligned}$$

3.29 First note that $(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{S_{xx}}\begin{pmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$. When x_i moves further from $\bar{x}$, both h_{ii} and h_{ij} increase.

$$
\begin{aligned}
h_{ii} &= (1 \quad x_i)\left[\frac{1}{S_{xx}}\begin{pmatrix}\sum x_i^2/n & -\bar{x}\\ -\bar{x} & 1\end{pmatrix}\right]\begin{pmatrix}1\\ x_i\end{pmatrix}\\
&= \left[\frac{1}{S_{xx}}\right]\left[\frac{\sum x_i^2}{n} - x_i\bar{x} - x_i\bar{x} + x_i^2\right]\\
&= \left[\frac{1}{S_{xx}}\right]\left[\frac{\sum x_i^2}{n} - (\bar{x}^2) + (\bar{x}^2) - 2x_i\bar{x} + x_i^2\right]\\
&= \left[\frac{1}{S_{xx}}\right]\left[\frac{\sum x_i^2 - n\bar{x}^2}{n} + (x_i - \bar{x})^2\right]\\
&= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}
\end{aligned}
$$

and

$$
\begin{aligned}
h_{ij} &= (1 \quad x_i)\left[\frac{1}{S_{xx}}\begin{pmatrix}\sum x_i^2/n & -\bar{x}\\ -\bar{x} & 1\end{pmatrix}\right]\begin{pmatrix}1\\ x_j\end{pmatrix}\\
&= \left[\frac{1}{S_{xx}}\right]\left[\frac{\sum x_i^2}{n} - x_i\bar{x} - x_j\bar{x} + x_ix_j\right]\\
&= \left[\frac{1}{S_{xx}}\right]\left[\frac{\sum x_i^2}{n} - (\bar{x}^2) + (\bar{x}^2) - x_i\bar{x} - x_j\bar{x} + x_ix_j\right]\\
&= \left[\frac{1}{S_{xx}}\right]\left[\frac{\sum x_i^2 - n\bar{x}^2}{n} + (x_i - \bar{x})(x_j - \bar{x})\right]\\
&= \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}
\end{aligned}
$$

3.30

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}]\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\beta}\\
&= \boldsymbol{\beta} + \mathbf{R}\boldsymbol{\beta}
\end{aligned}
$$

3.31 From equation 3.15b, we get that $\boldsymbol{\beta} = (\mathbf{I} - \mathbf{H})\mathbf{y}$. So substituting for $\mathbf{y}$, we get

$$
\begin{aligned}
(\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}) &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\boldsymbol{\beta}\\
&= (\mathbf{I} - \mathbf{H})\boldsymbol{\beta}
\end{aligned}
$$

3.32

$$\begin{aligned} SS_R(\boldsymbol{\beta}) &= \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \\ &= \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{y}'\mathbf{H}\mathbf{y} \end{aligned}$$

3.33

$$\begin{aligned} [\text{Corr}(\mathbf{y}, \widehat{\mathbf{y}})]^2 &= \frac{[\mathbf{y}'\widehat{\mathbf{y}}]^2}{(\mathbf{y}'\mathbf{y})(\widehat{\mathbf{y}}'\widehat{\mathbf{y}})} \\ &= \frac{[\mathbf{y}'\mathbf{H}\mathbf{y}]^2}{(\mathbf{y}'\mathbf{y})(\mathbf{y}'\mathbf{H}\mathbf{y})} \\ &= \frac{(SS_R)^2}{(S_{yy})(SS_R)} \\ &= SS_R/S_{yy} = R^2 \end{aligned}$$

3.34 $S = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - 2\lambda(\mathbf{T}\boldsymbol{\beta} - \mathbf{c})$. Then take the derivative of S with respect to $\boldsymbol{\beta}$ and λ and set them equal to zero.

$$\frac{\partial S}{\boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2(\mathbf{X}'\mathbf{X})^{-1}\tilde{\boldsymbol{\beta}} - 2\lambda\mathbf{T}' = \mathbf{0}, \qquad \frac{\partial S}{\lambda} = 2\left(\mathbf{T}\tilde{\boldsymbol{\beta}} - \mathbf{c}\right) = \mathbf{0}$$

This yields $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}'\lambda$. Now substitute this expression for $\tilde{\boldsymbol{\beta}}$ into $\frac{\partial S}{\lambda}$ and solve for λ.

$$\begin{aligned} \mathbf{T}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}'\lambda\right] - \mathbf{c} &= \mathbf{0} \\ \mathbf{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}' &= \mathbf{T}\widehat{\boldsymbol{\beta}} - \mathbf{c} \\ \lambda &= [\mathbf{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}']^{-1}\left(\mathbf{T}\widehat{\boldsymbol{\beta}} - \mathbf{c}\right) \end{aligned}$$

Finally, substitute λ back into the equation for $\tilde{\boldsymbol{\beta}}$ which gives the desired result. Note that the sign will change when you write the last part as $\mathbf{c} - \mathbf{T}\widehat{\boldsymbol{\beta}}$.

3.35

The variance of $\hat{\beta}_j$ is the j^{th} diagonal element of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Let $\mathbf{x}_j$ be the column of $\mathbf{X}$ associated with the j^{th} regressor, and let $\mathbf{X}_{-j}$ be the rest of $\mathbf{X}$. Therefore,

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} \mathbf{X}'_{-j}\mathbf{X}_{-j} & \mathbf{X}'_{-j}\mathbf{x}_j \\ \mathbf{x}'_j\mathbf{X}_{-j} & \mathbf{x}'_j\mathbf{x}_j \end{bmatrix}.$$

From Appendix C.2.1.13, the j^{th} diagonal element of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is

$$\begin{aligned} \mathbf{var}(\hat{\beta}_j) &= \sigma^2 \left[\mathbf{x}'_j[\mathbf{I} - \mathbf{X}_{-j}(\mathbf{X}'_{-j}\mathbf{X}_{-j})^{-1}\mathbf{X}'_{-j}]\mathbf{x}_j\right]^{-1} \\ &= \sigma^2 \left[\mathbf{x}'_j\mathbf{x}_j - \mathbf{x}'_j\mathbf{X}_{-j}(\mathbf{X}'_{-j}\mathbf{X}_{-j})^{-1}\mathbf{X}'_{-j}\mathbf{x}_j\right]^{-1} \end{aligned}$$

3.36 Since $R^2 = SS_R/S_{yy}$, we need to show that the sum of squares for regression for model B, SS_{Rb} is greater than the sum of squares for regression for model A, SS_{Ra}. We can do this using partitioning SS_R into sequential sums of squares. Consider i parameters in $\boldsymbol{\beta}_1$ and j parameters in $\boldsymbol{\beta}_2$. Then model B is using $(i \times j)$ parameters of which the first i are the same as model A. Then SS_{Rb} equals

$$\begin{aligned} R(\beta_{i1}, \beta_{i2}, \ldots, \beta_{ii}, \beta_{j1}, \beta_{j2}, \ldots, \beta_{jj}|\beta_0) &= R(\beta_{i1}, \beta_{i2}, \ldots, \beta_{ii}|\beta_0) \\ &\quad + R(\beta_{j1}, \beta_{j2}, \ldots, \beta_{jj}|\beta_0, \beta_{i1}, \beta_{i2}, \ldots, \beta_{ii}) \end{aligned}$$

Since the second term on the right is a sum of squares, it must be greater than or equal to zero. Thus, $SS_{Rb} \geq SS_{Ra}$ which is equivalent to $R^2_B \geq R^2_A$.

3.37 $\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{y}$. Therefore,

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}_1) &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' E(\mathbf{y}) \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1\boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 \end{aligned}$$

The estimate is unbiased if $\mathbf{X}_1'\mathbf{X}_2$ is $\mathbf{0}$, which happens if $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal.

3.38

$$\begin{aligned} \sum_{i-1}^{n} Var(\widehat{y}_i) &= \sum_{i=1}^{n} \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i(\sigma^2) \\ &= \sigma^2\left(\sum_{i=1}^{n} h_{ii}\right) \\ &= \sigma^2(\text{rank of } \mathbf{X}) = p\sigma^2 \end{aligned}$$

3.39 The j^{th} VIF is the j^{th} diagonal element of $(\mathbf{W}'\mathbf{W})^{-1}$, where $\mathbf{W}'\mathbf{W}$ is the correlation matrix. Let $\mathbf{w}_j$ be the column of $\mathbf{W}$ associated with the j^{th} regressor, and let $\mathbf{W}_{-j}$ be the rest of $\mathbf{W}$. Therefore,

$$(\mathbf{W}'\mathbf{W})^{-1} = \begin{bmatrix} \mathbf{W}_{-j}'\mathbf{W}_{-j} & \mathbf{W}_{-j}'\mathbf{w}_j \\ \mathbf{w}_j'\mathbf{W}_{-j} & \mathbf{w}_j'\mathbf{w}_j \end{bmatrix}.$$

We note that $\mathbf{W}'\mathbf{W}$ is the correlation matrix. As a result $\mathbf{w}_j'\mathbf{w}_j = 1$. From Appendix C.2.1.13, the j^{th} diagonal element of $(\mathbf{W}'\mathbf{W})^{-1}$ is

$$\begin{aligned} \left[\mathbf{w}_j'[\mathbf{I} - \mathbf{W}_{-j}(\mathbf{W}_{-j}'\mathbf{W}_{-j})^{-1}\mathbf{W}_{-j}']\mathbf{w}_j\right]^{-1} &= \left[\mathbf{w}_j'\mathbf{w}_j - \mathbf{w}_j'\mathbf{W}_{-j}(\mathbf{W}_{-j}'\mathbf{W}_{-j})^{-1}\mathbf{W}_{-j}'\mathbf{w}_j\right]^{-1} \\ &= \left[1 - \mathbf{w}_j'\mathbf{W}_{-j}(\mathbf{W}_{-j}'\mathbf{W}_{-j})^{-1}\mathbf{W}_{-j}'\mathbf{w}_j\right]^{-1}. \end{aligned}$$

Since $\mathbf{1}'\mathbf{w}_j = 0$. If we regress $\mathbf{w}$ on $\mathbf{W}_{-j}$, we obtain that

$$\mathbf{SS}_{total} = \mathbf{w}_j'\mathbf{w}_j - \mathbf{w}_j'\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{w}_j = 1,$$

and

$$\mathbf{SS}_{reg} = \mathbf{w}_j'\mathbf{W}_{-j}(\mathbf{W}_{-j}'\mathbf{W}_{-j})^{-1}\mathbf{W}_{-j}'\mathbf{w}_j.$$

As a result, if we regress $\mathbf{w}_j$ on $\mathbf{W}_{-j}$, the resulting R_j^2 is

$$\begin{aligned} R_j^2 &= \frac{\mathbf{SS}_{reg}}{\mathbf{SS}_{total}} \\ &= \mathbf{w}_j'\mathbf{W}_{-j}(\mathbf{W}_{-j}'\mathbf{W}_{-j})^{-1}\mathbf{W}_{-j}'\mathbf{w}_j. \end{aligned}$$

As a result, the j^{th} diagonal element of $(\mathbf{W}'\mathbf{W})^{-1}$ is

$$\left[1 - \mathbf{w}_j'\mathbf{W}_{-j}(\mathbf{W}_{-j}'\mathbf{W}_{-j})^{-1}\mathbf{W}_{-j}'\mathbf{w}_j\right]^{-1} = [1 - R_j^2]^{-1} = \frac{1}{1 - R_j^2}.$$

3.40 If $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, then $\mathbf{T}\widehat{\boldsymbol{\beta}} - \mathbf{c} \sim N\left(\mathbf{T}\boldsymbol{\beta} - \mathbf{c}, \mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{T}'(\sigma^2)\right)$. Note that the $\text{rank}\left[\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{T}'\right] = \text{rank}[\mathbf{T}] = q$. First, we need to show that

$$Q/\sigma^2 = \left(\mathbf{T}\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left[\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{T}'\right]^{-1}\left(\mathbf{T}\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)/\sigma^2$$

is distributed as χ_q^2 under H_0. Since $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$, then

$$Q = \left(\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} - \mathbf{c}\right)'\left(\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{T}'\right)^{-1}\left(\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} - \mathbf{c}\right)$$

Now $\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} - \mathbf{c} = \mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left[\mathbf{y} - \mathbf{X}\mathbf{T}'\left(\mathbf{T}\mathbf{T}'\right)^{-1}\mathbf{c}\right]$. Hence,

$$Q = \left[\mathbf{y} - \mathbf{X}\mathbf{T}'\left(\mathbf{T}\mathbf{T}'\right)^{-1}\mathbf{c}\right]' \left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{T}'\right)\left(\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{T}'\right)^{-1}\left(\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right) \left[\mathbf{y} - \mathbf{X}\mathbf{T}'\left(\mathbf{T}\mathbf{T}'\right)^{-1}\mathbf{c}\right]$$

Thus Q is expressed as a quadratic form in the vector $\mathbf{y} - \mathbf{X}\mathbf{T}'\left(\mathbf{T}\mathbf{T}'\right)^{-1}\mathbf{c}$. It is straightforward to verify that the inner matrix of Q is idempotent. Also, since under H_0, $\mathbf{T}\boldsymbol{\beta} = \mathbf{c}$, the noncentrality parameter λ is zero. Thus $Q/\sigma^2 \sim \chi^2_q$. Now we consider $SS_E = \mathbf{y}\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{y}$. SS_E can also be written as a quadratic form in terms of the vector $\mathbf{y} - \mathbf{X}\mathbf{T}'\left(\mathbf{T}\mathbf{T}'\right)^{-1}\mathbf{c}$:

$$SS_E = \left[\mathbf{y} - \mathbf{X}\mathbf{T}'\left(\mathbf{T}\mathbf{T}'\right)^{-1}\mathbf{c}\right]'\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\left[\mathbf{y} - \mathbf{X}\mathbf{T}'\left(\mathbf{T}\mathbf{T}'\right)^{-1}\mathbf{c}\right].$$

Since the matrix in this quadratic form is still $(\mathbf{I} - \mathbf{H})$, is it clear it is idempotent and $\lambda = 0$. Thus SS_E/σ^2 is distributed χ^2_{n-p}. Note that

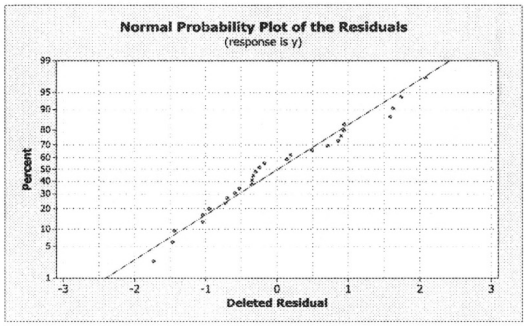
$$\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{T}'\right)\left(\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{T}'\right)^{-1}\left(\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right) = \mathbf{0}$$

Therefore, SS_E/σ^2 and Q/σ^2 are independently distributed as central chi-square variables under H_0. Hence, $F = \dfrac{Q/q}{MS_E} \overset{H_0}{\sim} F_{q,n-p}$.

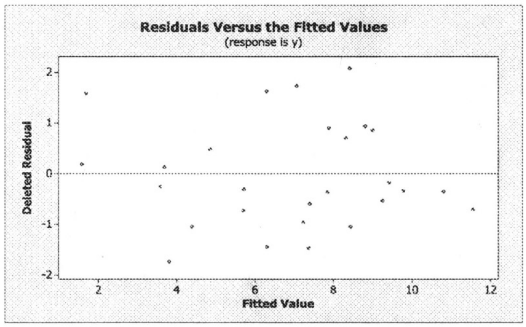
Now under the alternative, $\mathbf{T}\boldsymbol{\beta} \neq \mathbf{c}$. Therefore, we get $\lambda = \left(\mathbf{T}\boldsymbol{\beta} - \mathbf{c}\right)'\left(\mathbf{T}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{T}'\right)^{-1}\left(\mathbf{T}\boldsymbol{\beta} - \mathbf{c}\right)$. Hence, $Q/\sigma^2 \overset{H_1}{\sim} \left(\chi^2_q\right)^*$ which is a noncentral chi-square. Thus $F = \dfrac{Q/q}{MS_E} \overset{H_1}{\sim} F^*_{q,n-p}$ which is a noncentral F-distribution.

Chapter 4: Model Adequacy Checking

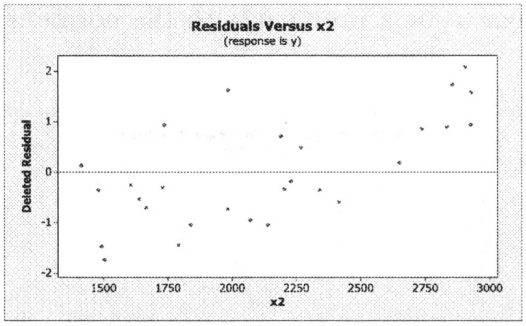
4.1 a. There does not seem to be a problem with the normality assumption.



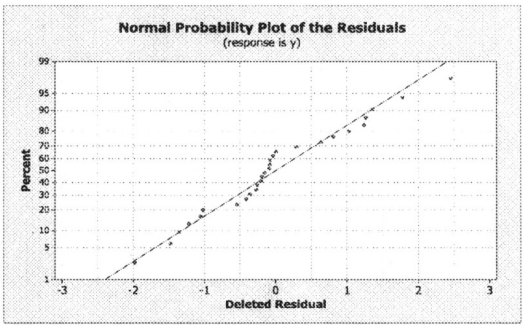
b. The model seems adequate.



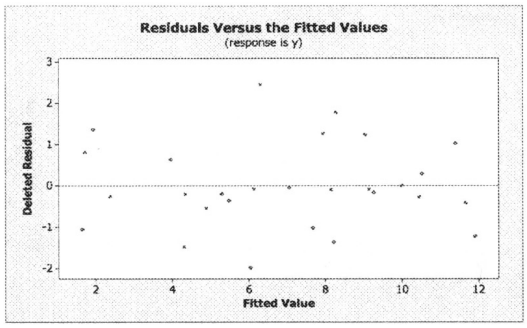
c. It appears that the model will be improved by adding x_2.



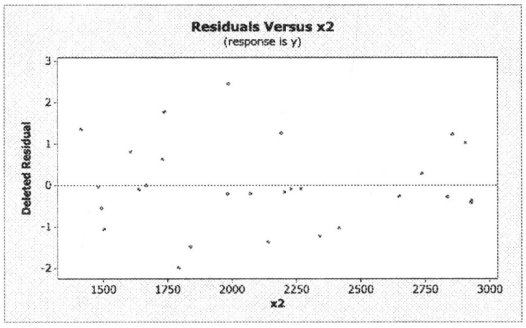
4.2 a. There looks to be a slight problem with normality.

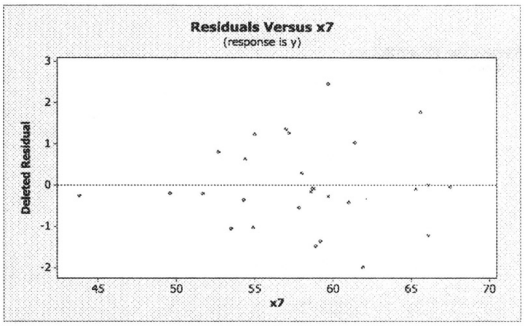


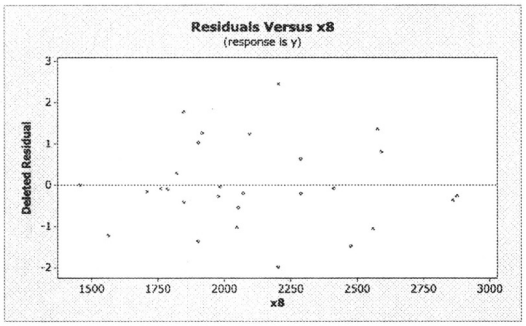
b. The plot looks good.



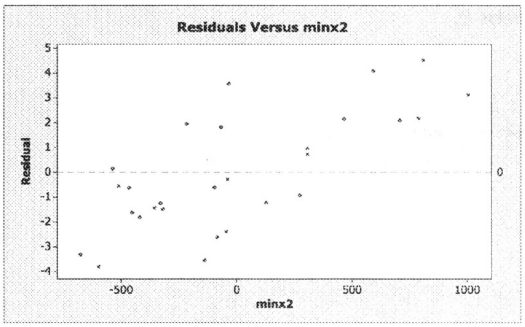
c. The plot for x_8 looks ok, the plot for x_2 shows mild nonconstant variance, and the plot for x_7 exhibits nonconstant variance.

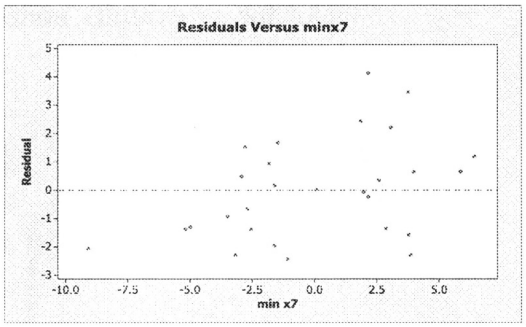


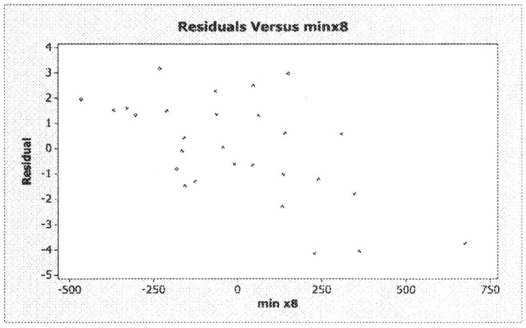




d. These plots indicate whether the relationships between the response and the regressor variables are correct. They show that there is not a strong linear relationship between the response and x_7.

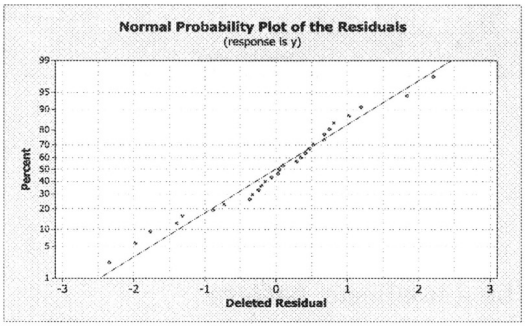




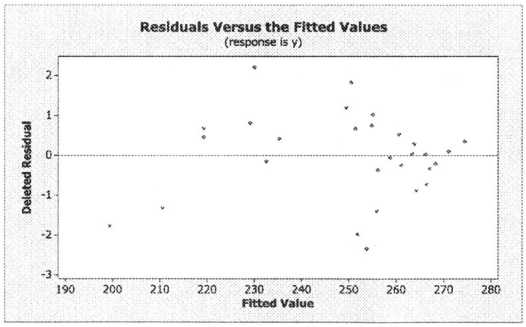


e. They can be used to determine influential points and outliers. For this example, the first observation is identified as a possible outlier.

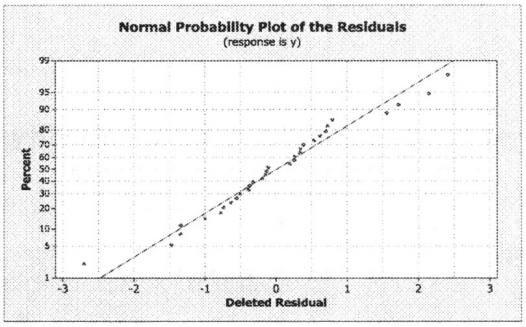
4.3 a. There does not seem to be any problem with normality.



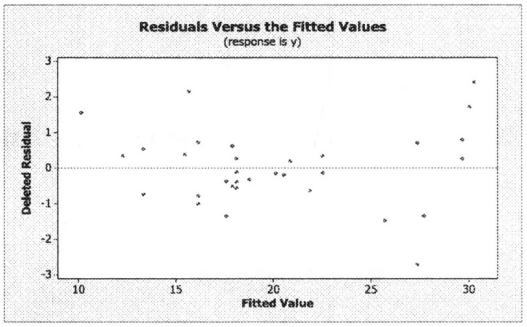
b. There appears to be a pattern and possible nonconstant variance.



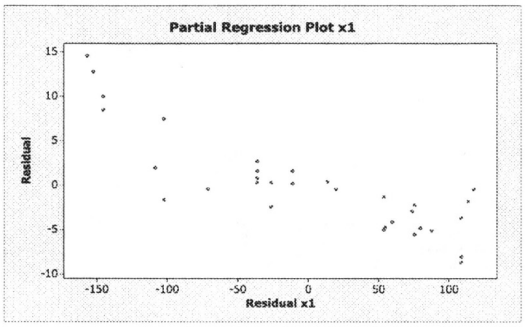
4.4 a. There seems to be a slight problem with normality.

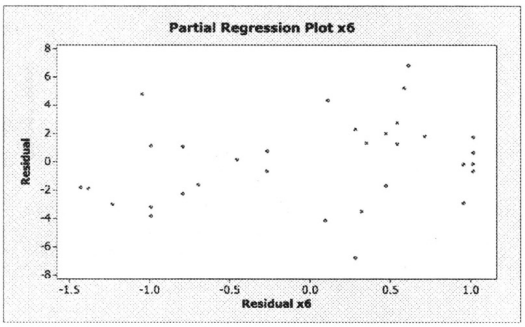


b. There appears to be a nonlinear pattern.



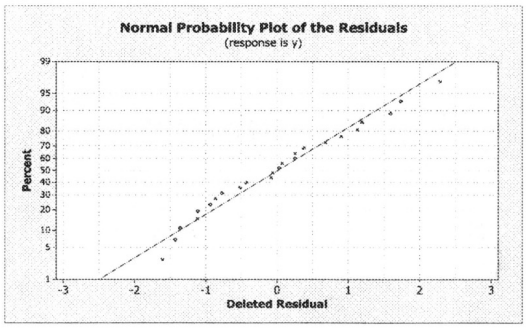
c. There is a linear pattern for x_1. The graph for x_6 shows no pattern and indicates it might be unnecessary to include it in the model.



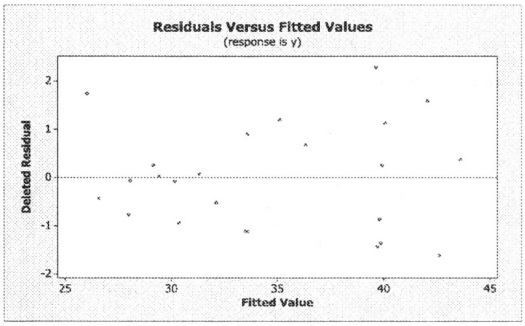


d. These residual indicate that observations 12 and 15 are possible outliers.

4.5 a. There does not appear to be a problem with normality.



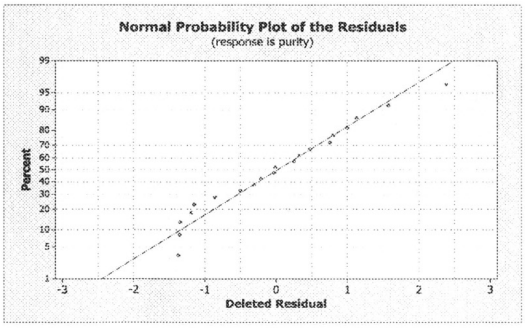
b. There is a slight drift upward in the plot.



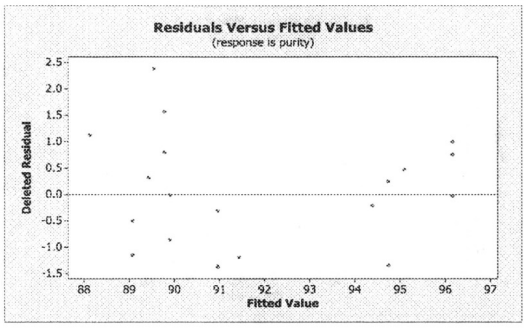
c. Yes, after x_1 is in the model, most of the other variables contribute very little.

d. They indicate observation 16 is a possible outlier.

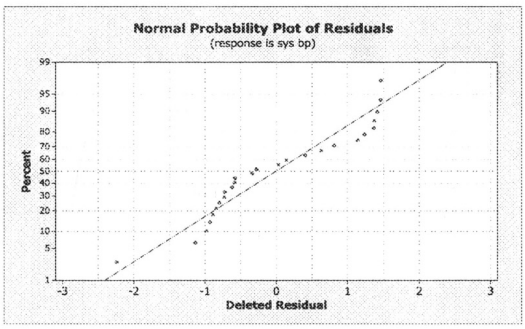
4.6 a. There is a evidence of a problem with normality.



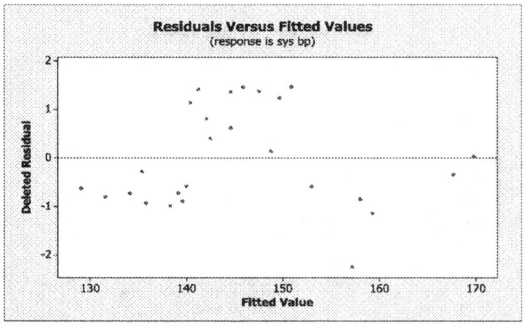
b. There is a nonlinear pattern.



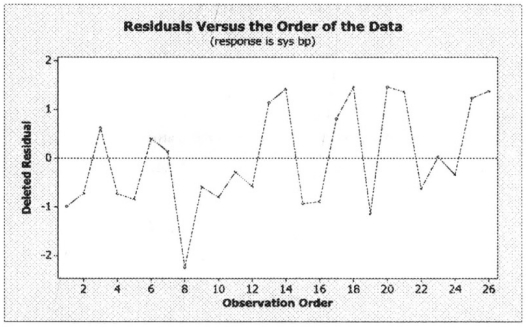
4.7 a. There is a serious problem with normality.



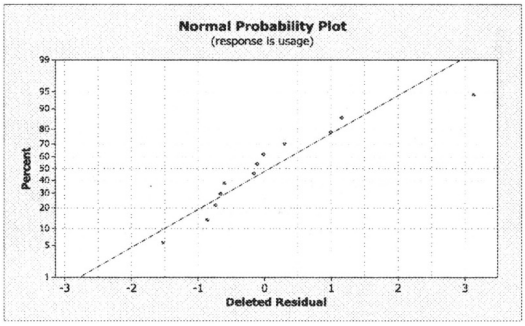
b. There is a nonlinear pattern.



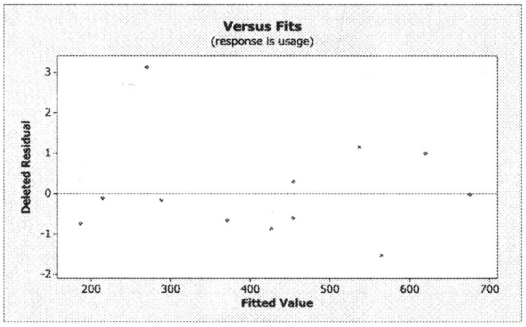
c. There does not appear to be any pattern with time.



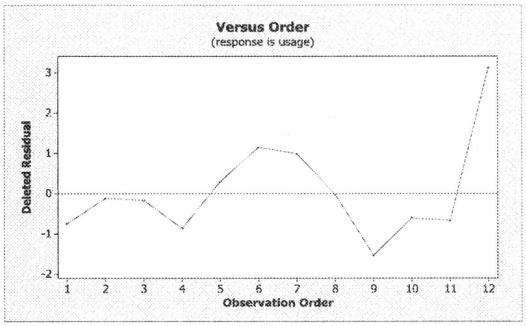
4.8 a. The plot shows normality is not a big problem.



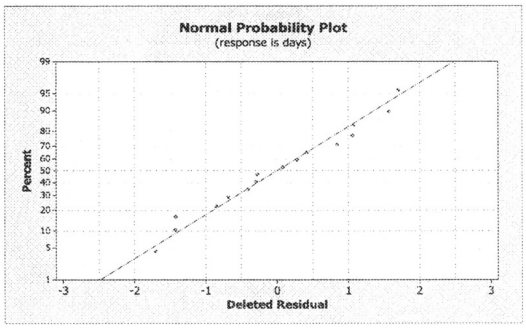
b. There is a pattern.



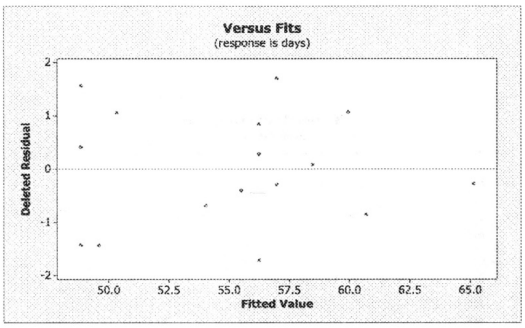
c. The plot shows positive autocorrelation.



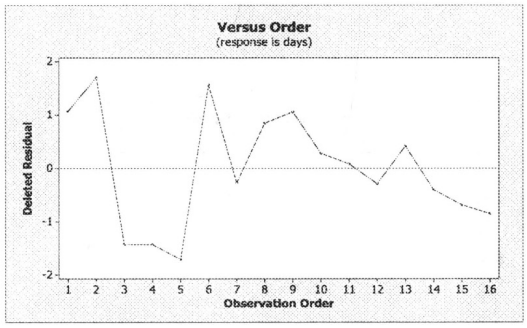
4.9 a. There appears to be no problem with normality.



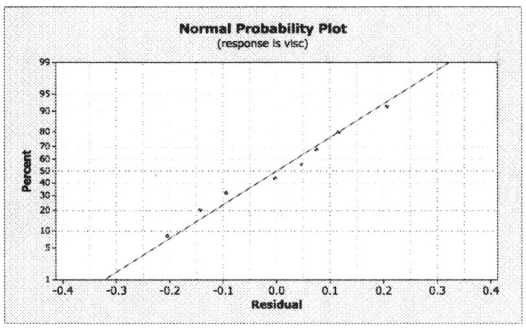
b. There is a pattern.



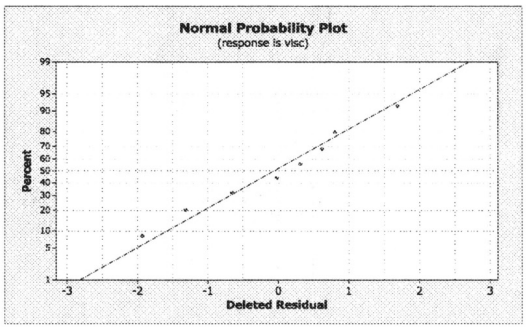
c. The plot shows positive autocorrelation.



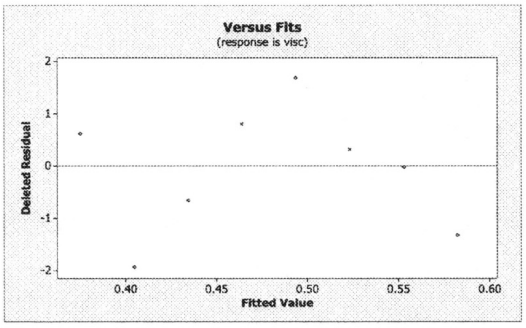
4.10 a. There appears to be no problem with normality.



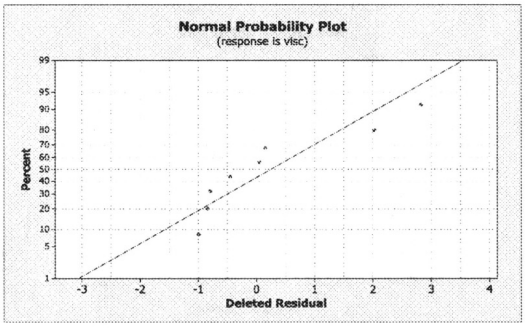
b. There is no real difference between the two plots.



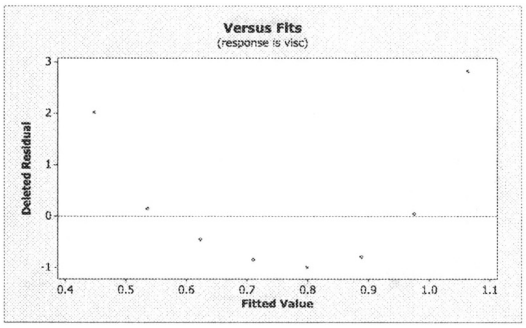
c. The plot shows a definite pattern.



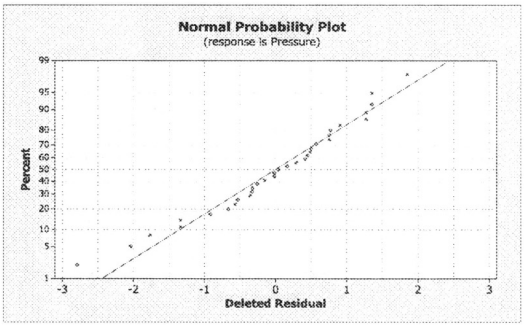
4.11 a. There is a slight problem with normality.



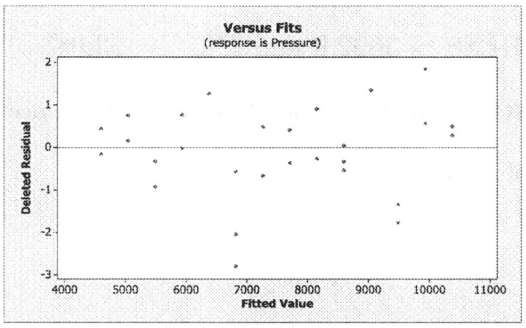
b. There is a quadratic pattern indicating that a second-order term is needed.



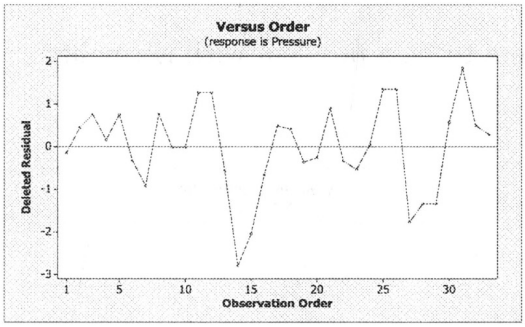
4.12 a. There is no problem with normality.



b. There is no pattern.



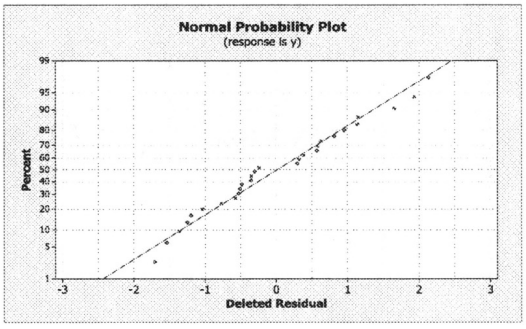
c. There is no pattern.



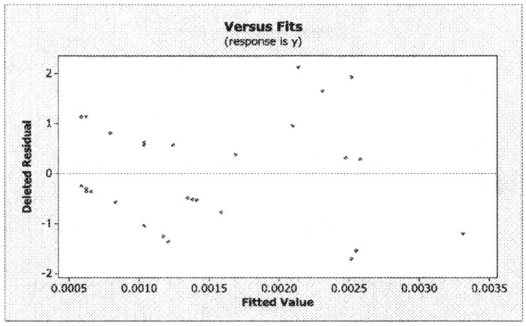
4.13 When x_7 and x_6 are in the model, PRESS = 3388.6 and $R^2_{Pred} = 56.94\%$. When just x_6 is in the model, PRESS = 3692.9 and $R^2_{Pred} = 53.08\%$. The residual plots for both models show nonconstant variance and departure from normality. There is no insight into the best choice of model.

4.14 When x_1 and x_6 are in the model, PRESS = 328.8 and $R^2_{Pred} = 73.43\%$. When just x_1 is in the model, PRESS = 337.2 and $R^2_{Pred} = 72.75\%$. Both models give basically the same values.

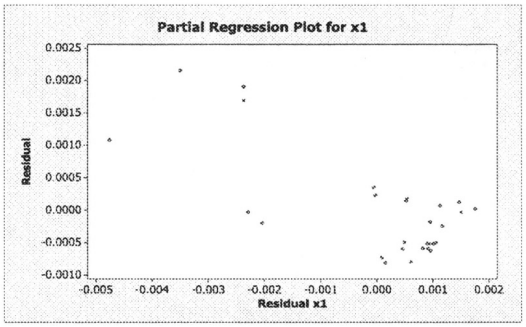
4.15 a. There does not seem to be a problem with normality.

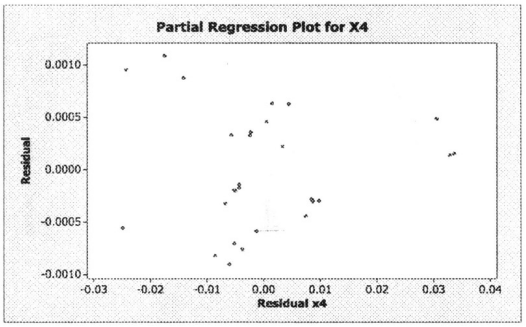


b. There is a nonlinear pattern.

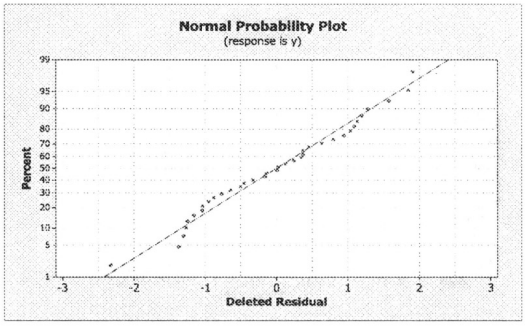


c. x_1 shows a linear pattern but x_4 does not.

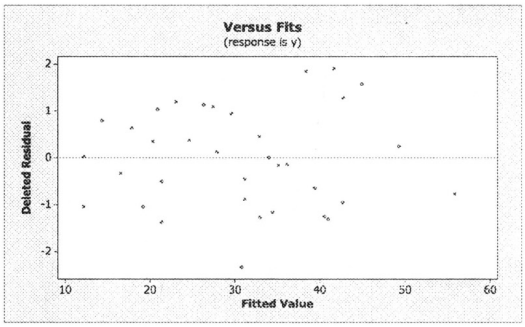




4.16 a. There is some problems in the tails.

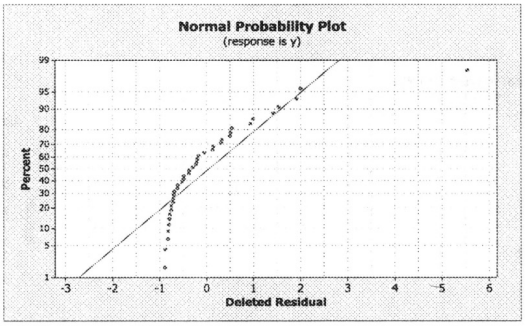


b. The fit seems pretty good.

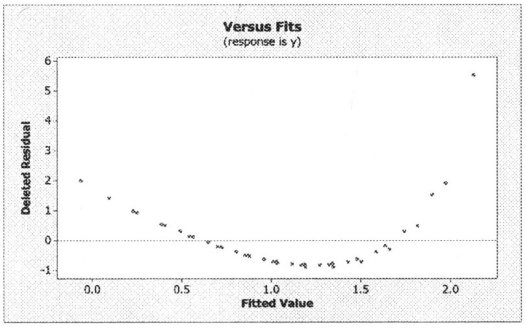


c. When x_1 and x_2 are in the model, PRESS = 916.41 and $R^2_{Pred} = 80.76\%$. When just x_2 is in the model, PRESS = 2825.62 and $R^2_{Pred} = 40.66\%$. The model with both x_1 and x_2 is more likely to provide better prediction of new data.

4.17 a. There is a serious problem with normality.

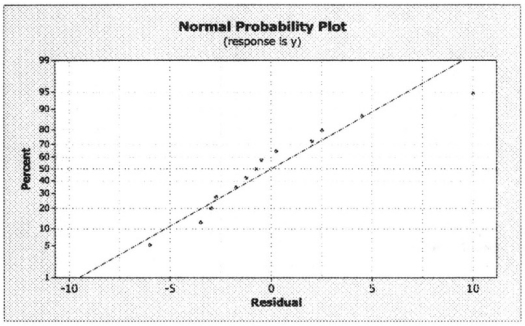


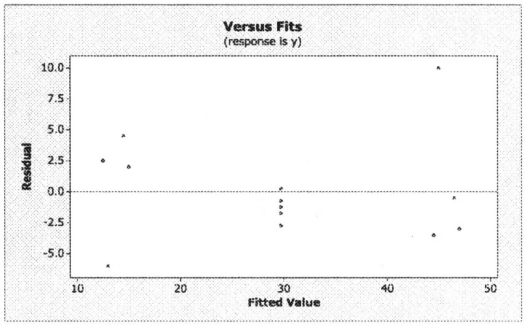
b. There is a nonlinear pattern.



c. When x_1 and x_2 are in the model, PRESS = 3.11 and $R^2_{Pred} = 77.75\%$. When just x_2 is in the model, PRESS = 6.77 and $R^2_{Pred} = 51.54\%$. The model with both x_1 and x_2 is more likely to provide better prediction of new data.

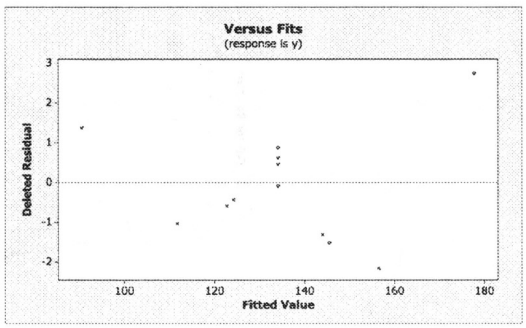
4.18 a. Normality seems ok. There is a nonconstant variance problem. There is very little variability at the center points. The observation with $y = 55$ is a potential outlier.

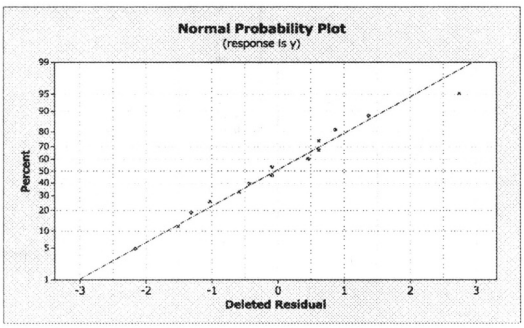




b. For lack of fit, $F_0 = \frac{39.16}{1.25} = 31.33$ with $p = 0.003$. There is evidence of lack of fit of the linear model.

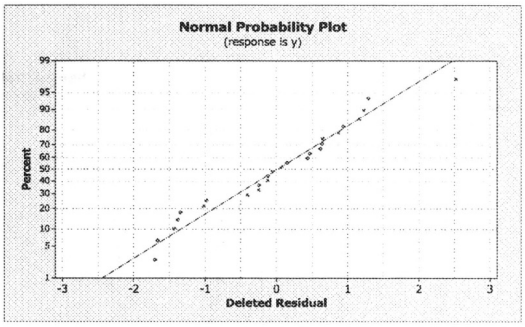
4.19 a. Normality does not seem to be a problem. There is a nonlinear pattern in the residual plot versus the fitted values. The observation with $y = 198$ is a potential outlier.

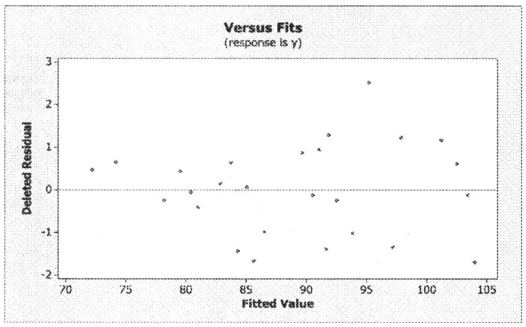




b. For lack of fit, $F_0 = \frac{299.5}{25.4} = 11.81$ with $p = 0.008$. There is evidence of lack of fit of the linear model.

4.20 a. There is a problem with normality. There is a problem with nonconstant variance. The observation with $y = 115.2$ is a potential outlier.





b. There is no test for lack of fit since there are no replicate points. It is possible to use the near-neighbor approach.

4.21 $E(MS_{PE}) = \frac{1}{n-m} E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2\right]$ where

$$E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2\right] = E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}^2 - 2y_{ij}\bar{y}_i + \bar{y}_i^2\right)\right]$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[E(y_{ij}^2) - 2E\left(y_{ij}\sum_{j^*=1}^{n_i}\frac{y_{ij^*}}{n_i}\right) + E(\bar{y}_i^2)\right]$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[\sigma^2 - 2E\left(y_{ij}\sum_{j^*=1}^{n_i}\frac{y_{ij^*}}{n_i}\right) + \frac{\sigma^2}{n_i}\right]$$

$$= n\sigma^2 - 2\sum_{i=1}^{m}\left(\sum_{j=1}^{n_i}\sum_{j^*=1}^{n_i}\frac{y_{ij}y_{ij^*}}{n_i}\right) + m\sigma^2$$

$$= n\sigma^2 - 2\sum_{i=1}^{m}\frac{n_i\sigma^2}{n_i} + m\sigma^2$$

$$= n\sigma^2 - 2m\sigma^2 + m\sigma^2$$

$$= (n-m)\sigma^2$$

Therefore, $E(MS_{PE}) = \sigma^2$.

Now, $SS_{Res} = SS_{PE} + SS_{LOF}$ and so $SS_{LOF} = SS_{Res} - SS_{PE}$. Using Appendix C for $E(SS_{Res})$ when the model is under specified and using $E(SS_{PE}) = (n-m)\sigma^2$ from above, we get

$$E(SS_{Res}) - E(SSPE) = (n-2)\sigma^2 + \sum_{i=1}^{m}\left[E(y_i) - \beta_0 - \beta_1 x_i\right]^2 - (n-m)\sigma^2$$

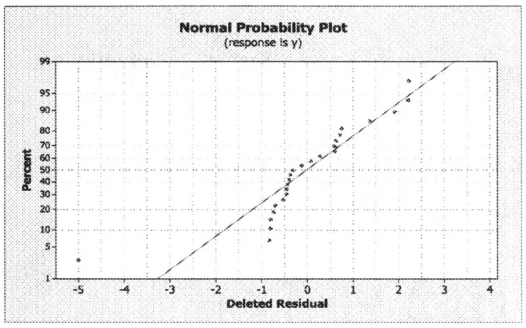
$$= (m-2)\sigma^2 + \sum_{i=1}^{m}\left[E(y_i) - \beta_0 - \beta_1 x_i\right]^2$$

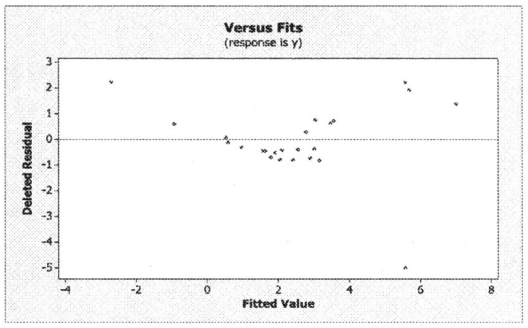
Therefore,

$$E(MS_{LOF}) = E\left(\frac{SS_{LOF}}{m-2}\right)$$

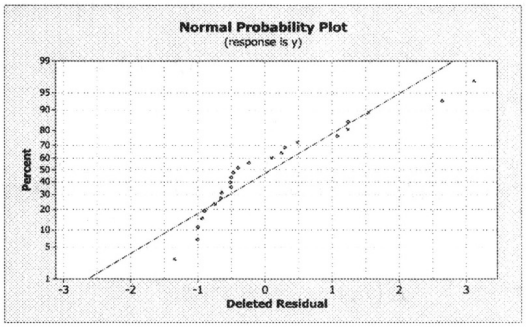
$$= \sigma^2 + \frac{\sum_{i=1}^{m}\left[E(y_i) - \beta_0 - \beta_1 x_i\right]^2}{m-2}$$

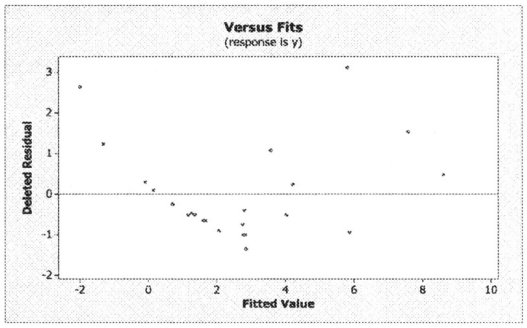
4.22 a. There is a problem with normality and there is a nonlinear pattern to the residual plot. Observation 2 is a potential outlier. The model does not fit well.



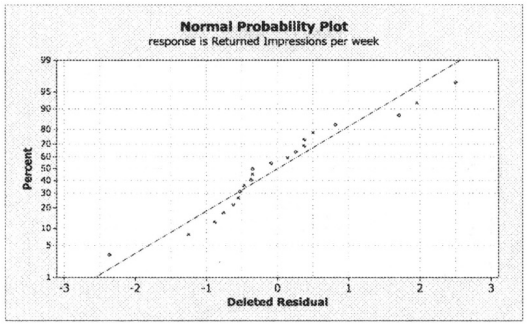


b. There is still a problem with normality and there is still a nonlinear pattern to the residual plot. Several observations are potential outliers. The model still does not fit well.

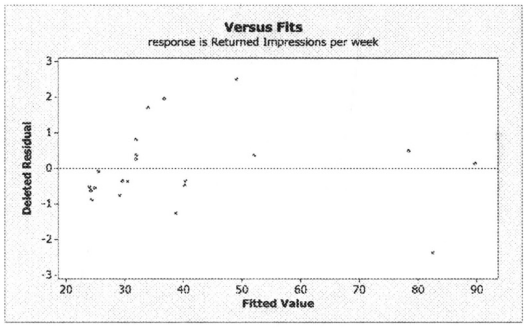




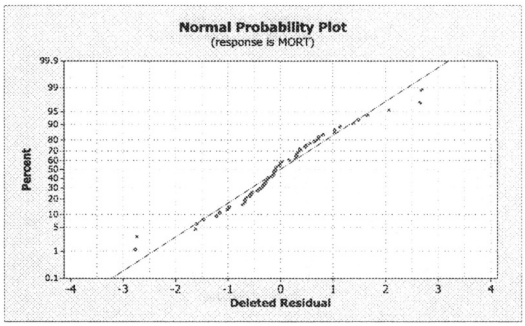
4.23 a. There does not appear to be a problem with normality.



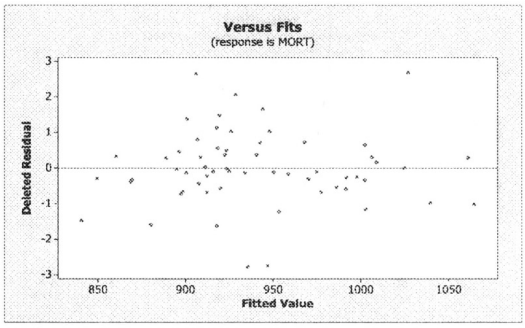
b. There appears to be a slight pattern and possible nonconstant variance.



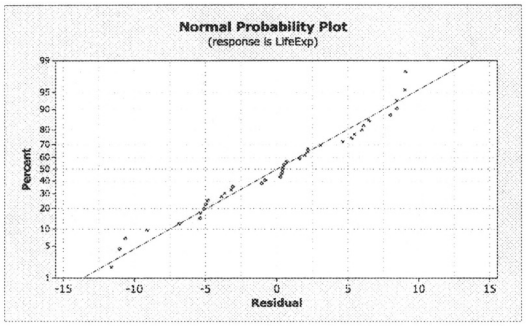
4.24 a. There does not appear to be a problem with normality.

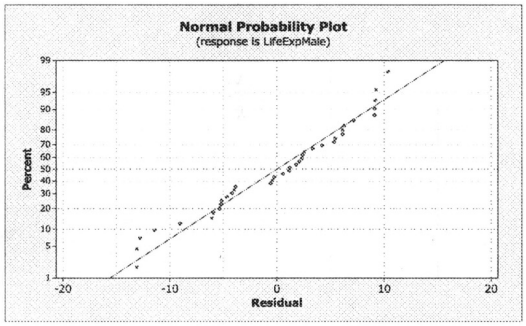


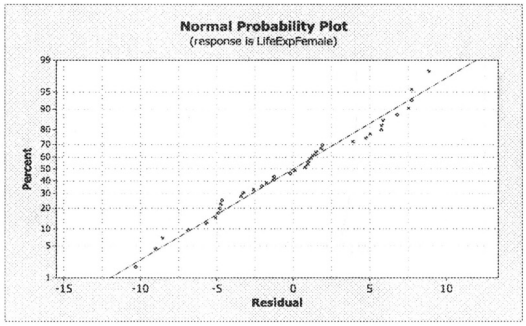
b. There is no apparent pattern.



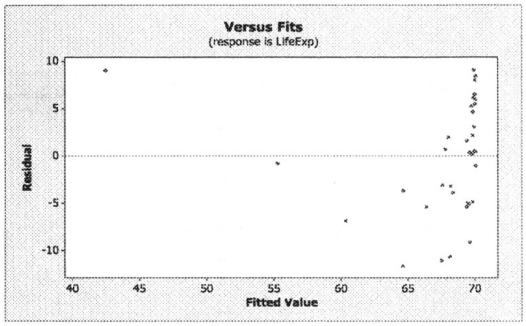
4.25 a. The plots for LifeExp and LifeExpMale show problems in the tails, but the LifeExpFemale plot shows no problems in normality.

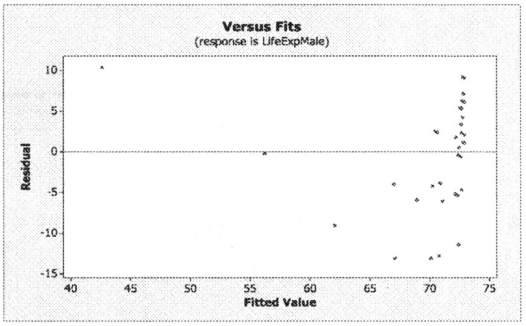


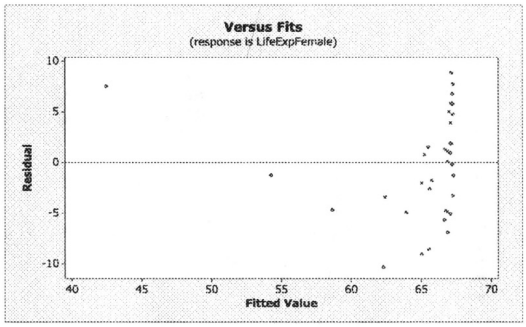




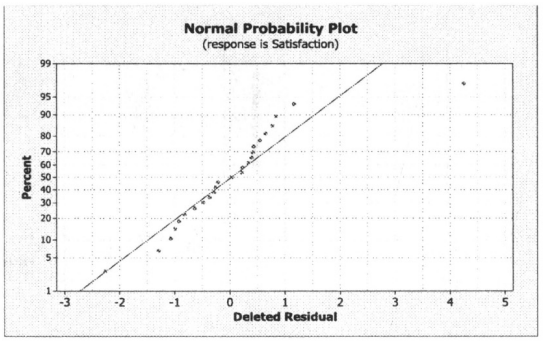
b. All three plots show a nonlinear pattern.

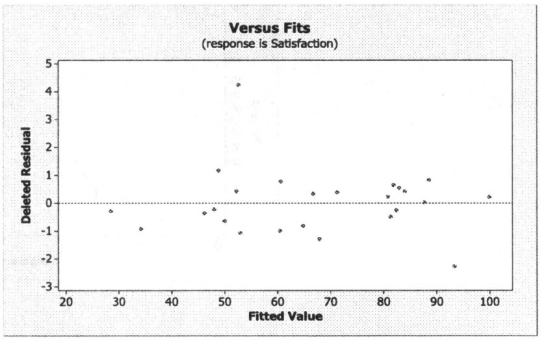




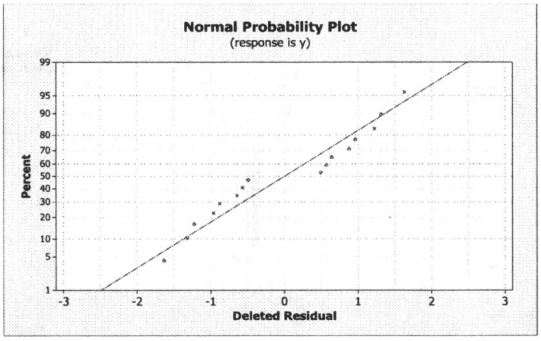


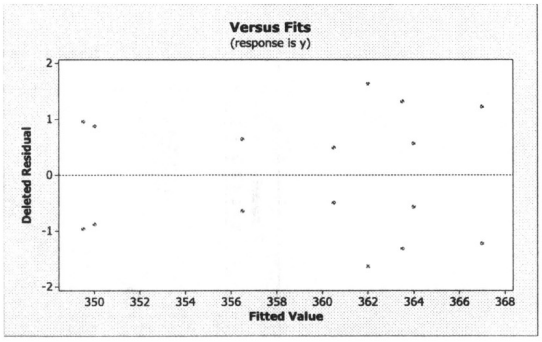
4.26 The normal probability plot indicates some possible deviations from normality in the tails of the distribution; however this may be a result of observations 9 and 17 being possible outliers. The Deleted Residual versus Fit plot also indicates that observations 9 and 17 are possible outliers but otherwise there is no apparent pattern.



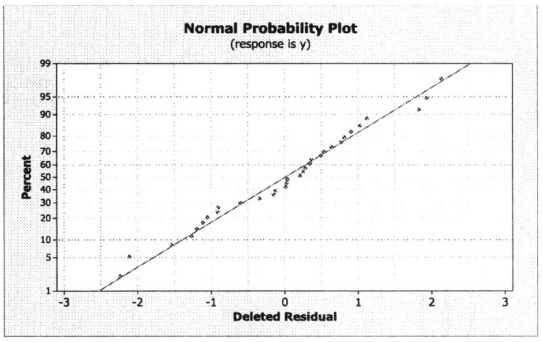


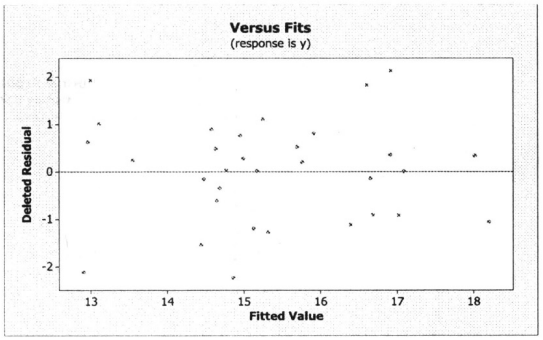
4.27 The residual analysis for the fuel consumption data indicates separation which may be a result of a variable missing from the model. There is also a pattern in the Deleted Residual versus Fit plot indicating the model is not adequate.



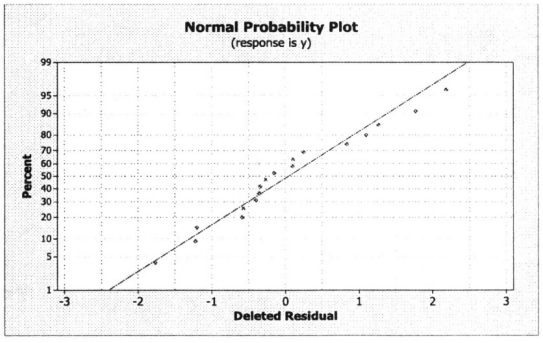


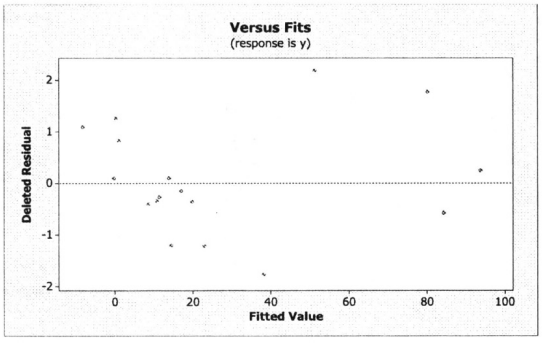
4.28 The residual analysis for the wine quality of young red wines data indicates an adequate model. There appears to be no problem with normality based on the normal probability plot and there is also no apparent pattern in the Deleted Residuals versus Fit plot.





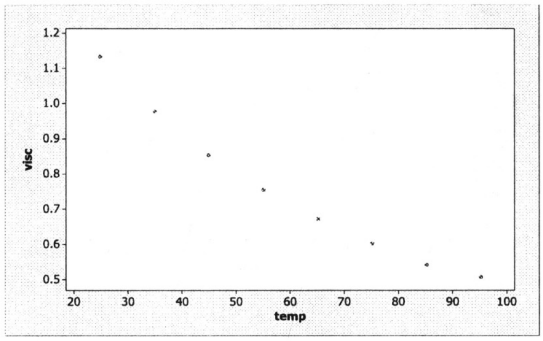
4.29 The residual analysis for the methanol oxidation data indicates no major problems with normality from the normal probability plot. However, the Deleted Residuals versus Fits plot shows a nonlinear pattern indicating the model does not fit the data well.



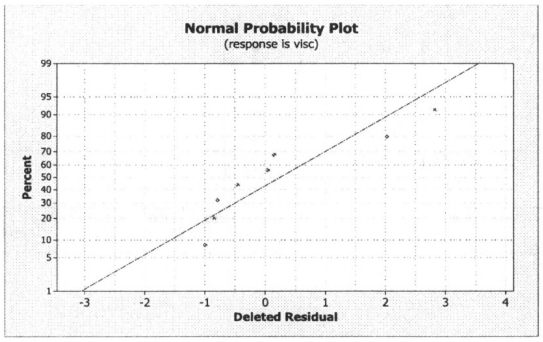


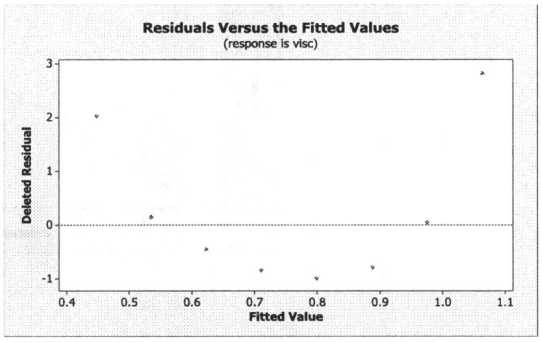
Chapter 5: Transformations and Weighting to Correct Model Inadequacies

5.1 a. It has a nonlinear pattern.



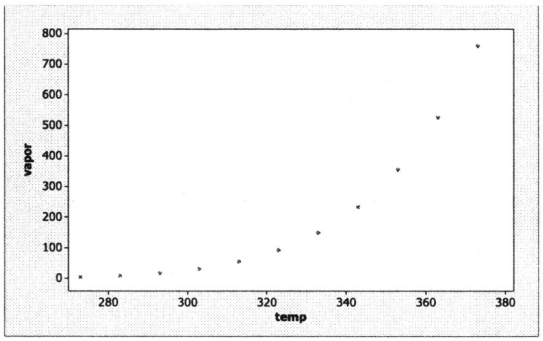
b. While $R^2 = 96\%$, the residual plot shows a nonlinear pattern and normality is violated.



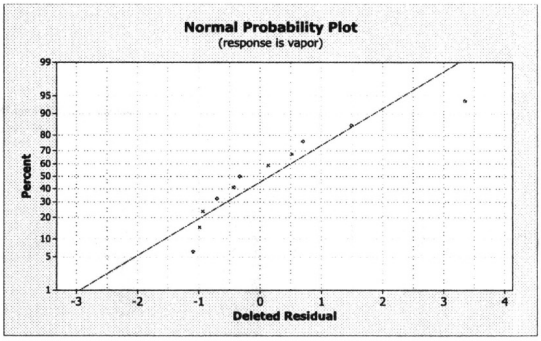


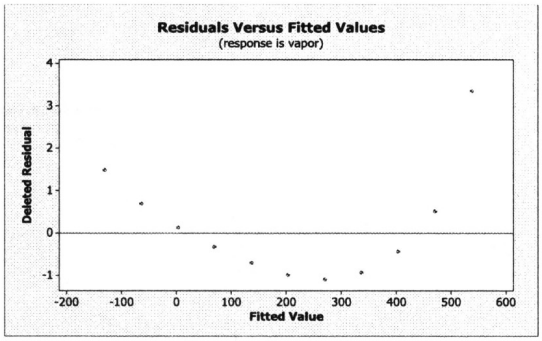
c. There is a slight improvement in the model.

5.2 a. There is a nonlinear pattern.



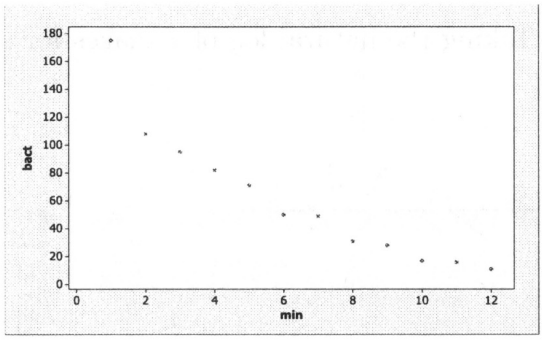
b. There is a problem with normality and a nonlinear pattern in the residuals.



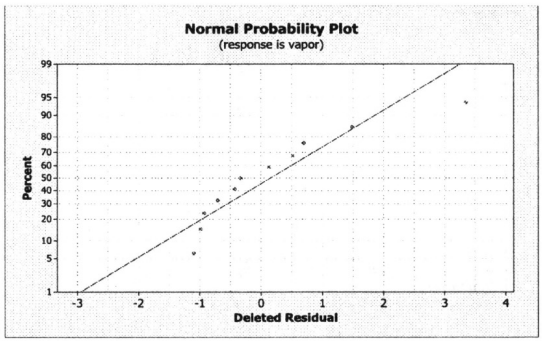


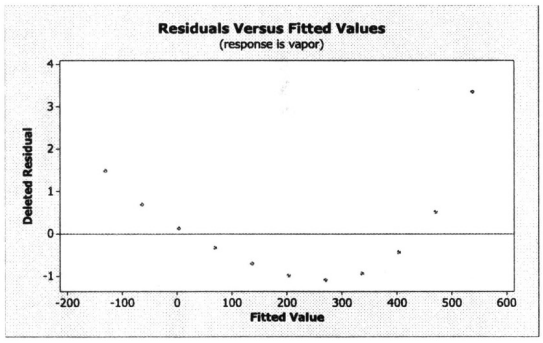
c. There is a slight improvement in the model.

5.3 a. There is a nonlinear pattern.



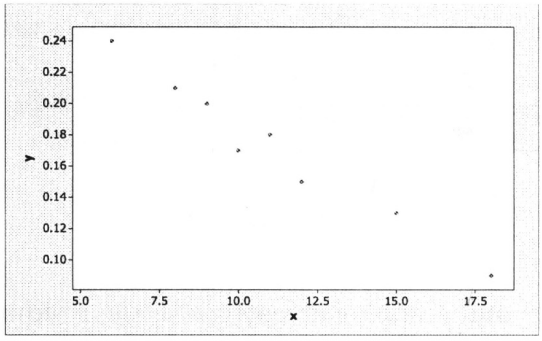
b.There is a problem with normality and a nonlinear pattern in the residuals. Observation 1 is an outlier.

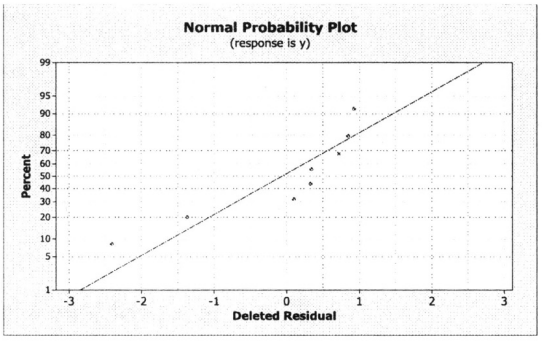


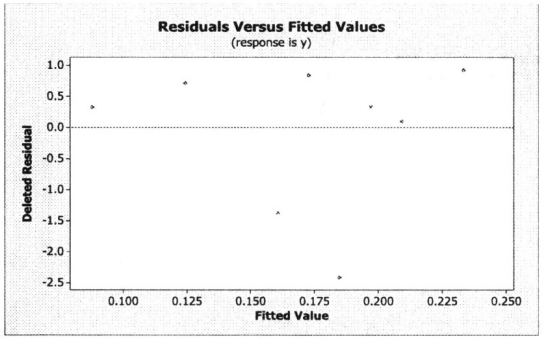


c. Fit the number of bacteria versus the natural log of the minutes. The first observation is still an outlier but otherwise the model fits fine.

5.4 The scatterplot looks fine. There is a problem with normality and the residual plot does not look good. Taking the natural log of x makes for a better model.



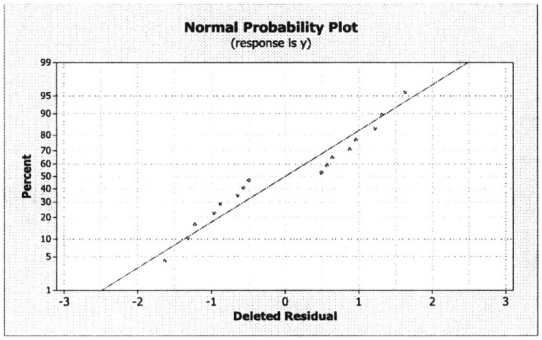


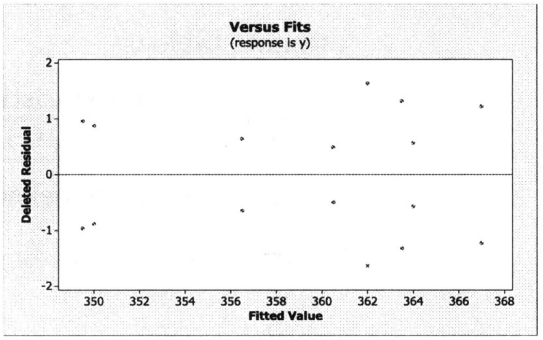


5.5 a. $\hat{y} = -31.698 + 7.277x$. There is a nonlinear pattern to the residuals.

b. Taking the natural log of defects versus weeks makes for a better model.

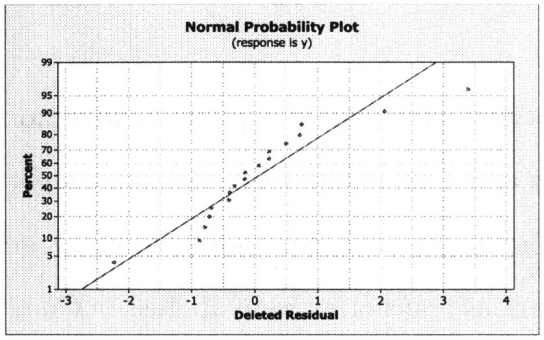
5.6 a. The residual analysis from Exercise 4.27 indicates a problem with normality and a pattern in the residuals versus fits graph that indicates the model is not fitting the fuel consumption data well. However, this pattern does not suggest a transformation that would improve the analysis. Various transforms were applied but none improved the fit of the model. See problem 5.20 for an appropriate analysis of these data.

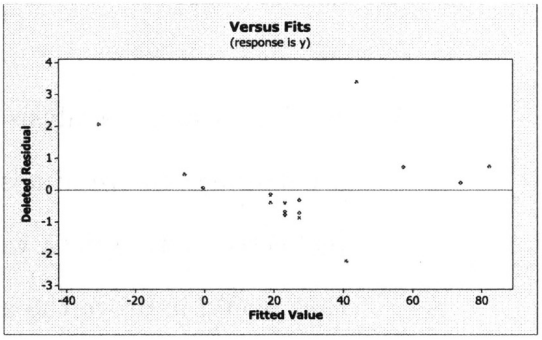




5.7 Prior to the residual analysis for the methanol oxidation data, the original model was reduced to only the significant regressors. This reduces the model from 5 regressors down to 2. This leaves regressors x_1 and x_3 in the model, reactor system and reactor residence time (seconds).

The residual plots for this reduced model are seen below. There is a problem with both the normality ansumption and there is also a pattern in the residual versus fits plot.





A log transformation was performed on the response percent conversion. Regressor x_1 is not longer significant. The new regression equation is $log(\widehat{y}) = 21.4 - 2.49x_3$. The estimate table is:

Coefficient	test statistic	p-value
β_3	-10.13	0.000

The residuals plots below show no problem with the normality assumption and also show less of a pattern in the residual versus fits plot.