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

**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.



For general information on our other products and services please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

***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

Solutions Manual to Accompany

# Introduction to Linear Regression Analysis

Fifth Edition

# 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 $\text{in}^3$ 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 $\text{in}^3$ 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 $\text{in}^3$ while part e. is an interval estimator on a future observation at 275 $\text{in}^3$. 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 $\rho$ 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 $\rho$ 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 $\beta_0$ is significant. The model with the intercept is superior.

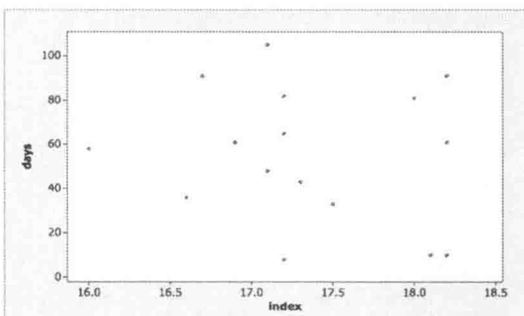
2.12 a. $\widehat{y} = -6.33 + 9.21x$

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 $527.759 \pm 3.169(2.063) = (521.22, 534.29)$.

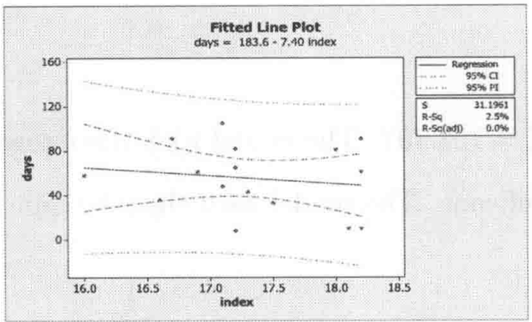
2.13 a.



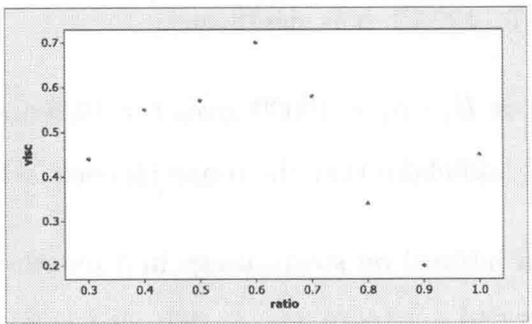

b. $\widehat{y} = 183.596 - 7.404x$

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

d.



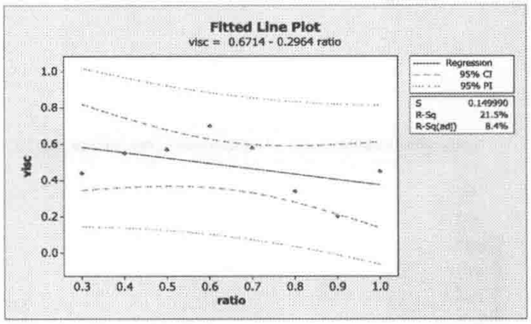

2.14 a.




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.

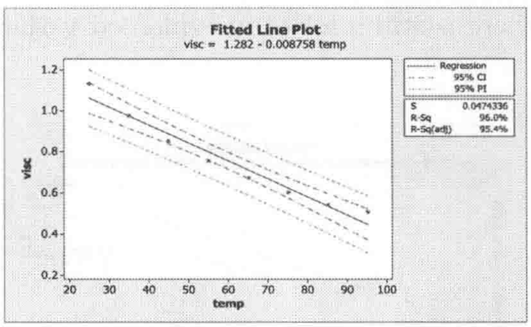
d.




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

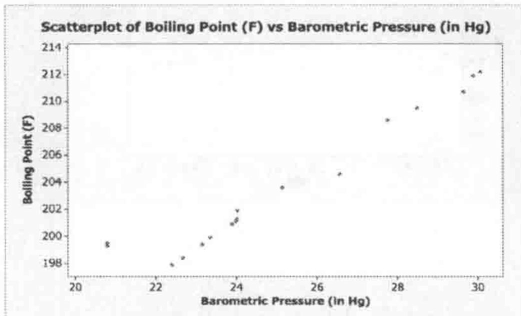
b. $F = .32529..00225 = 144.58$ with $p = 0.000$. $R^2 = 96\%$. There is a linear association between viscosity and temperature.

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.

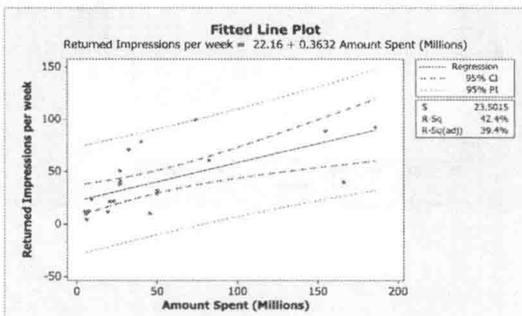
2.17 $\widehat{y} = 163.931 + 1.5796x$, $F = 226.4$ with $p = 0.000$, $R^2 = 93.8\%$. The model is a good fit of the data.




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

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.

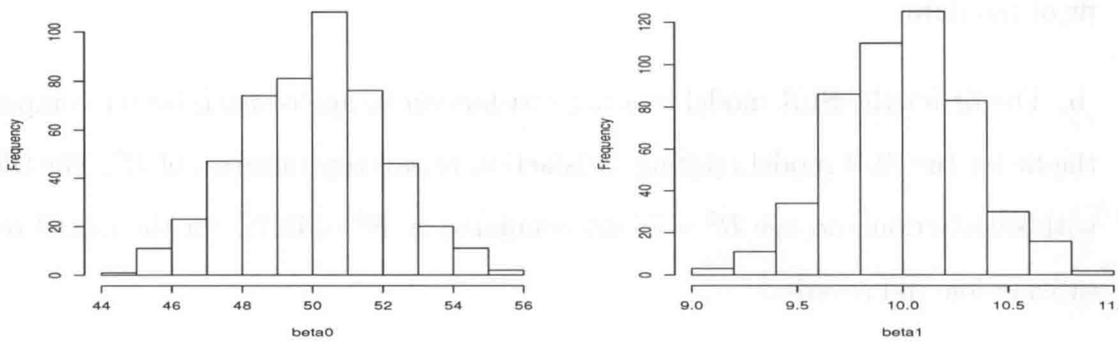
b. The fit for the SLR model relating satisfaction to age is much better compared to the fit for the SLR model relating satisfaction to severity in terms of $R^2$. For the SLR with satisfaction and age $R^2 = 75.8\%$ compared to $R^2 = 42.7\%$ for the model relating satisfaction and severity.

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.

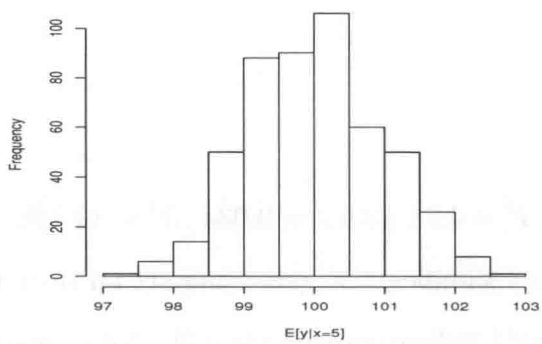
2.22 $\widehat{y} = 21.25 + 7.80x$, $F = 0.22$ with $p = 0.648$ ,$R^2 = 1.3\%$. The chemist's belief is incorrect. There is no relationship between the ratio of inlet oxygen to inlet methanol and percent conversion ($p - value = 0.648$). The $R^2 = 1.3\%$, which indicates that the ratio explains virtually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

2.23 a.




Both histograms are bell-shaped. The one for $\beta_0$ is centered around 50 and the one for $\beta_1$ 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.

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

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,

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

b. The bias is

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

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