12 Optional Sections: Multiple Linear Regression

12.6 The Polynomial and Qualitative Predictor Models

The multiple linear regression model, given in Equation 12.13, suggests that all of the predictor variables, $x_1, x_2, \ldots, x_k$, are different. However, the more general linear regression model allows for polynomial equations, qualitative variables, and even interaction terms. The purpose of this section is to extend the multiple linear regression model to include some of these other, more general models.

Can you write a cubic polynomial regression model? How about a fifth-degree polynomial model?

A polynomial model contains quadratic or higher-degree terms. For example, the following is a (quadratic) polynomial (linear) regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i \tag{12.14}$$

There is only one independent variable in this model example, so there is no need for a double subscript on x.

Remember, this is still a *linear* regression model because the expression on the right in Equation 12.14 is a linear combination of the regression coefficients.

A CLOSER LOOK

1. Polynomial regression models may contain more than one predictor variable. The highest power, or degree, of each predictor variable included in the model may be different, and intermediate-degree terms may be omitted. For example, here is a polynomial regression model with two predictor variables:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \beta_4 x_{2i}^3 + E_i \tag{12.15}$$

   Notice that the highest degree on one predictor is 2, and on the other, 3. And the *squared* term associated with the second predictor is omitted.

2. There is a danger in using *any* regression model to predict a mean value or an observed value outside the range of data (used to obtain the estimated regression coefficients). This is especially true for polynomial regression. If the values of the dependent variable and the predictor variable lie along a curve, the relationship may vary considerably outside the range of data.

3. A predictor variable in a polynomial regression model with a degree higher than 3 should be used rarely. It is difficult to interpret the regression coefficients in terms of degree 4 or higher.

4. Suppose there are n observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We can always find an estimated polynomial regression, with one predictor variable, of degree $n - 1$, that will *pass through*, or contain, all n observations. However, it is very unlikely that this type of a model will convey the true relationship between the predictor and the dependent variables. For example, Figure 12.94 shows a polynomial of degree 5 passing through six points.

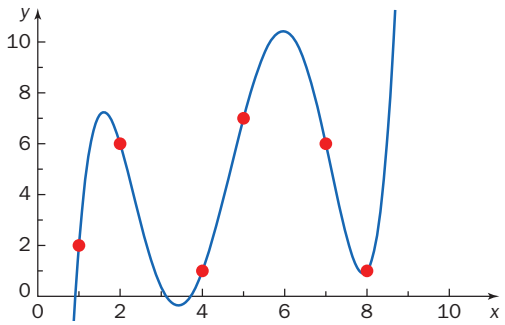


Figure 12.94 An illustration of an exact polynomial fit.

The curve fits the data exactly. However, this type of complex model is rare; it would be difficult to find a practical example and justification for a fifth-degree polynomial model.

5. For a set of n observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, there may be many justifiable mathematical models that could be used to characterize the relationship between the independent and dependent variables. However, often the best model is based in reality and tangible, physical phenomena.

All of the results presented earlier in this chapter—how to find the estimated regression coefficients, inference procedures, and regression diagnostics—apply to polynomial regression models. Technology should be used to compute the estimates for the true regression parameters.

Example 12.15 Cerebral Blood Flow

DATA SET
CEREBRAL

A recent medical study suggested that the cerebral blood flow velocity in certain patients is affected by age.[42] A random sample of patients was obtained and the middle cerebral artery maximum flow velocity (MFV, in cm/sec) and the age (in years) was measured for each. The data are given in the following table.

MFV	Age	MFV	Age	MFV	Age	MFV	Age
84.84	41	59.57	78	76.63	31	74.46	61
51.34	74	71.60	28	78.50	42	68.97	25
67.86	58	75.13	54	78.77	29	72.63	39
83.42	37	74.82	44	55.45	72	54.64	77
67.78	52	79.54	42	77.33	25	82.32	45

a. Construct a scatter plot and consider a quadratic model. Find the estimated regression equation.

b. Verify that the regression is significant at the $\alpha = 0.01$ level.

c. Conduct separate hypothesis tests to determine whether the linear and/or quadratic terms contribute to the overall significant regression. Use $\alpha = 0.05$ in each test.

SOLUTION

Solution Trail 12.15a

KEYWORDS
- Quadratic model
- Estimated regression equation
- Random sample

TRANSLATION
- Polynomial regression

CONCEPTS
- Least-squares estimates

VISION
Compute age^2 for each patient. Use technology to find the estimates of the regression coefficients.

a. MFV (y) is the dependent variable and age (x) is the independent variable. Figure 12.95 shows a scatter plot of MFV versus age. This plot suggests that a quadratic model might be appropriate. The points tend to fall along a parabola, concave down.

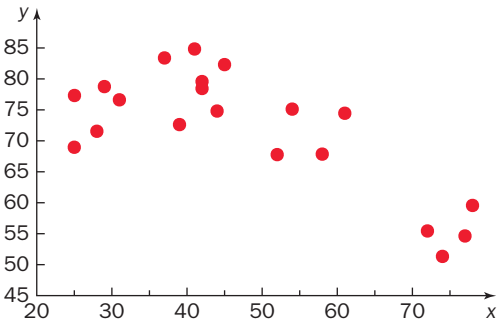
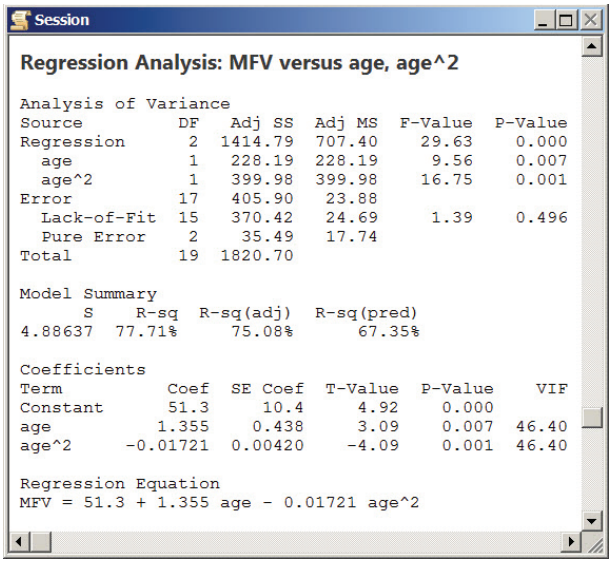


Figure 12.95 Scatter plot of MFV versus age.

The quadratic model is

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$$

Use technology to find the estimated regression coefficients. The results from the Minitab regression analysis are shown in Figure 12.96.



Session

Regression Analysis: MFV versus age, age^2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	1414.79	707.40	29.63	0.000
age	1	228.19	228.19	9.56	0.007
age^2	1	399.98	399.98	16.75	0.001
Error	17	405.90	23.88		
Lack-of-Fit	15	370.42	24.69	1.39	0.496
Pure Error	2	35.49	17.74		
Total	19	1820.70			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.88637	77.71%	75.08%	67.35%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	51.3	10.4	4.92	0.000	
age	1.355	0.438	3.09	0.007	46.40
age^2	-0.01721	0.00420	-4.09	0.001	46.40

Regression Equation

MFV = 51.3 + 1.355 age - 0.01721 age^2

Figure 12.96 Minitab regression analysis.

The estimated regression equation is $y = 51.27 + 1.3548x - 0.017209x^2$. Here is the ANOVA table.

ANOVA summary table for multiple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	F	p Value
Regression	1414.79	2	707.40	29.63	0.0000028
Error	405.90	17	23.88		
Total	1820.69	19			

The coefficient of determination is $r^2 = 0.777$. The model can be used to explain approximately 78% of the variation in the dependent variable.

b. There are $k = 2$ predictor variables (x and x^2), and $n = 20$ observations. The F test for a significant regression with $\alpha = 0.01$ is

H_0: $\beta_1 = \beta_2 = 0$

H_a: $\beta_i \neq 0$ for at least one i

TS: $F = \text{MSR/MSE}$

RR: $F \geq F_{\alpha,k,n-k-1} = F_{0.01,2,17} = 6.11$

Using the ANOVA table, the value of the test statistic is

$f = \text{MSR/MSE} = 29.63\ (\geq 6.11)$

Because f lies in the rejection region (or, equivalently, $p = 0.0000028 \leq 0.05$), there is evidence to suggest that at least one of the regression coefficients is different from 0.

c. To test whether the linear term, x, is a significant predictor, conduct a hypothesis test concerning the regression coefficient β_1.

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

TS: $T = \dfrac{B_1 - 0}{S_{B_1}}$

RR: $|T| \geq t_{\alpha/2,n-k-1} = t_{0.025,17} = 2.1098$

Solution Trail 12.15b

KEYWORDS

- Regression is significant

TRANSLATION

- F test for a significant regression

CONCEPTS

- Hypothesis test for a significant regression

VISION

Use the template for a hypothesis test for a significant regression.

Remember that k is the number of predictor variables, or terms, in the model.

Solution Trail 12.15c

KEYWORDS

- Determine whether the linear and/or quadratic terms contribute

TRANSLATION

- Separate hypothesis test concerning β_1 and β_2

CONCEPTS

- Hypothesis tests concerning β_1 and β_2

VISION

Test the hypotheses H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$.

Using the Minitab output,

$$t = \frac{\hat{\beta}_1 - 0}{s_{B_1}} = 3.09$$

The value of the test statistic lies in the rejection region, $|t| = |3.09| = 3.09 \geq 2.1098$ (or, equivalently, $p = 0.007 \leq 0.05$). There is evidence to suggest that the regression coefficient on the linear term, β_1, is different from 0.

Conduct a similar hypothesis test concerning the regression coefficient β_2.

H_0: $\beta_2 = 0$

H_a: $\beta_2 \neq 0$

TS: $T = \dfrac{B_2 - 0}{S_{B_2}}$

RR: $|T| \geq t_{\alpha/2,n-k-1} = t_{0.025,17} = 2.1098$

Using the Minitab output, the value of the test statistic is

$$t = \frac{\hat{\beta}_2 - 0}{s_{B_2}} = -4.09$$

The value of the test statistic lies in the rejection region, $|t| = |-4.09| = 4.09 \geq 2.1098$ (or, equivalently, $p = 0.001 \leq 0.05$). There is evidence to suggest that the regression coefficient on the quadratic term, β_2, is different from 0.

Figure 12.97 shows a graph of the estimated regression equation and a scatter plot of the data.

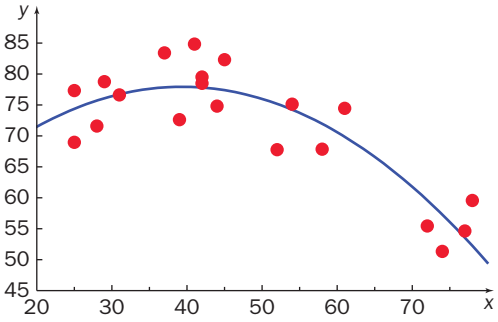


Figure 12.97 Scatter plot of MFV versus age and a graph of the estimated regression equation.

Note that in order to check the regression assumptions, we might consider a normal probability plot of the residuals along with the usual residual plots. ■

Example 12.16 Citrus Pests

DATA SET
CITRUS

The Asian citrus psyllid, *Diaphorian citri*, is present in southern Asia and other citrus-growing regions, even Florida and California. This citrus pest produces a toxin that affects plant tips and normal leaf growth. A recent study suggests that the number of eggs laid by female *Diaphorian citri* is affected by temperature.[43] Suppose a random sample of 48-hour periods in Florida was obtained. The temperature at 24 hours (x, in °C) and the number of eggs on certain leaves (y) was measured for each time period. The data are given in the following table.

x	17	20	23	26	29	32	35	37	41	21	25	24	29	38	30	34
y	7	12	41	70	72	85	67	30	4	33	50	60	85	60	55	43

Figure 12.98 shows a scatter plot of the data. This graph suggests a quadratic model of the form $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$

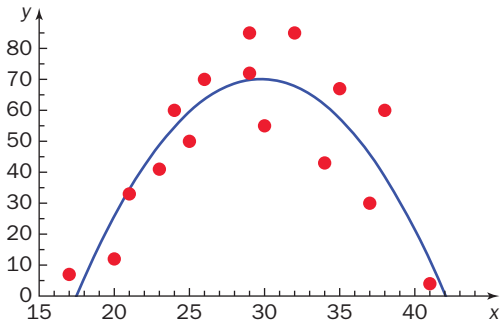


Figure 12.98 Scatter plot and graph of the estimated regression equation for the citrus pest data.

a. Construct a 95% confidence interval for the mean number of eggs when the temperature is 27°C. Use this confidence interval to determine whether there is any evidence to suggest that the mean number of eggs for a temperature of 27 is different from 80.

b. Find a 95% prediction interval for an observed value of the number of eggs when the temperature is 27°C.

SOLUTION

a. Recall that a confidence interval concerning the mean value of Y for $x = x^*$ and a prediction interval for an observed value of Y when $x = x^*$ are based on the t distribution. Use technology to find the estimated regressions coefficients, the confidence interval, and the prediction interval. The results from the Minitab regression analysis are shown in Figure 12.99.

Session

Regression Analysis: Eggs versus Temp, Temp^2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	7492.7	3746.37	19.17	0.000
Temp	1	7426.3	7426.33	37.99	0.000
Temp^2	1	7168.3	7168.31	36.67	0.000
Error	13	2541.0	195.46		
Lack-of-Fit	12	2456.5	204.71	2.42	0.467
Pure Error	1	84.5	84.50		
Total	15	10033.8			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
13.9808	74.68%	70.78%	57.31%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-341.2	63.0	-5.41	0.000	
Temp	27.63	4.48	6.16	0.000	75.51
Temp^2	-0.4641	0.0766	-6.06	0.000	75.51

Regression Equation

Eggs = -341.2 + 27.63 Temp - 0.4641 Temp^2

Variable	Setting
Temp	27
Temp^2	729

Fit	SE Fit	95% CI	95% PI
66.5495	4.83023	(56.1144, 76.9845)	(34.5940, 98.5050)

Figure 12.99 Minitab regression analysis.

Challenge: Use the estimated regression equation to find the temperature that produces the maximum number of eggs.

The estimated regression equation is $y = -341.2 + 27.63x - 0.4641x^2$. Note that the coefficient of determination is $r^2 = 0.7468$; Figure 12.98 also shows a graph of the regression equation.

Using the Minitab output, a 95% confidence interval for the true mean number of eggs when the temperature is 27°C is (56.11, 76.98). Because 80 is not included in this interval, there is evidence to suggest that the mean number of eggs is different from 80.

b. Using the Minitab output, a 95% prediction interval for a single observation of the number of eggs when $x = 27$ (and $x^2 = 729$) is (34.59, 98.51). Notice that this prediction interval is wider that the corresponding confidence interval.

A CLOSER LOOK

1. If the effect of a predictor variable depends on the value of a second predictor variable, then the model is not additive. In this case, an *interaction*, or *linear-by-linear*, or *bilinear* term may be appropriate. For example, the following regression model has two predictor variables and an interaction term:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + E_i \tag{12.16}$$

If an interaction term is included, the regression coefficients have a slightly different meaning. In a simple linear regression model, β_1 is the change in the response variable if the first predictor variable increases by 1 (with the second predictor variable held constant). However, in Equation 12.16, if x_1 increases by 1 and x_2 is held constant, the response variable changes by $\beta_1 + \beta_3 x_2$.

Suppose x_2 increases by 1 and x_1 is held constant. How does the response variable change?

2. Just a reminder: A regression model may contain several of the components presented in this and earlier sections. For example, the following model includes a cubic term and an interaction term:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^3 + \beta_3 x_{2i} + \beta_4 x_{1i} x_{2i} + E_i \tag{12.17}$$

Suppose a response variable is affected by a single predictor variable and a categorical variable, for example sex, or age group, or insurance risk (low, medium, or high). A qualitative, or indicator, variable may be included in the model to help explain the variation in the response variable. For example, consider a model used to predict the maximum wind speed of a hurricane (Y, in mph) based on the diameter of the eye of the hurricane (x, in miles). Suppose the maximum wind speed is affected by the presence (or absence) of El Niño. Let x_2 be an indicator variable with value 0 if El Niño is absent and value 1 if present. The regression model is

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i \tag{12.18}$$

Notice that if El Niño is absent, $x_2 = 0$ and the true regression equation becomes

$$y = \beta_0 + \beta_1 x_1 \tag{12.19}$$

If El Niño is present, $x_2 = 1$ and the true regression equation is

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 \tag{12.20}$$

The regression coefficient represents the effect due to the presence of El Niño. The graph of each of these equations is a straight line. Both lines have the same slope (a different y intercept) and are therefore parallel.

Suppose there are c categories to account for in a regression model. The model is adjusted by adding $c - 1$ indicator variables. For example, in the hurricane maximum wind speed example, suppose the diameter of the eye and sunspot activity over the last year (low, medium, high) are the predictors. Sunspot activity is a qualitative variable

with $c = 3$ classes. Therefore, $c - 1 = 2$ indicator variables are necessary, and are defined as

$$x_2 = \begin{cases} 1 & \text{if low sunspot activity} \\ 0 & \text{otherwise} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{if medium sunspot activity} \\ 0 & \text{otherwise} \end{cases} \tag{12.21}$$

Notice that $x_2 = 0$ and $x_3 = 0$ when the sunspot activity is high. The regression model becomes

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + E_i \tag{12.22}$$

Example 12.17 Sleep Duration and Body Mass

DATA SET

SLEEP

Research studies suggest that sleep duration is associated with body mass index (BMI). A random sample of men and women from a sleep cohort study was obtained. The BMI (y, in kg/m^2) and daily sleep duration (x_1, in hours) was recorded for each person. The data are given in the following table.

y	x_1	x_2	y	x_1	x_2
26.8	6.3	0	29.1	6.6	1
24.7	7.1	0	25.3	7.0	0
31.4	4.6	0	33.7	5.2	1
30.8	5.3	0	32.8	5.3	1
31.6	6.3	1	27.5	6.5	0
25.9	7.6	1	28.5	6.3	0
31.2	5.6	1	24.3	7.8	0
27.2	7.2	1	28.8	6.3	0
29.6	5.7	0	28.8	5.8	0
24.2	8.0	1	31.9	5.0	1
26.5	7.6	1	27.3	7.4	1
32.4	5.4	1	34.4	4.7	1
33.7	4.9	1	33.0	5.7	1
34.3	4.8	1	31.8	5.8	1
28.9	6.9	1	28.7	6.8	1

There is additional evidence to suggest that the relationship between BMI and sleep duration is different for men and women, shifted up for men.

a. Construct a scatter plot of the data (BMI versus sleep duration). Describe the relationship between sleep duration and BMI.

b. Let x_2 be an indicator variable, 0 for a woman and 1 for a man. Consider the following regression model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i$$

Find the estimated regression equation.

c. Verify that the regression is significant at the $\alpha = 0.05$ level. Interpret the estimate of the regression coefficient β_2.

SOLUTION

a. Plot all of the data on the same axes with a different plot symbol for women and men. See Figure 12.100.

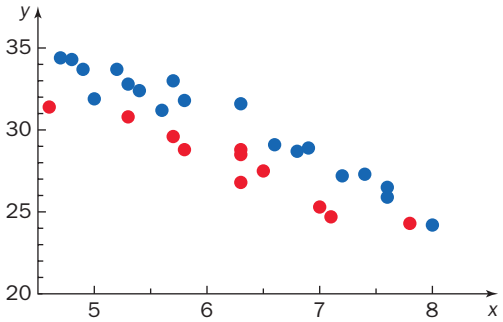


Figure 12.100 Scatter plot of BMI (y) versus sleep duration (x_1). Red points correspond to observations for women; blue points correspond to observations for men.

The scatter plot suggests that the relationship between BMI and sleep duration is linear. As sleep duration increases, BMI tends to decrease. The scatter plot also suggests that the linear relationship is shifted up for men.

Solution Trail 12.17b

KEYWORDS
- Indicator variable
- Estimated regression equation

TRANSLATION
- Multiple linear regression

CONCEPTS
- Least-squares estimates

VISION
Use technology to find the estimates of the regression coefficients.

b. Use technology to find the estimated regression coefficients. The results from the JMP Fit Least Squares command are shown in Figure 12.101.

Response BMI

Whole Model

Summary of Fit

RSquare	0.9434
RSquare Adj	0.939208
Root Mean Square Error	0.765954
Mean of Response	29.50333
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	264.02915	132.015	225.0175
Error	27	15.84052	0.587	Prob > F
C. Total	29	279.86967		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	44.956609	0.917629	48.99	<.0001*
Sleep	-2.736866	0.142198	-19.25	<.0001*
W/M	2.3205466	0.29053	7.99	<.0001*

Figure 12.101 JMP Fit Least Squares analysis.

The estimated regression equation is $y = 44.9566 - 2.7369x_1 + 2.3205x_2$. Here is the ANOVA table.

ANOVA summary table for multiple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	F	p Value
Regression	264.03	2	132.01	225.02	< 0.0001
Error	15.84	27	0.59		
Total	279.87	29			

The coefficient of determination is $r^2 = 0.943$. The model can be used to explain approximately 94% of the variation in the dependent variable.

Solution Trail 12.17c

KEYWORDS
- Regression is significant

TRANSLATION
- F test for a significant regression

CONCEPTS
- Hypothesis test for a significant regression

VISION
Use the template for a hypothesis test for a significant regression.

c. There are $k = 2$ predictor variables (x_1 and x_2), and $n = 30$ observations. The F test for a significant regression with $\alpha = 0.05$ is

H_0: $\beta_1 = \beta_2 = 0$

H_a: $\beta_i \neq 0$ for at least one i

TS: $F = \text{MSR/MSE}$

RR: $F \geq F_{\alpha,k,n-k-1} = F_{0.05,2,27} = 3.35$

Using the ANOVA table, the value of the test statistic is

$f = \text{MSR/MSE} = 225.02\ (\geq 3.35)$

Because f lies in the rejection region (or, equivalently, $p \leq 0.05$), there is evidence to suggest that at least one of the regression coefficients is different from 0. The JMP output suggests that both regression coefficients are significant at the $p < 0.001$ level.

The estimated regression coefficient, $\hat{\beta}_2 = 2.3205$, suggests that the regression equation for men is shifted vertically 2.3205 kg/m^2. ■

Technology Corner

Procedure: Multiple linear regression, polynomial and qualitative predictor models.
Reconsider: Example 12.17, solution, and interpretations.

CrunchIt!

Use the Multiple Linear regression command to find the estimated regression line, conduct hypothesis tests concerning the coefficients, and construct a plot of the residuals versus the dependent variable and a normal probability plot of the residuals.

1. Enter the data into columns Var1, Var2, and Var3. Rename the columns if desired.
2. Select Statistics; Regression; Multiple Linear. Choose the Dependent Variable from the pull-down menu and check the Independent variables. Select Numeric results from the Display pull-down menu to view the estimated regression line and hypothesis tests concerning the coefficients. The CI Level box is used to obtain a confidence interval for each coefficient. Click Calculate to display the results. See Figure 12.102.

Results - Multiple Linear Regression

Export

Fitted Equation:	BMI = 44.96 - 2.737 * Sleep + 2.321 * Gender				
	Estimate	Std. Error	t value	Pr(>|t|)	CI
(Intercept)	44.96	0.9176	48.99	<0.0001	(43.07, 46.84)
Sleep	-2.737	0.1422	-19.25	<0.0001	(-3.029, -2.445)
Gender	2.321	0.2905	7.987	<0.0001	(1.724, 2.917)

r-Squared:	0.9434
Adjusted r-Squared:	0.9392
estimated sigma:	0.7660

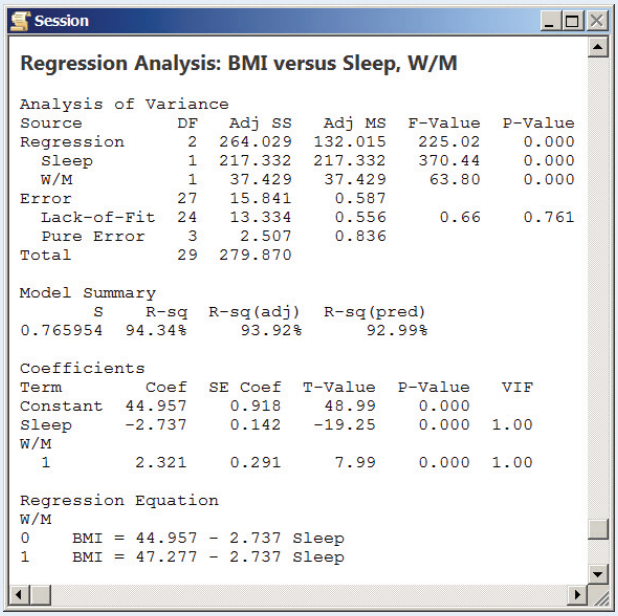
Figure 12.102 CrunchIt! Multiple Linear Regression summary output.

Minitab

There are several ways to conduct a multiple linear regression analysis in Minitab, in a session window and in the Stat; Regression menu.

1. Enter values for the dependent variable into column C1, and values for the predictor variables into columns C2 and C3. Name the columns if desired.

2. Select Stat; Regression; Regression; Fit Regression Model. Enter the Response column, C1, the Continuous predictors, C2, and the Categorical predictors, C3. Use the Storage option to save the residuals in a Worksheet column. Click OK to display the regression analysis. See Figure 12.103.



Session

Regression Analysis: BMI versus Sleep, W/M

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	264.029	132.015	225.02	0.000
Sleep	1	217.332	217.332	370.44	0.000
W/M	1	37.429	37.429	63.80	0.000
Error	27	15.841	0.587		
Lack-of-Fit	24	13.334	0.556	0.66	0.761
Pure Error	3	2.507	0.836		
Total	29	279.870			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.765954	94.34%	93.92%	92.99%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	44.957	0.918	48.99	0.000	
Sleep	-2.737	0.142	-19.25	0.000	1.00
W/M					
1	2.321	0.291	7.99	0.000	1.00

Regression Equation

W/M

0 BMI = 44.957 - 2.737 Sleep

1 BMI = 47.277 - 2.737 Sleep

Figure 12.103 Minitab regression analysis.

Excel

Regression analysis is included in the Data Analysis tool pack.

1. Enter the dependent variable into column A, and values for the predictor variables into columns B and C.
2. Under the Data tab, select Data Analysis; Regression. Enter the Y Range, X Range, and the Output Range. Check Residuals Plot to construct a plot of the residuals versus each predictor variable. Check Residuals to save the residuals. The summary output is shown in Figure 12.104.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9713
R Square	0.9434
Adjusted R Square	0.9392
Standard Error	0.7660
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	264.0292	132.0146	225.02	0.0000
Residual	27	15.8405	0.5867		
Total	29	279.8697			

	Coefficients	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	44.9566	0.9176	48.9921	0.0000	43.0738	46.8394	43.0738	46.8394
Sleep	-2.7369	0.1422	-19.2468	0.0000	-3.0286	-2.4451	-3.0286	-2.4451
Gender	2.3205	0.2905	7.9873	0.0000	1.7244	2.9167	1.7244	2.9167

Figure 12.104 Excel regression analysis summary output.

SECTION 12.6 EXERCISES

Concept Check

12.174 True/False A polynomial regression model may contain more than one predictor variable.

12.175 True/False In a polynomial regression model, the highest degree on a predictor variable is 3.

12.176 True/False Any multiple linear regression model can be used to predict values outside the range of data.

12.177 True/False For any set of n observations, we can always find an estimated polynomial regression that will pass through all n observations.

12.178 True/False All predictor variables in a multiple linear regression model must be continuous.

12.179 Fill in the Blank If the effect of a predictor variable depends on the value of a second predictor variable, then the model should include a(n) _____ term.

12.180 Fill in the Blank If there are c categories to account for in a regression model, then there are _____ indicator variables.

Practice

12.181 Write the regression model that includes predictor variables as described in each situation.

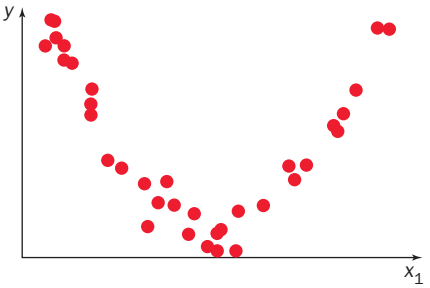
a. Cubic in x_1 and an indicator variable x_2.
b. Quadratic in x_1 and quadratic in x_2.
c. A fourth-degree polynomial in x_1 and an interaction term between x_1 and x_2.
d. Linear in x_1 and the appropriate indicator variables to delineate a health assessment (excellent, good, fair, poor).

12.182 Write the regression model that includes predictor variables as described in each situation.

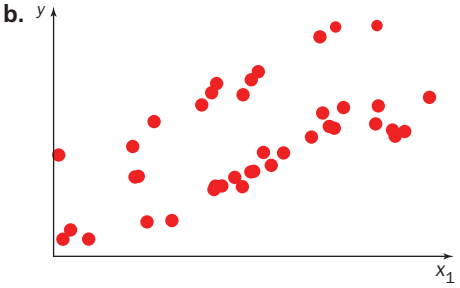
a. Linear in x_1, x_2, and x_3, and three interaction terms.
b. A fourth-degree polynomial in x_1, and three additional predictors, x_2, x_3, and x_4.
c. Quadratic in x_1 and the appropriate indicator variables to differentiate between the categories of an investment risk (high, medium, low).
d. Cubic in x_1, cubic in x_2, and an interaction term.

12.183 Write a regression model that could be used to describe the relationship suggested in each scatter plot.

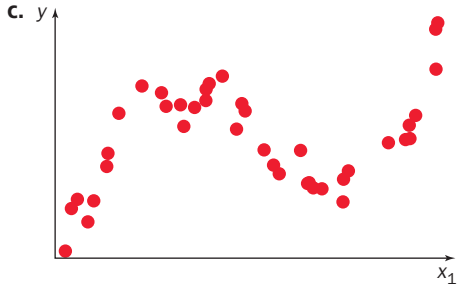
a.



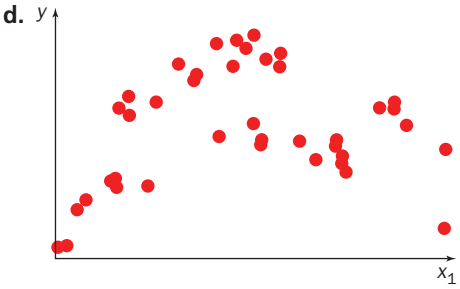
b.



c.



d.



12.184 Suppose the true regression equation relating the variables x_1 and y for values of x_1 between 10 and 30 is $y = 179.7 - 17.2x_1 + 0.42x_1^2$.

a. Find the expected value of Y when $x_1 = 21$.
b. Estimate the minimum value of Y.
c. Explain the relationship between x_1 and y as the values of x_1 increase from 25 to 30.
d. Suppose $\sigma = 2.2$. Find the probability an observed value of Y is greater than 18 when $x_1 = 15$.

12.185 Suppose the true regression equation relating the variables x_1, x_2, x_3 and y is $y = 7.8 + 4.2x_1 + 5.7x_2 - 3.6x_3$, where x_1 varies between 0 and 20, and x_2 and x_3 are indicator variables defined by

Category	x_2	x_3
1	1	0
2	0	1
3	0	0

a. Find the expected value of Y when $x_1 = 6.2$ in category 3.
b. Find the expected value of Y when $x_1 = 17$ in category 1.
c. For any value of x_1 in the interval 0 to 20, which category has the largest expected value of Y? Justify your answer.
d. How much change in the dependent variable is expected when x_1 decreases by 3 units in category 1? Category 2?

e. Suppose $\sigma = 7.8$. Find the probability that an observed value of Y is less than 30 when $x_1 = 6$ in category 1.

12.186 An experiment resulted in the following observations on a dependent variable and a single independent variable. EX12.186

x	y	x	y
99.8	90.4	79.1	150.3
90.1	146.2	58.0	92.0
60.7	100.5	77.6	171.9
51.2	53.8	50.1	54.2
88.3	147.6	70.4	146.5
91.2	137.0	73.3	152.3

a. Construct a scatter plot of the data. Construct an appropriate model and estimate the regression coefficients.
b. Estimate the true mean value of y when $x = 65$.
c. Find the residuals. Construct a normal probability plot for the residuals. Is there any evidence to suggest that the random errors are not normal? Justify your answer.
d. Carefully sketch a graph of the residuals versus the predictor variable, x. Does this graph suggest any evidence of a violation of the regression assumptions? Justify your answer.

12.187 An experiment resulted in observations on a single dependent variable and one independent variable for three categories. EX12.187

a. Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + E_i$, where x_2 and x_3 are indicator variables. Find the coefficient of determination.
b. Conduct an F test for a significant regression. Use $\alpha = 0.01$. Use technology to find an exact p value.
c. Conduct the appropriate hypothesis tests to determine which predictor variables are significant.
d. Explain the meaning of the estimated regression coefficient $\hat{\beta}_1$. Use the estimated regression coefficients $\hat{\beta}_2$ and $\hat{\beta}_3$ to explain the effect of each category on the dependent variable y.

12.188 An experiment resulted in observations on a single dependent variable and two independent variables. EX12.188

a. Estimate the regression coefficients in the model $Y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + E_i$. Conduct an F test for a significant regression. Use $\alpha = 0.05$. Is there any evidence to suggest that x_1 and/or x_2 can be used to explain the variation in the dependent variable? Justify your answer.
b. Estimate the regression coefficients in the model $Y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i$. Conduct an F test for a significant regression. Use $\alpha = 0.05$. Is there any evidence to suggest that x_1 and/or x_2 can be used to explain the variation in the dependent variable? Justify your answer.
c. Using the second model, find a 95% confidence interval for the true mean value of Y when $x = (3.05, 15.75)$.

Applications

12.189 Medicine and Clinical Studies Emergency medical technicians (EMTs) often suffer from acute stress due to the nature of their jobs. They routinely make split-second medical decisions on patients involved in traumatic accidents. There is some evidence to suggest that experience is related to stress, and that the relationship between these two variables is not necessarily linear.[44] A random sample of EMTs was obtained and the number of years experience (x) and systolic blood pressure (y, in mmHg), an indicator of stress, were measured for each. The data are given in the following table. EMT

x	y	x	y	x	y
5.3	111	20.0	163	4.1	134
7.5	108	8.2	111	19.8	168
7.6	119	14.7	110	3.0	146
8.3	130	15.3	110	18.0	145
2.8	151	6.7	122	3.3	133
4.2	124	11.8	127	2.6	142
9.7	131	3.1	152		

a. Construct a scatter plot of the data. Use this plot to explain the relationship between stress and years of experience. Write an appropriate regression model.
b. Find the estimated regression equation.
c. Conduct an F test for a significant regression with $\alpha = 0.01$.
d. Find an estimate of the true mean systolic blood pressure for $x = 11$ years of experience.

12.190 Public Health and Nutrition There is some evidence to suggest that the amount of fluoride in drinking water may affect a child's IQ level.[45] Although fluoride is recommended to strengthen teeth and fight tooth decay, there is evidence to suggest that high levels can be toxic. Suppose a random sample of children in the United States aged 13–18 years was obtained. The fluoride level (x, in mg/L) in the water and the IQ was measured for each. FLUOR

a. Construct a scatter plot of the data. Write an appropriate regression model.
b. Find the estimated regression equation.
c. Does this model explain a significant amount of the variation in IQ? Conduct the appropriate model utility test using $\alpha = 0.05$.
d. Find an estimate of an observed value of IQ for $x = 4.0$ mg/L of fluoride, the U.S. Environmental Protection Agency's standard for the maximum amount of fluoride allowed in drinking water.

12.191 Public Health and Nutrition There is evidence to suggest that diet soda is linked to heart risks.[46] One theory suggests that those people who drink diet soda crave sweets; eat a lot of pastry, candy, and desserts; and, therefore, increase their risk of heart disease. A random sample of middle-aged adults was obtained. The amount of diet soda per day (x, in ounces) and the Heart Risk Assessment Number

(HRAN, a unitless number between 0 and 100; higher numbers indicate increased risk) were recorded for each. The data obtained were used to fit the multiple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$. Minitab was used to estimate the regression coefficients, and the output is shown below. SODA

a. Is the overall regression significant? Conduct the appropriate hypothesis test and justify your answer.

b. Conduct two hypothesis tests with H_0: $\beta_i = 0$, for $i = 1, 2$ and $\alpha = 0.05$. Which regression coefficients are significantly different from 0?

c. Using the results from part (b), suggest a different regression model.

Session

Regression Analysis: HRAN versus Ounces, Ounces^2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	9006.0	4503.02	40.53	0.000
Error	37	4111.1	111.11		
Total	39	13117.1			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
10.5409	68.66%	66.96%	65.02%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	12.91	6.38	2.02	0.050	
Ounces	-1.512	0.674	-2.24	0.031	15.80
Ounces^2	0.0710	0.0163	4.37	0.000	15.80

Regression Equation

HRAN = 12.91 - 1.512 Ounces + 0.0710 Ounces^2

12.192 Biology and Environmental Science The Jackson ratio is used to determine whether certain tortoises are ready for hibernation. It is computed by dividing weight by carapace length cubed. A researcher believes that the Jackson ratio can be used to predict the length of hibernation. A random sample of male tortoises ready for hibernation was obtained. The Jackson ratio (x) and the length of hibernation (y, in days) was recorded for each. JRATIO

a. Construct a scatter plot of the data. Describe the relationship.

b. Consider an appropriate regression model and find the estimated regression coefficients.

c. Find a 95% confidence interval for the true mean hibernation time when the Jackson ratio is 0.150.

d. Carefully sketch a normal probability plot for the residuals. Is there any evidence to suggest that the random error terms are not normal? Justify your answer.

12.193 Medicine and Clinical Studies Researchers have concluded that nicotine adversely affects certain regions of the brain that control emotions, REM sleep, and seizures.[47] Suppose an experiment was conducted to examine the relationship between the dose of nicotine per day (x, in mg/kg of body weight) and the volume of a certain area of the brain (y, in mm^3) in rats. Animals were studied over a five-day period, and the data were recorded. BRAIN

a. Consider a quadratic regression model for this data. Find the estimated regression line.

b. Conduct an F test for a significant regression with $\alpha = 0.05$.

c. Find a 95% prediction interval for an observed value of Y when $x = 12$.

d. Estimate the nicotine dose that would result in complete degeneration of this part of the brain.

12.194 Public Health and Nutrition Yerba Maté tea is a popular herbal drink because of its alleged health benefits. This tea reportedly contains several antioxidants, and is even being sold as a bottled energy drink. A study was conducted to investigate the relationship between the steeping time of Yerba Maté tea and the amount of manganese, an essential element that helps brain functions and is also used in the production of certain enzymes. A random sample of Yerba Maté teabags was obtained, and eight ounces of boiling water was used to brew each cup of tea. The brewing times (x, in minutes) were randomly selected and the amount of manganese (y, in mg) was measured in each drink. YERBA

a. Construct a scatter plot for these data and write an appropriate regression model.

b. Find the estimated regression coefficients.

c. Conduct an F test for a significant regression with $\alpha = 0.01$.

d. Estimate the optimum time to brew one cup of Yerba Maté, i.e., the time that yields the highest amount of manganese.

12.195 Travel and Transportation One of the main reasons an aircraft may roll off the end of a runway or slide laterally is the accumulation of rubber (from tires) on the runway surface. The U.S. Federal Aviation Administration has issued recommendations for minimum friction testing frequency and has assigned friction-level classifications that specify planning and action levels. A study was conducted to investigate the relationship between runway friction (y), temperature (x_1, in degrees Fahrenheit), and relative humidity (x_2, a percentage). A random sample of runways, days, and times was selected, and the data were recorded. Consider the regression model, which includes an interaction term: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + E_i$. RUNWAY

a. Find the estimated regression coefficients. Conduct a model utility test with $\alpha = 0.01$. Find r^2 and interpret this value.

b. Which regression coefficients are significantly different from 0? Justify your answer.

c. Based on your answer to part (b), can you suggest a different model? Find the estimated regression coefficients for this (improved) model, conduct a model utility test, and find r^2. Which model is better for predicting runway friction? Why?

12.196 Psychology and Human Behavior A study was conducted to examine the relationship between involvement in sports and depression levels among adolescents. A random sample of teenagers was selected. Sports involvement was measured in hours per week (x_1), and depression was measured using the Center for Epidemiological Studies Depression Scale (CES-D) (y). The scores on the CES-D range from 0 to 60, with higher scores indicating depression. The following regression model was considered: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i$, where x_2 is an indicator variable (0 for female, 1 for male). Minitab was

used to estimate the regression coefficients, and part of the output is shown below. SPORTS

```
Session
Regression Analysis: CES-D versus Hrs/wk, Gender

Analysis of Variance
Source          DF   Adj SS   Adj MS  F-Value  P-Value
Regression
Error           47  1450.20
Total           49  3478.50

Model Summary
      S    R-sq  R-sq(adj)  R-sq(pred)
5.55475  58.31%     56.54%      53.79%

Coefficients
Term         Coef  SE Coef  T-Value  P-Value
Constant    33.73     1.86    18.13    0.000
Hrs/wk     -1.967    0.247   -7.97    0.000
Gender       3.90     1.67     2.34    0.024

Regression Equation
CES-D = 33.73 - 1.967 Hrs/wk + 3.90 Gender
```

a. Complete the ANOVA table and conduct an F test for a significant regression. Use technology to find the exact p value.
b. Does sex affect the relationship between depression and sports involvement? Justify your answer.
c. Estimate the true mean depression score for a female who participates in sports for four hours per week.
d. Suppose a CES-D score of 20 or greater indicates depression. Use this model to predict the number of hours per week a male and a female should participate in sports in order to avoid depression.

12.197 Biology and Environmental Science In 1851, Lorenzo Lorraine Langstroth discovered the concept of "bee space," the distance between wax combs. Modern beekeepers believe there is an optimum distance between wax combs that promotes maximum honey production. A random sample of bee colonies was obtained, and the bee space (x, in cm) and the yearly honey production (y, in kg) were measured for each. BSPACE

a. Construct a scatter plot for the data and write an appropriate regression model.
b. Find the estimated regression coefficients.
c. Conduct an F test for a significant regression. Find the exact p value.
d. Remove the single outlier from the data set. Find the estimated regression coefficients for this reduced data set. Which set of regression coefficients do you believe is more appropriate? Justify your answer.

12.198 Biology and Environmental Science A researcher believes there is a linear relationship between the amount of phosphorus (x_1, in mg/L) and nitrogen (y, in mg/L) in freshwater lakes. In addition, this relationship may be affected by the annual rainfall (dry, normal, wet). A random sample of lakes in the United States was obtained, and a sample of water was obtained from each during the month of September. LAKES

a. Write the appropriate regression model and find the estimated regression coefficients.
b. Is the overall regression significant? Justify your answer.
c. Is there any evidence to suggest that the annual rainfall affects the relationship between phosphorus and nitrogen? Justify your answer.
d. Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_{1i} + E_i$. Find the estimated regression coefficients. Which model is better for predicting the amount of nitrogen in a lake, given the amount of phosphorus? Why?

12.199 Biology and Environmental Science A study was conducted to determine the effect of certain herbicides on weeds in farmland in Australia. A random sample of plots was obtained, and each was treated with the herbicide pendimethalin in various concentrations. Oat was planted in all plots using a low-soil-disturbance disc. Twenty-five days after sowing, soil samples were used to determine the length of each plant root (y, in cm) and the herbicide concentration in the soil (x, in μg/g). WEEDS

a. Construct a scatter plot of the data. Describe the relationship.
b. Consider a simple linear regression model for this data. Find the estimated regression line. Conduct an F test for a significant regression with $\alpha = 0.01$. Carefully sketch a normal probability plot of the residuals and a plot of the residuals versus x. Is there any evidence to suggest any violations of the regression assumptions? Justify your answer.
c. Consider a quadratic regression model for this data. Find the estimated regression line. Conduct an F test for a significant regression with $\alpha = 0.01$. Carefully sketch a normal probability plot of the residuals and a plot of the residuals versus x. Is there any evidence to suggest any violations of the regression assumptions? Justify your answer.
d. Which model, linear or quadratic, do you think is better, and why?

Extended Applications

12.200 Physical Sciences A researcher believes that many factors affect the magnitude of an earthquake, including the length of the fault line, the duration of the first shaking portion of the earthquake, and the rock formation in the area. Suppose a random sample of recent earthquakes in the United States was obtained. The magnitude (y, based on the Richter scale), fault length (x_1, in kilometers), duration (x_2, in seconds), and the rock formation (x_3, x_4, indicator variables) were recorded for each. The values of the indicator variables were defined as QUAKE

Rock formation	x_3	x_4
Igneous	1	0
Metaphoric	0	1
Sedimentary	1	1

Consider the regression model $Y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + E_i$.

a. Find the estimated regression coefficients. Conduct appropriate hypothesis tests to determine which regression coefficients are significantly different from 0.

b. Estimate the mean magnitude for $x_1 = 400$ and $x_2 = 45$ in an igneous rock formation.
c. Write a new regression model based on your results from part (a). Find the estimated regression coefficients in this model.
d. For this new model, estimate the mean magnitude for $x_1 = 400$ and $x_2 = 45$ in an igneous rock formation. Which estimate do you think is more accurate? Why?

12.201 Marketing and Consumer Behavior A Seattle pump station supervisor is trying to model the amount of water (y, in thousands of gallons/hour) as a function of hours after midnight (x). A random sample of times on various days was obtained, and the flow rate for each time was recorded. PUMP

a. Construct a scatter plot of the data. Explain the relationship between water flow rate and hours after midnight.
b. Consider the regression model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + E_i$. Find the estimated regression coefficients.
c. Construct the ANOVA table and conduct the model utility test. Use $\alpha = 0.01$.
d. Find the value of r^2 and interpret this value.

12.202 Biology and Environmental Science An experiment was conducted to study the relationship between the proportion of fresh grass in the diet of cows and the milk yield. A random sample of two types of dairy cows (Guernsey and Holstein-Friesian) in upstate New York was obtained. The proportion of fresh grass in the diet (x) and the amount of milk produced per week (y, in liters) was recorded for each cow.[48] COWS

a. Construct a scatter plot of the data, without regard to cow type. Describe the relationship between the proportion of fresh grass and milk yield.
b. Consider a simple linear regression model with an indicator variable (for cow type). Estimate the regression coefficients in this model.
c. Conduct an F test for a significant regression ($\alpha = 0.05$) and an appropriate test to determine whether the indicator variable is significant.
d. Consider a simple linear regression model without an indicator variable. Estimate the regression coefficient in this model. Which model is more appropriate? Why?

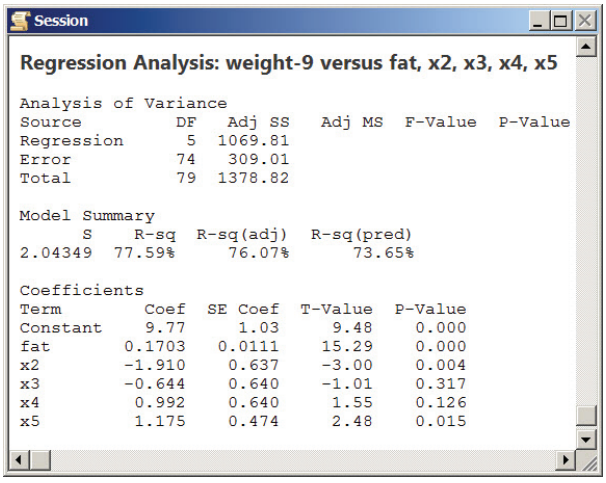
12.203 Biology and Environmental Science According to GlobalSecurity.org, the straits of Malacca, Sunda, and Lombok are heavily used by merchant fleets to transport raw materials, oil, and liquid natural gas. Environmentalists are concerned that the surface concentration of aluminum in these waterways is increasing, and may be affected by the water temperature. A random sample of days and waterways was selected. The surface water temperature (x, in degrees Fahrenheit) and the surface aluminum concentration (y, in μg/L) were measured. STRAITS

a. Construct a scatter plot of the data, by strait type. Describe the relationship.
b. Consider a simple linear regression model with the appropriate number of indicator variables to account for strait type. Estimate the regression coefficients in this model.
c. Complete the summary ANOVA table.
d. Conduct the model utility test. Use $\alpha = 0.01$.
e. Is there any indication that the linear relationship varies due to strait? Justify your answer.
f. Find an estimate for the true mean aluminum concentration in the strait of Sunda for a water temperature of 85°F.

12.204 Education and Child Development There is some evidence to suggest that the diet of a woman while pregnant may affect the weight, or tendency toward obesity, of her child. One theory is that if a woman's diet is high in fat, then the child may be destined for adult obesity.[49] A follow-up long-term study was conducted in which a random sample of 80 pregnant women from four races was obtained. The daily fat intake during pregnancy (x_1, g/day) was recorded and the weight of the child (y, in kg) was measured at age 9. The following simple linear regression model with indicator variables for race and for sex of the child was used to fit the data: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + E_i$. The values of the indicator variables were defined as follows: DIET

Race	x_2	x_3	x_4	Sex	x_5
White	0	0	0	Male	0
African American	1	0	0	Female	1
Hispanic	0	1	0		
Asian American	0	0	1		

Minitab was used to estimate the regression coefficients, and part of the output is shown below.


Session

Regression Analysis: weight-9 versus fat, x2, x3, x4, x5

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	1069.81			
Error	74	309.01			
Total	79	1378.82			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.04349	77.59%	76.07%	73.65%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	9.77	1.03	9.48	0.000
fat	0.1703	0.0111	15.29	0.000
x2	-1.910	0.637	-3.00	0.004
x3	-0.644	0.640	-1.01	0.317
x4	0.992	0.640	1.55	0.126
x5	1.175	0.474	2.48	0.015

a. Complete the ANOVA table and conduct an F test for a significant regression with $\alpha = 0.01$.
b. Does sex affect the weight of the child? Justify your answer.
c. Estimate the mean weight for an Asian American female with a daily fat intake of 88 grams.
d. How does the regression line shift from a White male to an African American female? To an Asian American female? Justify your answers.

12.205 Physical Sciences Ham radio operators from all over the world communicate with one another and are often very

helpful during emergencies. The communication distance is related to the radio signal power at the antenna, which is related to the transmitter power and the feed line loss. A random sample of amateur radio operators was obtained. The transmitter power (x_1, in watts), the feed line loss (x_2, in dB), and the output power from the feed line (y, in watts) was measured for each. Consider the regression model (with an interaction term) $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + E_i$. RADIO

a. Find the estimated regression coefficients.
b. Estimate the mean output power for $x_1 = 40$ and $x_2 = 5$.
c. Conduct the appropriate hypothesis tests to determine which regression coefficients are significantly different from 0.
d. Write a new regression model based on your results from part (c). Find the estimated regression coefficients in this model.
e. Estimate the mean output power for $x_1 = 40$ and $x_2 = 5$ using the second model. Which estimate do you think is more accurate? Why?

12.206 Manufacturing and Product Development It is important for textile manufacturers to understand the effect of various factors on the tensile properties of woven fabric to achieve certain fabric performance. A random sample of 100% cotton fabrics was obtained. The following characteristics were measured for each.[50] FABRIC

y = percent utilization of single-yarn strength
x_1 = number of load-bearing yarns per centimeter
x_2 = number of transverse yarns per centimeter
x_3 = linear density of load-bearing yarns
x_4 = linear density of transverse yarns
x_5 = strength of single load-bearing yarn in newtons
x_6 = strength of single transverse yarn in newtons
x_7 = float length
x_8 = crimp percentage in the load-bearing yarn
x_9 = crimp percentage in the transverse yarn

a. Find the estimated regression coefficients with all variables included in the model. Conduct the model utility test with $\alpha = 0.01$.
b. Conduct the appropriate hypothesis test to determine which regression coefficients are significantly different from 0.
c. Write a new model based on your results in part (b). Find the estimated regression coefficients in this new model. Find r^2 and interpret this value. Carefully sketch a normal probability plot for the residuals. Is there any evidence that the random error terms are not normal? Justify your answer.

Challenge

12.207 Biology and Environmental Science A logistic curve is often used to model the spread of a disease (or a rumor), the growth of a particular population, a biological response, or the cumulative sales of a specific product. A logistic curve is nonlinear, and the equation may be written as

$$y = \frac{L}{1 + ae^{bx}}$$

where L is called the *carrying capacity* (for example, the maximum population an area could support). Use a little algebra and take the logarithm of both sides to produce

$$\ln\left(\frac{L}{y} - 1\right) = \ln(ae^{bx}) = \ln a + bx$$

If we know, or can estimate, L, then this last equation is *linear*.

$$Y = \ln a + bx \quad \text{where} \quad Y = \ln\left(\frac{L}{y} - 1\right)$$

The town of Anaconda, Montana, has a population of approximately 10,000. When one person in town gets the flu during the winter, the virus spreads through the town according to a logistic curve. The local health clinic records all cases of the flu in town. The total number of people who have had the flu (y), and the days since the first reported case (x) are given on the text website. On the first day of reported flu cases, day 0, 160 people had the flu. FLU

a. Construct a scatter plot of the data.
b. Use $L = 10{,}000$ and transform the data. Find estimates of the parameters a and b.
c. Find an estimate of the number of people in the town who will have had the flu 80 days after the first reported case.
d. The rate of maximum growth of the number of flu cases is the point of inflection on the curve, the point where the concavity changes. Estimate the number of days after the first reported case when the rate of maximum growth occurs. How many cases of the flu had been recorded by that time? Use your estimated logistic equation to find the number of cases of the flu by that time.

12.208 Biology and Environmental Science The median–median line is an alternative to the least-squares regression line. Given n pairs of observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, here's how it works.

a. Arrange the observations in order, smallest to largest, based on the x values. Divide the ordered observations into three parts: group 1, the smallest third; group 2, the middle third; and group 3, the largest third. If the number of observations is not divisible by 3:
  i. If there is one extra pair, include it in the middle group.
  ii. If there are two extra pairs, include one each in group 1 and group 3.
b. Find the median values of x and y for each group: $(\tilde{x}_1, \tilde{y}_1), (\tilde{x}_2, \tilde{y}_2), (\tilde{x}_3, \tilde{y}_3)$,
c. Find the equation of the line through $(\tilde{x}_1, \tilde{y}_1)$ and $(\tilde{x}_3, \tilde{y}_3)$. (Remember point-slope format?)
d. Adjust this line one-third of the distance to $(\tilde{x}_2, \tilde{y}_2)$. This is the median–median line.

Results of a study published in the *American Journal of Botany*[51] suggested that the vessel diameter in trees is affected by elevation.

A random sample of maple tree seedlings was obtained, and the vessel diameter (y, in microns) and the elevation (x, height above sea level, in feet) was recorded for each. MAPLE

a. Construct a scatter plot of the data. Describe the relationship between elevation and vessel diameter.
b. Find the median–median line.
c. Find the regression equation using the method of least squares.
d. Add both lines to the scatter plot. Which do you think is the better model? Why?
e. Use both equations to estimate the mean vessel diameter for a maple tree seedling 1000 feet above sea level.

12.7 Model Selection Procedures

In a multiple linear regression model, several independent variables are used to describe the variation in a dependent variable. Frequently, there are many independent variable *candidates* for inclusion in a model. As we have seen in previous examples and exercises, a hypothesis test of H_0: $\beta_i = 0$ may not be rejected. If there is no evidence to suggest that the regression coefficient is different from 0, the multiple regression model may be *better* without the corresponding independent variable candidate.

Consider a multiple regression model with dependent variable ocean water temperature (at a certain location). This value may be affected by (the independent variables) speed of the surrounding current, air temperature, salinity, humidity, air pressure, and cloud cover. Some of these independent variables might *move together*. Therefore, the final model may need to include only one of these variables. And other variables may simply have no effect on water temperature.

The purpose of this section is to present methods for selecting the *best* multiple regression model. Given a collection of independent variables, we need to decide which ones contribute the most to the variation in y. Using the list of independent variable candidates, the goal is to construct (build) the *best* multiple regression model.

Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + E_i$$

in which there are k independent variable candidates for inclusion in the final model. One technique for building a model is to consider all possible subsets (or collections) of the independent variable candidates. One could use some reasonable measure of model *goodness*, and based on this value, select the best model.

Using this all-possible-subsets approach, we need to consider all models with one independent variable, all models with two independent variables, etc., and finally the model with all k independent variables. If we assume that β_0 is included in the final model, then there are 2^k possible models to consider. For example, if there are $k = 5$ independent variable candidates, then there are $2^5 = 32$ possible subset models.

We assume that the number of observations, n, is greater than k, and it is advantageous to have n a lot bigger than k.

One reasonable measure of model goodness is r^2, the coefficient of determination, a measure of the proportion of the variation in the data that is explained by the regression model. If model 1 has a larger r^2 than model 2, then model 1 can be used to explain more of the variation in the dependent variable. Recall, however, that adding *any* additional independent variable to a model will always increase r^2. Therefore, we want to consider *significant* increases in the value of r^2 when an additional independent variable is added to the model. A significant increase is usually a subjective judgment.

Example 12.18 Exposure to Manganese

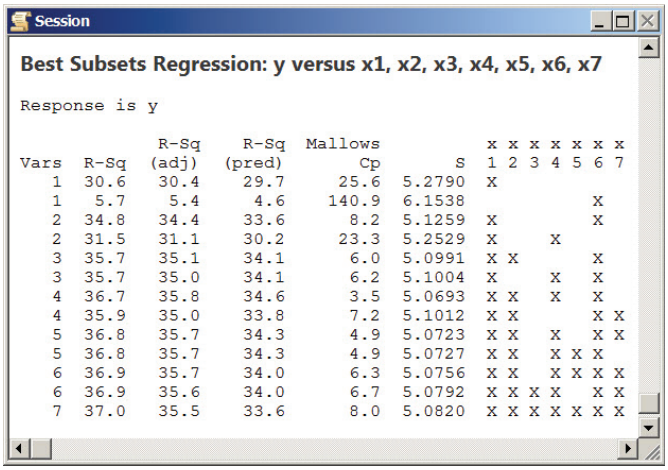
DATA SET
MANGANES

Manganese is an essential trace element. Found mostly in bones, the liver, kidneys, and the pancreas, manganese helps our bodies form connective tissue and bones. However, excessive exposure to manganese can be toxic, especially if it is inhaled. A study involving Mexican children living near mines suggested that intellectual function is inversely related to airborne manganese exposure.[52] The goal was to build a multiple regression model to predict the total IQ of a child based on other independent observations. Suppose

a random sample of children living in Mexican mining towns was obtained, and the following variables were measured for each child:

Variable	Description
y	Total IQ level
x_1	Age of the child, in years
x_2	Concentration of manganese in the blood, in μg/L
x_3	Fuel for cooking: 0, wood; 1, gas
x_4	Father a miner: 0, no; 1, yes
x_5	Mother's education in years
x_6	Number of miles from home to the nearest mine
x_7	Sex: 0, male; 1, female

The Minitab output using the Stat; Regression; Regression; Best Subsets command is shown in Figure 12.105 (the response variable is y, and the predictor variables are x_1–x_7). Use the value of r^2 to select the best multiple linear regression model from all of the possible subsets of independent variables.


Session

Best Subsets Regression: y versus x1, x2, x3, x4, x5, x6, x7

Response is y

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	x 1	x 2	x 3	x 4	x 5	x 6	x 7
1	30.6	30.4	29.7	25.6	5.2790	X						
1	5.7	5.4	4.6	140.9	6.1538						X	
2	34.8	34.4	33.6	8.2	5.1259	X					X	
2	31.5	31.1	30.2	23.3	5.2529	X			X			
3	35.7	35.1	34.1	6.0	5.0991	X	X				X	
3	35.7	35.0	34.1	6.2	5.1004	X			X		X	
4	36.7	35.8	34.6	3.5	5.0693	X	X		X		X	
4	35.9	35.0	33.8	7.2	5.1012	X	X				X	X
5	36.8	35.7	34.3	4.9	5.0723	X	X		X		X	X
5	36.8	35.7	34.3	4.9	5.0727	X	X		X	X	X	
6	36.9	35.7	34.0	6.3	5.0756	X	X		X	X	X	X
6	36.9	35.6	34.0	6.7	5.0792	X	X	X	X		X	X
7	37.0	35.5	33.6	8.0	5.0820	X	X	X	X	X	X	X

Figure 12.105 Minitab Best Subsets Regression output.

SOLUTION

STEP 1 All possible subsets of the independent variables are considered when using the Best Subsets Regression command. First, all models with a single independent variable are considered, then all models with two independent variables, etc., until the final model with all independent variables included. The calculations are all completed in the background, and Minitab uses r^2 to determine which models are the best. Minitab displays the two best models by default, but will display up to five models for each value of k, the number of independent variables in the model.

STEP 2 The other statistics reported in the Minitab output are

R-Sq(adj): Adjusted r^2. This is a modification of r^2 that takes into account the number of independent variables in the model. The adjusted r^2 can actually be negative, it will always be less than or equal to r^2, and it does not have the same interpretation as r^2.

R-Sq (pred): Predicted r^2. This is a measure of how well a model predicts new responses. A regression model could fit the given data very well but predict the value of new responses inadequately. The predicted r^2 can be negative, and a value much less than r^2 suggests that there are too many terms in the model.

Mallows C_p: This measure involves the mean square due to error, MSE. Values of C_p that indicate a good regression model include small values around k, the number of independent variables in the model.

s: The estimate of the standard deviation of the error terms. Small values of s are desirable, indicating that the observations do not vary much from the estimated regression line.

STEP 3 Using the r^2 criterion, with a check on Mallows C_p and s, the first model with four independent variables seems to be the best. $r^2 = 0.367$, and it does not increase significantly with the addition of any other independent variables. Mallows C_p is *close* to 4, and s does not decrease significantly with additional predictors in the model. The estimated regression equation is

$$y = 109.48 - 2.768x_1 - 0.313x_2 - 1.245x_4 - 0.474x_6$$

The Minitab output is shown in Figure 12.106.

Session

Regression Analysis: y versus x1, x2, x4, x6

```
Analysis of Variance
Source       DF   Adj SS   Adj MS  F-Value  P-Value
Regression    4   4389.4  1097.35    42.70    0.000
Error       295   7580.9    25.70
Total       299  11970.3

Model Summary
      S    R-sq  R-sq(adj)  R-sq(pred)
5.06930  36.67%     35.81%      34.64%

Coefficients
Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  109.48     2.56    42.79    0.000
x1        -2.768    0.242   -11.42    0.000  1.01
x2        -0.313    0.145    -2.16    0.032  1.01
x4        -1.245    0.587    -2.12    0.035  1.00
x6        -0.474    0.107    -4.41    0.000  1.00

Regression Equation
y = 109.48 - 2.768 x1 - 0.313 x2 - 1.245 x4 - 0.474 x6
```

Figure 12.106 Minitab regression analysis output.

The most important variables in predicting a child's IQ are age of the child, concentration of manganese in the blood, whether the child's father is a miner, and the number of miles from home to the nearest mine.

STEP 4 The graph of r^2, or s, versus k, for various models may be helpful in selecting the best regression model. See Figures 12.107 and 12.108. The Best Subsets Regression command was used to generate more data for these graphs.

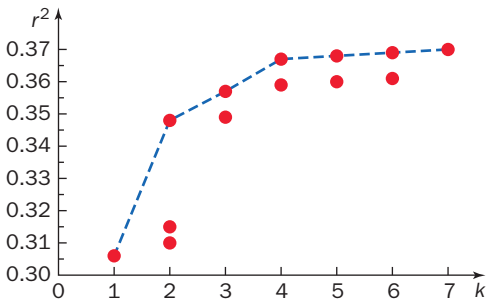


Figure 12.107 Graph of r^2 versus k, for various models.

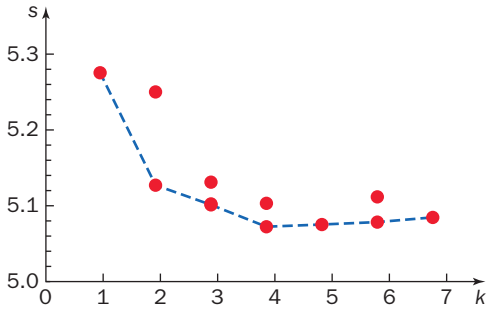


Figure 12.108 Graph of s versus k, for various models.

Although the model selection procedure described above is reasonable, it has several disadvantages. Selecting the best model is very subjective. There is no clear, concise method for deciding if an increase in r^2 is really significant. In addition, if the number of possible independent variables is large, for example 30, it would take lots of computer power and time to check all possible models.

There are several very prescriptive algorithms that produce the *best* multiple regression model. They may or may not all produce the same final model. One widely used procedure is **forward selection**. Using this technique, the single most significant independent variable, based on the t tests for significant regression coefficients, is added to the model at each step. Another similar procedure is backward elimination.

The following example demonstrates the method of forward selection.

Example 12.19 Forward Selection

Recall that in a simple linear regression model, the t test of H_0: $\beta_1 = 0$ is equivalent to the F test for an overall significant regression.

It is possible that no predictor variable produces a significant regression. And, often, the traditional p-value cutoff value of 0.05 is relaxed to 0.10 in forward selection.

Suppose there are five possible independent variables in a multiple linear regression model: x_1, x_2, x_3, x_4, x_5.

SOLUTION

STEP 1 Consider all five simple linear regression models: $Y_i = \beta_0 + \beta_1 x_{ki} + E_i$, for $k = 1, 2, 3, 4, 5$. In each case, conduct the hypothesis test to determine if the predictor variable is significant. That is, consider the test of H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. The following table shows the t statistic and the p value for each of these tests.

Model variable	t Statistic	p Value
x_1	3.34	0.0016
x_2	3.47	0.0011
x_3	−0.18	0.8579
x_4	−3.90	0.0003
x_5	−1.20	0.2359

Add the *most significant* predictor variable to the model, the independent variable that produces the smallest p value (below some threshold value, for example, ≤ 0.05). In this example, add the variable x_4 to the model.

STEP 2 Consider the four regression models with x_4 to determine whether any other predictor helps to explain the variation in the dependent variable: $Y_i = \beta_0 + \beta_1 x_{4i} + \beta_2 x_{ki} + E_i$ for $k = 1, 2, 3, 5$. Conduct the hypothesis test to determine if the additional predictor variable is significant (H_0: $\beta_2 = 0$ versus $\beta_2 \neq 0$). The following table shows the t statistic and p value for each of these tests.

Added variable	t Statistic	p Value
x_1	1.64	0.1075
x_2	4.24	0.0001
x_3	−0.38	0.7056
x_5	−0.67	0.5061

Add the variable with the smallest p value (≤ 0.05). Therefore, add x_2 to the model.

STEP 3 Continue in this manner until no additional variable is significant.

Consider the three regression models with x_4 and x_2 to determine if any other predictor helps to explain the variation in the dependent variable: $Y_i = \beta_0 +$

$\beta_1 x_{4i} + \beta_2 x_{2i} + \beta_3 x_{ki} + E_i$ for $k = 1, 3, 5$. Conduct the hypothesis test to determine whether the additional predictor variable is significant (H_0: $\beta_3 = 0$ versus $\beta_3 \neq 0$). The following table shows the t statistic and p value for each of these tests.

Added variable	t Statistic	p Value
x_1	1.41	0.1651
x_3	−1.22	0.2286
x_5	0.05	0.9603

All p values are greater than 0.05. No additional variable is added to the model. The forward selection procedure yields the final multiple regression model:

$$Y_i = \beta_0 + \beta_1 x_{4i} + \beta_2 x_{2i} + E_i$$

Example 12.20 Exposure to Manganese (Continued)

Use the forward selection procedure to find the best multiple linear regression model to predict IQ level.

SOLUTION

STEP 1 Minitab output using forward selection is shown in Figure 12.109. Select Stat; Regression; Regression; Fit Regression Model. Enter the response variable (column), y, the Continuous predictors (x_1, x_2, x_5, x_6), and the Categorical predictors (x_3, x_4, x_7). In the Stepwise menu, select Forward selection from the Method pull-down menu, enter 0.05 for Alpha to enter, and select Include details for each step from the Display menu.

Session

Regression Analysis: y versus x1, x2, x5, x6, x3, x4, x7

```
Method
Forward Selection of Terms
Candidate terms: x1, x2, x4, x6, x5, x3, x7
             -----Step 1----    -----Step 2----    -----Step 3----
               Coef        P      Coef        P      Coef        P
Constant     104.67             106.33             108.77
x1           -2.879    0.000    -2.811    0.000    -2.769    0.000
x6                              -0.475    0.000    -0.477    0.000
x2                                                 -0.296    0.043
x4

S                    5.27905            5.12589            5.09910
R-sq                  30.62%             34.81%             35.71%
R-sq(adj)             30.39%             34.37%             35.05%
R-sq(pred)            29.74%             33.63%             34.10%
Mallows' Cp            25.56               8.15               5.99

             -----Step 4----
               Coef        P
Constant     109.48
x1           -2.768    0.000
x6           -0.474    0.000
x2           -0.313    0.032
x4           -1.245    0.035

S                    5.06930
R-sq                  36.67%
R-sq(adj)             35.81%
R-sq(pred)            34.64%
Mallows' Cp            3.53

Coefficients
Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant   109.48     2.56    42.79    0.000
x1        -2.768    0.242   -11.42    0.000  1.01
x2        -0.313    0.145    -2.16    0.032  1.01
x4        -1.245    0.587    -2.12    0.035  1.00
x6        -0.474    0.107    -4.41    0.000  1.00
```

Figure 12.109 Minitab output using forward selection.

STEP 2 The first variable included in the model is x_1. The variables added in order via forward selection are x_6, x_2, x_4. Each p value to enter the model is less than 0.05. At each step, the value of s, r^2, adjusted r^2, predicted r^2, and Mallows C_p are given.

STEP 3 The final estimated regression equation is

$$y = 109.48 - 2.768x_1 - 0.313x_2 - 1.245x_4 - 0.474x_6$$

Note that increasing the value of Alpha to enter the model may add predictor variables to the final estimated regression equation. ■

The method of **backward elimination** begins with the maximum model, a regression equation with all independent variables included. At each step, the single *least* significant variable, based on t tests for significant regression coefficients, is eliminated from the model. The following example demonstrates the method of backward elimination.

Example 12.21 Backward Elimination

Suppose there are four possible independent variables in a multiple linear regression model: x_1, x_2, x_3, x_4. (Assume the intercept term is in the model.)

SOLUTION

STEP 1 Consider the multiple linear regression model with all four predictors:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + E_i$$

Conduct each hypothesis test, H_0: $\beta_i = 0$ versus H_a: $\beta_i \neq 0$ for $i = 1, 2, 3, 4$. The following table shows the t statistic and the p value for each of these tests.

Model variable	t Statistic	p Value
x_1	2.24	0.0309
x_2	1.36	0.1816
x_3	0.25	0.8039
x_4	−4.30	0.0001

Eliminate the *least significant* predictor variable, the independent variable associated with the largest p value greater than 0.05. In this example, delete x_3 from the model. Note that all p values could be less than 0.05. This suggests that all predictors are significant, and the final regression equation will include all of the independent variables.

STEP 2 Consider the multiple linear regression model with the remaining three predictors:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_4 x_{4i} + E_i$$

Conduct each hypothesis test, H_0: $\beta_i = 0$ versus H_a: $\beta_i \neq 0$ for $i = 1, 2, 4$. The following table shows the t statistic and the p value for each of these tests.

Model variable	t Statistic	p Value
x_1	2.56	0.0146
x_2	1.02	0.3142
x_4	−4.34	0.0001

Eliminate the variable with the largest p value (> 0.05). Therefore, eliminate x_2 from the model.

STEP 3 Continue in this manner until no variable can be eliminated from the model.

Consider the multiple linear regression model with the x_1 and x_4:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_4 x_{4i} + E_i$$

Conduct each hypothesis test, H_0: $\beta_i = 0$ versus H_a: $\beta_i \neq 0$ for $i = 1, 4$. The following table shows the t statistic and the p value for each of these tests.

Model variable	t Statistic	p Value
x_1	3.01	0.0047
x_4	−4.57	0.0000

There are no p values greater than 0.05. No other variable is eliminated. The backward elimination procedure suggests that the best model includes only the variables x_1 and x_4. ■

Example 12.22 Exposure to Manganese (Continued)

Use the backward elimination procedure to find the best multiple linear regression model to predict IQ level.

SOLUTION

STEP 1 Minitab output from backward elimination is shown in Figure 12.110. Select Stat; Regression; Regression; Fit Regression Model. Enter the response variable

Session

Regression Analysis: y versus x1, x2, x5, x6, x3, x4, x7

```
Method
Backward Elimination of Terms
Candidate terms: x1, x2, x4, x6, x5, x3, x7
             -----Step 1----    -----Step 2----    -----Step 3----
               Coef        P      Coef        P      Coef        P
Constant     108.77            109.15            109.77
x1           -2.751    0.000   -2.760    0.000   -2.763    0.000
x2           -0.324    0.028   -0.330    0.024   -0.321    0.028
x4           -1.212    0.041   -1.218    0.039   -1.228    0.038
x6           -0.463    0.000   -0.473    0.000   -0.477    0.000
x5            0.117    0.410    0.112    0.430
x3            0.309    0.611
x7           -0.494    0.403   -0.482    0.414   -0.474    0.422

S                    5.08201           5.07559           5.07234
R-sq                  37.00%            36.94%            36.81%
R-sq(adj)             35.49%            35.65%            35.73%
R-sq(pred)            33.63%            34.01%            34.34%
Mallows' Cp             8.00              6.26              4.88
             -----Step 4----
               Coef        P
Constant     109.48
x1           -2.768    0.000
x2           -0.313    0.032
x4           -1.245    0.035
x6           -0.474    0.000
x5
x3
x7

S                    5.06930
R-sq                  36.67%
R-sq(adj)             35.81%
R-sq(pred)            34.64%
Mallows' Cp             3.53

Coefficients
Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  109.48     2.56    42.79    0.000
x1        -2.768    0.242   -11.42    0.000  1.01
x2        -0.313    0.145    -2.16    0.032  1.01
x4        -1.245    0.587    -2.12    0.035  1.00
```

Figure 12.110 Minitab output using backward elimination.

(column), y, the Continuous predictors (x_1, x_2, x_5, x_6), and the Categorical predictors (x_3, x_4, x_7). In the Stepwise menu, select Backward elimination from the Method pull-down menu, enter 0.05 for Alpha to enter, and select Include details for each step from the Display menu.

STEP 2 The first variable eliminated from the model is x_3: $p = 0.611 > 0.05$. The p value is the largest among the seven, and is greater than 0.05.

STEP 3 The other variables eliminated from the model in order are x_5 and x_7.

STEP 4 At Step 4, no p value is greater than 0.05. The final estimated regression equation is

$$y = 109.48 - 2.768x_1 - 0.313x_2 - 1.245x_4 - 0.474x_6$$

Note that this is the same model produced using forward selection (and best subsets). This is not always the case. Forward selection and backward elimination may result in different models. ■

Recall that if all possible subsets of predictor variables are considered in order to construct the best multiple linear regression model, this could take a lot of computer power and time. Forward selection and backward elimination greatly reduce the number of models considered and, therefore, the computation time. However, there is a chance of missing a better model, especially in cases with many possible predictor variables.

Using forward selection, once a variable is included in the model, it remains in the model regardless of other variables added later. Similarly, using backward elimination, if a variable is eliminated from the model, it can never be included in a later step.

Stepwise regression is a modification of forward selection, or backward elimination. At each step in the procedure, the entire model is re-evaluated. Here is how it works applied to forward selection.

Suppose the variable x_i is added to the model in the usual way at a certain step. Before another variable is added, compute the current estimated regression equation and test the significance of each variable in the model. Consider each hypothesis test for a significant regression coefficient. Delete the variable with the highest p value (above a threshold, for example, 0.05). Recompute the estimated regression equation if necessary, and proceed with the next step in forward selection.

There is an analogous procedure for stepwise regression applied to backward elimination. Stepwise regression is better than either forward selection or backward elimination for selecting the best model because it considers more models (but still not all possible subsets of predictor variables). Most statistical software packages have an option for stepwise regression applied to forward selection and/or backward elimination.

SECTION 12.7 EXERCISES

Concept Check

12.209 True/False Adding any additional independent variable to a regression model will always increase the value of r^2.

12.210 True/False Forward selection and backward elimination will always produce the same multiple linear regression model.

12.211 True/False Using forward selection, increasing the value of α to enter the model may add predictor variables to the final estimated regression equation.

12.212 True/False Using backward elimination, at each step, eliminate the variable with the largest t statistic greater than 1.96.

12.213 Short Answer Name one disadvantage to using the best-subsets method to select the best multiple linear regression model.

12.214 Short Answer Name two reasonable measures of multiple linear regression model goodness.

Practice

12.215 Describe stepwise regression applied to backward elimination.

12.216 Suppose an investigator wants to find the best model in which there are 10 possible predictor variables.

a. How many possible models are there if the researcher considers all possible subsets?

b. Suppose forward selection is used, and the final model contains all 10 independent variables. How many models were considered in the process?

c. Suppose backward elimination is used, and the final model contains a single independent variable. How many models were considered in the process?

12.217 Consider the following Minitab output using the Stat; Regression; Regression; Best Subsets. EX12.217

Session

Best Subsets Regression: y versus x1, x11, x12, x13, x14, x15, x16

Response is y

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	x1	x11	x12	x13	x14	x15	x16
1	64.9	64.4	62.3	61.7	9.1405			X				
1	36.8	35.8	32.5	160.8	12.267	X						
1	1.5	0.0	0.0	285.4	15.318		X					
1	0.2	0.0	0.0	289.9	15.419			X				
2	80.4	79.8	78.5	9.0	6.8821	X		X				
2	66.6	65.6	62.1	57.7	8.9865			X				X
2	66.2	65.1	63.0	59.3	9.0478	X		X				
2	65.6	64.5	62.1	61.1	9.1173				X	X		
3	82.0	81.1	79.0	5.5	6.6546		X	X				X
3	81.4	80.5	79.1	7.5	6.7598	X	X	X				
3	80.8	79.8	78.1	9.9	6.8785		X		X	X		
3	80.7	79.8	78.2	10.1	6.8884		X	X	X			
4	82.8	81.7	79.6	4.5	6.5473	X	X	X				X
4	82.5	81.3	79.1	5.8	6.6168		X	X	X			X
4	82.3	81.1	78.4	6.5	6.6577		X		X	X		X
4	82.0	80.9	78.5	7.3	6.6998		X		X		X	X
5	83.5	82.1	79.8	4.1	6.4727	X	X	X	X			X
5	83.0	81.5	78.9	6.1	6.5789	X	X		X	X		X
5	82.9	81.4	79.0	6.4	6.5989	X	X		X		X	X
5	82.6	81.1	78.3	7.4	6.6499		X	X	X	X		X
6	83.5	81.9	79.3	6.0	6.5207	X	X	X	X		X	X
6	83.5	81.8	79.1	6.1	6.5257	X	X	X	X	X		X
6	83.0	81.2	78.3	8.0	6.6312	X	X		X	X	X	X
6	82.7	80.9	77.7	9.1	6.6916		X	X	X	X	X	X
7	83.5	81.6	78.5	8.0	6.5751	X	X	X	X	X	X	X

a. Construct a graph of r^2 versus k, for various models.

b. Construct a graph of s versus k, for various models.

c. Which predictor variables do you think should be included in the model? Justify your answer.

12.218 Consider the data on the text website, where y is the dependent variable and x_1–x_5 are possible predictor variables. EX12.218

a. Use the value of r^2 to select the best multiple linear regression model from all possible subsets of independent variables.

b. Construct a graph of r^2 versus k to support your conclusion in part (a).

c. Use the best model to find an estimate of the mean value of Y for $x = (60, 51, 2.2, 35.67, 56)$.

12.219 Consider the data on the text website, where y is the dependent variable and x_1–x_4 are possible predictor variables. EX12.219

a. Use forward selection to find the best multiple linear regression model.

b. Use stepwise regression applied to forward selection to find the best multiple linear regression model. Is the resulting model any different from part (a)? If so, why?

12.220 Consider the following Minitab output using the Stat; Regression; Regression; Fit Regression Model; and the Stepwise Method: Backward elimination option. EX12.220

Session

Regression Analysis: y versus x1, x1^2, x2, x2^2, x1*x2

Backward Elimination of Terms

Candidate terms: x1, x1^2, x2, x2^2, x1*x2

	Step 1 Coef	Step 1 P	Step 2 Coef	Step 2 P	Step 3 Coef	Step 3 P
Constant	223		79.8		27.1	
x1	-38.1	0.595				
x1^2	-3.37	0.474	-5.81	0.000	-5.297	0.000
x2	-8.4	0.446	-7.6	0.484		
x2^2	2.062	0.000	2.055	0.000	1.948	0.000
x1*x2	0.87	0.402	0.78	0.444	0.224	0.723
S		58.5379		57.9869		57.6141
R-sq		98.10%		98.08%		98.06%
R-sq(adj)		97.84%		97.88%		97.91%
R-sq(pred)		97.42%		97.58%		97.64%
Mallows' Cp		6.00		4.29		2.78

	Step 4 Coef	Step 4 P
Constant	26.0	
x1		
x1^2	-5.089	0.000
x2		
x2^2	2.0030	0.000
x1*x2		
S	56.9824	
R-sq	98.05%	
R-sq(adj)	97.96%	
R-sq(pred)	97.74%	
Mallows' Cp	0.90	

The possible predictor variables are x_1, x_1^2, x_2, x_2^2, and x_1x_2. The first model considered in this process is quadratic in x_1 and x_2, and includes an interaction term.

a. Which variable is eliminated first? Second?

b. Which variables are included in the best multiple linear regression model?

Applications

12.221 Fuel Consumption and Cars Research suggests that it can cost significantly more to repair luxury cars (than non-luxury cars) involved in low-speed crashes. Accidents that occur in parking lots or commuter traffic usually cause damage to bumpers, grills, and headlights. To build a model to predict the cost of repairs, the Insurance Institute for Highway Safety obtained records from random luxury cars involved in low-speed crashes. The following variables were considered: REPAIRS

y = repair cost as a result of the accident damage
x_1 = list price of the automobile
x_2 = speed of the automobile at the time of the accident
x_3 = indicator variable: 0 = forward, 1 = reverse
x_4 = wheelbase, in inches
x_5 = curb weight, in pounds
x_6 = engine power, in HP

The data are given on the text website.

a. Use backward elimination to find the best multiple linear regression model.

b. Suppose a 2014 BMW 528i is involved in a 3-mph crash while backing into a parking spot at a mall. The other specifications include MSRP, $49,500; HP, 240; wheelbase, 116.9 inches; curb weight, 3814 pounds. Find a 95% confidence interval for the mean repair cost.

12.222 Public Health and Nutrition Suppose a two-year study was conducted to investigate the effect of nutrient intake

on changes in body composition. Thirty-six women, 18–31 years of age, were selected at random. A diet survey was used to determine daily nutrient intake. The data are given on the text website, where BODYCOM

y = percent change in body weight
x_1 = daily fat intake, in grams
x_2 = daily protein intake, in grams
x_3 = daily calcium/energy intake, mg/kcal
x_4 = daily sodium intake, in mg
x_5 = daily vitamin A intake, in IU
x_6 = daily carbohydrate intake, in grams

a. Use backward elimination to find the best model.
b. Use forward selection to find the best model. Compare this model with the one in part (a).
c. Add another possible predictor variable, an interaction term, x_3x_5. Use backward elimination again to find the best model.

12.223 Medicine and Clinical Studies Physicians at the Emergency Medicine Research (EMR) Group in Coventry, England, conducted a study to predict the number of days a patient stays in the hospital based on emergency department data. The following observations were recorded for each patient: PATIENT

y = number of days in the hospital
x_1 = elapsed time from arrival in the emergency department until seen by a doctor
x_2 = elapsed time from initial evaluation until decision to admit/not admit
x_3 = severity of the injury, using the EMR scale
x_4 = elapsed time from accident to arrival at the emergency department
x_5 = initial pulse rate
x_6 = initial respiratory rate

The Minitab outputs from Best Subsets, Backward Elimination, and Forward Selection are shown below.

Session

Best Subsets Regression: y versus x1, x2, x3, x4, x5, x6

Response is y

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	x1	x2	x3	x4	x5
1	57.9	57.1	54.6	19.7	24.461		X			
1	10.0	8.3	2.5	99.1	35.780	X				
1	1.3	0.0	0.0	113.6	37.475					
2	70.7	69.5	66.3	0.6	20.619	X	X			
2	60.2	58.7	55.2	17.9	24.013		X			X
2	59.1	57.5	54.2	19.7	24.347			X	X	
3	71.5	69.8	66.2	1.3	20.545	X	X	X		
3	70.7	69.0	64.4	2.5	20.810	X	X			X
3	70.7	69.0	65.2	2.5	20.813	X	X			X
4	71.6	69.3	64.4	3.1	20.713	X	X	X	X	
4	71.5	69.2	64.9	3.2	20.732	X	X	X		X
4	71.5	69.1	64.7	3.3	20.752	X	X	X		
5	71.6	68.7	63.0	5.0	20.913	X	X	X	X	X
5	71.6	68.6	62.9	5.1	20.926	X	X	X	X	
5	71.5	68.6	63.4	5.1	20.938	X	X	X		X
6	71.6	68.0	61.5	7.0	21.128	X	X	X	X	X

Session

Regression Analysis: y versus x1, x2, x3, x4, x5, x6

Backward Elimination of Terms
Candidate terms: x1, x2, x3, x4, x5, x6

	Step 1 Coef	Step 1 P	Step 2 Coef	Step 2 P	Step 3 Coef	Step 3 P
Constant	5.1		10.6		21.53	
x1	-1.508	0.000	-1.495	0.000	-1.497	0.000
x2	3.243	0.000	3.257	0.000	3.263	0.000
x3	1.22	0.235	1.22	0.231	1.203	0.231
x4	-0.243	0.707	-0.250	0.695	-0.274	0.662
x5	0.161	0.767	0.135	0.794		
x6	0.158	0.869				
S	21.1279		20.9128		20.7131	
R-sq	71.63%		71.62%		71.57%	
R-sq(adj)	68.01%		68.66%		69.25%	
R-sq(pred)	61.46%		62.96%		64.37%	
Mallows' Cp	7.00		5.03		3.10	

	Step 4 Coef	Step 4 P	Step 5 Coef	Step 5 P
Constant	21.77		29.21	
x1	-1.558	0.000	-1.581	0.000
x2	3.237	0.000	3.224	0.000
x3	1.137	0.248		
x4				
x5				
x6				
S	20.5454		20.6191	
R-sq	71.46%		70.68%	
R-sq(adj)	69.75%		69.53%	
R-sq(pred)	66.16%		66.28%	
Mallows' Cp	1.28		0.57	

α to remove = 0.05

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	29.21	7.21	4.05	0.000	
x1	-1.581	0.336	-4.71	0.000	1.00
x2	3.224	0.314	10.28	0.000	1.00

Regression Equation
y = 29.21 - 1.581 x1 + 3.224 x2

Session

Regression Analysis: y versus x1, x2, x3, x4, x5, x6

Forward Selection of Terms
Candidate terms: x1, x2, x3, x4, x5, x6

	Step 1 Coef	Step 1 P	Step 2 Coef	Step 2 P
Constant	6.71		29.21	
x2	3.145	0.000	3.224	0.000
x1			-1.581	0.000
S	24.4608		20.6191	
R-sq	57.93%		70.68%	
R-sq(adj)	57.12%		69.53%	
R-sq(pred)	54.60%		66.28%	
Mallows' Cp	19.70		0.57	

α to enter = 0.05

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	29.21	7.21	4.05	0.000	
x1	-1.581	0.336	-4.71	0.000	1.00
x2	3.224	0.314	10.28	0.000	1.00

Regression Equation
y = 29.21 - 1.581 x1 + 3.224 x2

a. Using the Best Subsets output, construct a graph of r^2 versus various models. Which predictor variables would you recommend for inclusion in the model? Why?
b. Using the Backward Elimination output, what predictor variables would you recommend for inclusion in the model? Which variable was excluded first in this process? Second?

c. Using the Forward Selection output, what predictor variables would you recommend for inclusion in the model?
d. Consider all three methods. What predictor variables would you recommend for inclusion in the model? Do these predictor variables seem reasonable? Are any variables left out of the model that might be good predictors?

12.224 Physical Sciences Coal is still an abundant energy source and is used in many generating plants in the United States. For example, the coal-burning Montour Power Plant in Washingtonville, Pennsylvania, has produced 360 million megawatt hours of electricity in 40 years of operation. To manage production and purchasing, plant officials studied coal production at various mines. A random sample of mines in the United States was obtained, and the following observations were obtained for each one: COAL

y = output of raw coal, in metric tons
x_1 = area of the mine, in km^2
x_2 = total reserve, in metric tons
x_3 = recoverable reserve, in metric tons
x_4 = depth of the shaft, in meters
x_5 = designed annual output of raw coal
x_6 = average thickness of the coal seam, in meters

The data are given on the text website.[53]

a. Use backward elimination to find the best multiple linear regression model. Find the estimated regression coefficients.
b. Estimate the true mean value of Y when $x = (35.2, 132.85, 28.17, 297, 2.4, 8.6)$.
c. Find the residuals associated with your model. Construct a normal probability plot for the residuals. Is there any evidence to suggest that the random errors are not normal? Justify your answer.
d. Carefully sketch a graph of the residuals versus each predictor variable in your model. Is there any evidence of a violation of the regression assumptions? Justify your answer.

12.225 Travel and Transportation Ferryboats are a major source of public transportation is some cities. For example, New York City and Miami operate ferryboats, and Seattle, Washington, has approximately 28 ferryboats in operation. A study was conducted to develop a model to predict total yearly passenger miles. The following variables were considered: FERRY

y = total passenger miles, in thousands
x_1 = transportation zone population
x_2 = maximum number of vehicles operated
x_3 = maximum number of vehicles available
x_4 = vehicle revenue miles, in thousands
x_5 = vehicle revenue hours, in thousands
x_6 = passenger unlinked trips in thousands

A random sample of primary cities was obtained, and the data were recorded.[54]

a. Use backward elimination with $\alpha = 0.05$ to find the best model.
b. Use forward selection with $\alpha = 0.05$ to find the best model.
c. Repeat parts (a) and (b) with $\alpha = 0.10$. Does either procedure produce a different model?
d. Of all four models, which would you recommend as the best? Justify your answer.

12.226 Fuel Consumption and Cars Zipcar is a self-serve hourly car rental agency. After paying an annual membership fee, customers use a swipe ID card to rent available cars, usually in urban areas. There are no lines, no paperwork, and no service people. The company owner decided to conduct a study to predict the number of hours each car is rented. A random sample of rentals was obtained, and observations were recorded for the following variables: RENTAL

y = number of hours rented
x_1 = time of day for rental, in hours after 12:00 A.M.
x_2 = indicator variable, 0 = weekday, 1 = weekend
x_3 = renter's annual income, in thousands of dollars
x_4 = hourly rate for the rental
x_5 = annual membership fee

a. Use backward elimination with $\alpha = 0.10$ to find the best model. Find the estimated regression coefficients.
b. Use forward selection with stepwise regression, $\alpha = 0.10$, to find the best model. Compare this model with the one found in part (a).
c. Use the estimated regression coefficients to explain the relationship between the indicator variable and the hours a car is rented, and between the hourly rate and the hours a car is rented.
d. Suppose the hourly rate is \$10 on a weekend day. Find a 95% confidence interval for the true mean number of hours the car will be rented.

Extended Applications

12.227 Business and Management Whenever a consumer makes a purchase with a credit card, it must be verified and authenticated. Usually the card is *swiped,* and the sale is approved within a few seconds. A consumer group is trying to develop a model to predict the amount of time it takes to verify a purchase. A random sample of sales was obtained, and observations were recorded for the following variables: CREDIT

y = time until the purchase is approved, in seconds
x_1 = amount of the purchase, in dollars
x_2 = type of store, 0 = convenience, 1 = restaurant, 2 = retail
x_3 = distance from home, in miles
x_4 = time of day, in hours after 12:00 A.M.
x_5 = number of days until Christmas

a. Use backward elimination to find the best regression model ($\alpha = 0.10$). Use forward selection to find the best regression model ($\alpha = 0.10$). Don't forget to use the appropriate indicator variables.
b. One member of the research team believes that the time of day raised to the fourth power should be in the model.

Add this as a possible predictor, and use the method of your choice to find the best model.

c. Using the model found in part (b), find an estimate of the mean time for credit card verification when $x = (152.00, 1, 20, 18.5, 120)$.

d. Consider a model with x_4, x_4^2, x_4^3, and x_4^4. What happens when you try any method with these as possible predictors? Why?

12.228 Biology and Environmental Science Reliable measurements of river beds are important for water management, shipping, and flood predictions. A random sample of rivers in the United States was obtained, and observations were recorded for the following variables:[55] RIVER

y = total river-bed material load, in kg/s
x_1 = discharge, in m^3/s
x_2 = average velocity, in m/s
x_3 = bottom width, in meters
x_4 = flow depth, in meters
x_5 = area, in m^2
x_6 = longitudinal slope

a. Use forward selection with $\alpha = 0.05$ to find the best model. Use the sign of each regression coefficient to explain the relationship between each predictor in the model and the total river-bed material load.

b. Use backward elimination with $\alpha = 0.05$ to find the best model. Compare this model with the model in part (a). Carefully sketch a normal probability plot of the residuals. Is there any evidence to suggest a violation of the regression assumptions?

c. Suppose that the river bed will be dredged if the total river-bed material load is greater than 50 kg/s. Measurements for the Kissimmee River in Florida showed $x = (40, 1.25, 27, 1.05, 31.35, 0.015)$. Using the model in part (b), find a 95% confidence interval for the mean total river-bed material load. Use this confidence interval to determine if there is any evidence to suggest that this river should be dredged. Justify your answer.

12.229 Technology and the Internet A recent research study examined the effects of economic and information/communication technology (ICT) on Internet purchases by individuals in European Union member states.[56] The following variables were considered: NETBUY

y = Internet purchases in the last 12 months, expressed as a percentage
x_1 = level of Internet access, percentage
x_2 = fixed broadband penetration rate per 100 inhabitants
x_3 = level of computer skills, as a percentage of individuals aged 16–74
x_4 = public expenditure on education as a percentage of GDP
x_5 = GDP per capita in Purchasing Power Standards
x_6 = individuals using the Internet for finding information, as a percentage of individuals aged 16–74
x_7 = individuals' level of Internet skills, as a percentage of all individuals aged 16–74
x_8 = individuals who ordered/bought goods or services over the Internet, as a percentage of individuals aged 16–74
x_9 = concern about possible problems related to Internet usage, as a percentage of all individuals

A random sample of individuals was obtained, and the data were recorded.

a. Use forward selection and backward elimination to determine the best multiple linear regression model. Use the sign of each estimated coefficient to explain the effect of each predictor variable on the percentage of Internet purchases.

b. What would you say is the most important variable in predicting the percentage of Internet purchases? Why?

c. Construct a normal probability plot of the residuals. Is there any evidence to suggest that the random errors are not normal? Justify your answer.

d. Carefully sketch a graph of the residuals versus each predictor variable in the model. Is there any evidence of a violation of the regression assumptions?

12.230 Economics and Finance Researchers A. Q. Do and G. Grudnitski have written several articles concerning the prediction of the selling price of a residential home. They indicate that some of the important variables are the age of the home and the lot size. A real estate agent in Napa Valley would like to develop a model to predict selling price in her area. She would like to consider the following variables: PRICE

y = selling price of the home, in thousands of dollars
x_1 = age of the home, in years
x_2 = number of bedrooms
x_3 = number of bathrooms
x_4 = square footage of living area
x_5 = number of garage stalls
x_6 = number of fireplaces
x_7 = number of stories
x_8 = lot size, in acres
x_9 = indicator variable, abuts golf course (0 = no, 1 = yes)

A random sample of home sales was obtained, and the data were recorded.

a. Use any method you wish to develop the best multiple linear regression model. What would you say is the most important variable in predicting the selling price of a home in Napa Valley? Why?

b. Find the estimated regression equation. Explain the relationship between each predictor in the model and the selling price.

c. Construct a normal probability plot of the residuals. Is there any evidence to suggest that the random errors are not normal? Justify your answer.

d. Carefully sketch a graph of the residuals versus each predictor variable in the model. Is there any evidence of a violation of the regression assumptions?

e. Are there any other variables that you believe should be considered in the model? If so, why?

CHAPTER 12 SUMMARY

Multiple Linear Regression Model

Let $(x_{11}, x_{21}, \ldots, x_{k1}, y_1), (x_{12}, x_{22}, \ldots, x_{k2}, y_2), \ldots, (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)$ be n sets of observations such that y_i is an observed value of the random variable Y_i. We assume that there exist constants $\beta_0, \beta_1, \ldots, \beta_k$ such that

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + E_i$$

where $E_1, E_2, \ldots, E_n$ are independent, normal random variables with mean 0 and variance σ^2. That is,

1. The E_i's are normally distributed (which means that the Y_i's are normally distributed).
2. The expected value of E_i is 0 [which implies that

$$E(Y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}].$$

3. $\text{Var}(E_i) = \sigma^2$ [which implies that $\text{Var}(Y_i) = \sigma^2$].
4. The E_i's are independent (which implies that the Y_i's are independent).

The E_i's are the random deviations or random error terms. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is the true regression line.

Principle of least squares

The estimated regression equation is obtained by minimizing the sum of the squared deviations between the observations and the estimated values.

The estimated regression equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$.

The ith predicted (fitted) value is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$ $(i = 1, 2, \ldots, n)$.

The ith residual is $\hat{e}_i = y_i - \hat{y}_i$.

The sum of squares

$$\underbrace{\Sigma(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\Sigma(\hat{y}_i - \bar{y})^2}_{\text{SSR}} + \underbrace{\Sigma(y_i - \hat{y}_i)^2}_{\text{SSE}}$$

Analysis of variance table for multiple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F	p Value
Regression	SSR	k	$\text{MSR} = \dfrac{\text{SSR}}{k}$	$\dfrac{\text{MSR}}{\text{MSE}}$	p
Error	SSE	$n - k - 1$	$\text{MSE} = \dfrac{\text{SSE}}{n - k - 1}$		
Total	SST	$n - 1$			

Coefficient of determination: $r^2 = \text{SSR}/\text{SST}$

Estimate of variance: $s^2 = \text{SSE}/(n - k - 1)$

F test for a significant multiple linear regression

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

(None of the predictor variables helps to explain the variation in y.)

H_a: $\beta_i \neq 0$ for at least one i

(At least one predictor variable helps to explain the variation in y.)

TS: $F = \dfrac{\text{MSR}}{\text{MSE}}$

RR: $F \geq F_{\alpha,k,n-k-1}$

Hypothesis test concerning β_i

H_0: $\beta_i = \beta_{i0}$

H_a: $\beta_i > \beta_{i0}$, $\beta_i < \beta_{i0}$, or $\beta_i \neq \beta_{i0}$

TS: $T = \dfrac{B_i - \beta_{i0}}{S_{B_i}}$

RR: $T \geq t_{\alpha,n-k-1}$, $T \leq -t_{\alpha,n-k-1}$, or $|T| \geq t_{\alpha/2,n-k-1}$

A $100(1-\alpha)\%$ confidence interval for β_i has as endpoints the values

$$\hat{\beta}_i \pm t_{\alpha/2,n-k-1} s_{B_i}$$

Hypothesis test concerning the mean value of Y for $x = x^*$

H_0: $y^* = y_0^*$

H_a: $y^* > y_0^*$, $y^* < y_0^*$, or $y^* \neq y_0^*$

TS: $T = \dfrac{(B_0 + B_1 x_1^* + \cdots + B_k x_k^*) - y_0^*}{S_{Y^*}}$

RR: $T \geq t_{\alpha,n-k-1}$, $T \leq -t_{\alpha,n-k-1}$, or $|T| \geq t_{\alpha/2,n-k-1}$

A $100(1-\alpha)\%$ confidence interval for $\mu_{Y|x^*}$, the mean value of Y for $x = x^*$, has as endpoints the values

$$(\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_k x_k^*) \pm t_{\alpha/2,n-k-1} s_{Y^*}$$

Prediction interval for an observed value of Y

A $100(1-\alpha)\%$ prediction interval for an observed value of Y when $x = x^*$ has as endpoints the values

$$(\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_k x_k^*) \pm t_{\alpha/2,n-k-1} \sqrt{s^2 + s_{Y^*}^2}$$

Regression diagnostics

1. Construct a histogram, stem-and-leaf plot, scatter plot, and/or normal probability plot of the residuals. These graphs are all used to check the normality assumption.
2. Construct a scatter plot of the residuals versus *each* independent variable. If there are no violations in assumptions, each scatter plot should appear as a horizontal band around 0. There should be no recognizable pattern.

Polynomial model

A polynomial regression model includes quadratic or higher-degree terms.

Interaction term

An interaction term is the product of two (or more) predictor variables, for example, $x_1 x_2$.

Indicator variables: An indicator variable takes on only the value 0 or 1. If there are c categories to account for in a regression model, then the model is adjusted by adding $c - 1$ indicator variables.

Model selection procedures

1. r^2: A larger r^2 indicates that the model can be used to explain more of the variation in the dependent variable. Mallows C_p: Small values, near k, indicate a good regression model.

2. *Forward selection*: The single most significant independent variable is added to the model at each step.
3. *Backward elimination*: The single least significant independent variable is eliminated from the model at each step.
4. *Stepwise regression*: A modification to forward selection and backward elimination. At each step in the procedure, the entire model is re-evaluated. Applied to forward selection, this allows variables already in the model to be eliminated. Applied to backward elimination, this allows variables already eliminated from the model to be added.

CHAPTER 12 EXERCISES

APPLICATIONS

12.231 Demographics and Population Statistics As urban areas expand and residents build more secluded homes outside of standard city and town infrastructure, owners face inadequate fire protection due to the distance to the nearest hydrant. Many zoning ordinances and the National Fire Protection Association recommend a fire hydrant within 1000 feet of a residential dwelling. An insurance company conducted a study to investigate the relationship between claims due to fire and fire hydrants. A random sample of residential home fires was obtained, and observations were recorded for the following variables:[57] FIRE

y = insurance claim due to the fire, in thousands of dollars
x_1 = distance to the nearest fire hydrant, in feet
x_2 = water pressure at the fire hydrant, in psi

a. Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i$. Interpret each estimated regression coefficient.
b. Conduct the model utility test. Use $\alpha = 0.01$. Use technology to find the exact p value associated with this test.
c. Find the value of r^2 and interpret this value.
d. Suppose a new home is constructed 750 feet away from the nearest fire hydrant and the water pressure is approximately 50 psi. If there is a fire in this home, what is the expected loss?

12.232 Manufacturing and Product Development A manufacturer of metal sheets would like to predict the springback angle from a given punch stroke in order to determine the final dimensions of the sheet accurately. A random sample of metal sheets was obtained, and observations were recorded for the following variables: METAL

y = springback angle, in degrees
x_1 = punch stroke, in mm
x_2 = initial length of the sheet, in mm
x_3 = sheet strength coefficient

The following multiple linear regression model was used: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + E_i$. Minitab was used to analyze the data, and a portion of the output is shown below.

Session

Regression Analysis: y versus x1, x2, x3

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	39.754			
Error					
Total	33	58.642			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.793471	67.79%	64.57%	57.56%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.93	1.32	1.46	0.154	
x1	0.1505	0.0645	2.34	0.026	1.00
x2	0.04830	0.00930	5.19	0.000	1.01
x3	-0.4276	0.0698	-6.12	0.000	1.01

Regression Equation

y = 1.93 + 0.1505 x1 + 0.04830 x2 - 0.4276 x3

a. Complete the ANOVA table and conduct a model utility test. Find the exact p value.
b. Explain the relationship between each predictor variable and the springback angle.
c. Suppose $x^* = (5, 125, 9)$. Estimate the mean springback angle for this value of x^*.

12.233 Physical Sciences Because of movies such as *Deep Impact* and *Armageddon*, and the recent meteor strike in Russia over the Ural Mountains, the public has become more aware and concerned about objects striking Earth. An astronomer selected several well-documented meteor impacts on Earth at random, and, using sophisticated scientific equipment, measured values for the following variables:[58] METEOR

y = diameter of the crater created by the impact, in meters
x_1 = diameter of the object, in meters
x_2 = density of the object, in kg/m^3
x_3 = velocity of the object, in km/s
x_4 = angle of the impact, in degrees
x_5 = elevation of impact, in km

a. Use forward selection, with $\alpha = 0.05$, to find the best model. Find the estimated regression coefficients. Explain the meaning of each estimated regression coefficient.

b. Estimate the mean crater diameter for $x^* = (50, 8, 60, 45)$. (*Note*: Your model should include x_1, x_2, x_3 and x_4.)
c. Carefully sketch a normal probability plot. Is there any evidence to suggest that the random errors are not normal? Justify your answer.

12.234 Medicine and Clinical Studies Most prescription and over-the-counter medications have expiration dates. Under reasonable storage conditions, medication should retain at least 90% of potency, or remain stable, if used prior to the expiration date. A consumer group recently conducted a study to predict the potency of prescription medications. A random sample of drugs was obtained, and the following variables were measured for each: MEDS

y = potency, a percentage
x_1 = temperature at which the drug was stored, in °F
x_2 = time since the drug was manufactured, in months

a. Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i$.
b. Explain the relationship between each independent variable and the potency.
c. Construct the ANOVA table and conduct the model utility test. Find the exact p value.
d. Conduct the necessary hypothesis tests to determine whether each regression coefficient is significantly different from 0.
e. Check the model assumptions by constructing a normal probability plot of the residuals and the appropriate scatter plots.

12.235 Marketing and Consumer Behavior ATM cash machines are readily available throughout the United States and around the world. To satisfy customers, banks must carefully plan when to restock ATM machines with cash, and with how much. A large U.S. bank conducted a study to predict the amount of cash withdrawn by its customers at ATMs. A random sample of withdrawal transactions was obtained, and the following variables were measured for each: ATMS

y = amount of cash withdrawn
x_1 = number of visits to an ATM the previous month
x_2 = amount of money in the user's account, in thousands of dollars
x_3 = indicator variable, 1 if Friday, 0 otherwise

a. Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + E_i$.
b. Use the estimated regression coefficients to explain the relationship between each independent variable and the amount of cash withdrawn. Why do you think an indicator variable for Friday should be included in the model?
c. Construct the ANOVA table and conduct the model utility test. Find the exact p value.
d. Conduct the necessary hypothesis tests to determine whether each regression coefficient is significantly different from 0.
e. Suppose a customer has \$20,000 in her account, used an ATM four times last month, and withdraws cash on a Friday evening. Find a 95% prediction interval for an observed value of the amount of cash withdrawn.
f. Carefully sketch a normal probability plot of the residuals. Is there any evidence to suggest that the residuals are not normally distributed? Justify your answer.

12.236 Psychology and Human Behavior The bounty hunter profession is very risky, but the reward for apprehending a criminal who has jumped bail can be huge. The Miami Dade County Bail Bond Company recently conducted a study to predict the time to apprehend a criminal based on the amount of the reward. A random sample of criminals who jumped bail was obtained, and the reward (y, in thousands of dollars) and time until apprehension (x, in days) was recorded for each. BOUNTY

a. Carefully sketch a scatter plot of the data. Write an appropriate polynomial regression model.
b. Find the estimated regression equation.
c. Conduct the model utility test using $\alpha = 0.01$.
d. Find an estimate of the time until apprehension for a criminal who has jumped bail with a reward of \$50,000.
e. Find an estimate of the time until apprehension for a criminal who has jumped bail with a reward of \$200,000. Why does this estimate seem inconsistent? What error is made in using the estimated regression equation to find this estimate?

12.237 Travel and Transportation Advances in technology have made turboprop airplanes quieter and more efficient. The noise inside the cabin is caused by the aerodynamic noise, engine exhaust, engine vibration, auxiliary systems, and most of all, by the propeller. A random sample of turboprop airplanes was obtained, and tests were conducted to measure the resonant frequency of the propellers (x, in thousands of rpm) and the noise level (y, in dB). The data are given in the following table.[59] PROPS

y	x	y	x	y	x	y	x
60	54.5	52	56.7	44	57.9	88	52.2
91	51.8	69	54.5	96	51.4	58	46.8
89	48.3	81	51.6	75	49.1	51	43.6
61	55.6	64	53.7	81	52.9	67	51.7
80	45.2	72	53.8	70	46.3	68	47.3
46	55.9						

a. Carefully sketch a scatter plot of the data. Consider a quadratic regression model and find the estimated regression line.
b. Conduct the hypothesis tests with H_0: $\beta_i = 0$, $i = 1, 2$ and $\alpha = 0.05$. Are both regression coefficients significantly different from 0?
c. Find a 95% confidence interval for the mean sound level when the frequency is 57,000 rpm.
d. Estimate the frequency at which the maximum sound level occurs.
e. Construct a normal probability plot of the residuals and a scatter plot of the residuals versus resonant frequency. Is there any evidence of a violation of the regression assumptions? Justify your answer.

12.238 Economics and Finance A study was conducted by a research team at Fidelity Investments concerning the amount of

money individuals have saved for retirement. A random sample of working adults was obtained, and the percentage of income saved for retirement last year (y, in dollars), age (x_1, in years), and yearly salary (x_2, in thousands of dollars) was recorded for each. The data are given in the following table. INVEST

y	x_1	x_2	y	x_1	x_2
11	38	87	11	46	64
22	29	119	5	38	49
11	59	46	17	47	114
9	27	103	12	45	76
12	29	82	13	25	110
12	48	84	14	26	102
18	60	87	12	45	41
14	40	50	13	36	60
18	35	98	15	30	99
14	36	93	14	26	79
12	47	67	17	30	75
22	57	116	16	48	69

a. Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i$. Conduct the model utility test with $\alpha = 0.05$.

b. Compute the residuals and construct a normal probability plot for the residuals. Is there any evidence to suggest that the random error terms are not normal?

c. Construct a graph of the residuals versus each predictor variable. Is there any evidence of a violation of the regression assumptions? Justify your answer.

d. Suppose a 45-year-old is earning $80,000 per year. Find a 95% confidence interval for the mean percentage of yearly salary saved. Fidelity suggests an individual at this age and salary should save 17% of his or her yearly income. Using the confidence interval, is there any evidence to suggest that adults in this situation are not saving enough for retirement? Justify your answer.

12.239 The Eyes Have It A person's intraocular pressure (IOP) is an indicator of risk of glaucoma. There is some evidence to suggest that other health factors may also be related to IOP. A random sample of American male adults was obtained. Each person was subject to a variety of medical tests, and the following variables were measured:[60] EYES

y = IOP, right eye (mmHg)
x_1 = age (years)
x_2 = total cholesterol (mg/dL)
x_3 = high-density lipoprotein (mg/dL)
x_4 = triglyceride (mg/dL)
x_5 = body mass index (kg/m^2)

a. Use backward elimination with $\alpha = 0.05$ to find the best multiple linear regression model.

b. Use the estimated regression coefficients to explain the relationship between each independent variable in the final model and the IOP in the right eye of American males.

c. Construct the ANOVA table and conduct the model utility test. Find the exact p value.

d. Carefully sketch a normal probability plot of the residuals. Is there any evidence to suggest that the residuals are not normally distributed? Justify your answer.

e. Suppose an American male has the following measurements: $x = (40, 280, 42, 114, 24)$. Use your model to find a 95% confidence interval for the mean IOP in his right eye. Suppose an IOP value of 16 mmHg or greater is a good indicator of glaucoma. Use the confidence interval to determine whether there is evidence that this person has glaucoma in his right eye.

EXTENDED APPLICATIONS

12.240 Manufacturing and Product Development Noise-canceling headphones/earphones have become very popular for ordinary use but also in cars, planes, and even at the office. These devices work to block out external noise using active noise-reduction technology. A study was conducted to predict the percentage of sound eliminated. A random sample of headphones was selected. Each was subject to a noise-canceling test, and the following variables were measured: NOISE

y = percentage of sound eliminated
x_1 = impedance, in ohms
x_2 = sensitivity, in dB
x_3 = driver units, in mm
x_4 = noise control range at 300 Hz, in dB
x_5 = form factor, earbud = 0, headphone = 1

a. Use forward selection to find the best multiple linear regression model.

b. If you were to advise someone interested in buying a set of noise-canceling headphones, what characteristics would you recommend? Why?

c. Estimate the true mean percentage of sound reduction when $x = (18, 121, 40, 15, 1)$.

d. Construct a normal probability plot for the residuals. Is there any evidence to suggest that the random error terms are not normally distributed? Justify your answer.

e. Delete observation 21 (82, 25, 122, 40, 21, 0) from the data set. Use forward selection on this reduced data set to find the best multiple linear regression model. Explain any differences between this model and the one in part (a).

12.241 Public Health and Nutrition Many breakfast cereal companies advertise brands with high fiber, which can lower the risk of heart disease, cancer, and diabetes. The U.S. Food and Drug Administration recently conducted research to predict the amount of fiber in one cup of various breakfast cereals. A random sample of cereals was obtained, and the following variables were measured for each one-cup serving:[61] CEREAL

y = fiber, in grams
x_1 = calories
x_2 = fat, in grams
x_3 = protein, in grams
x_4 = carbohydrates, in grams
x_5 = sodium, in mg
x_6 = calcium, in mg

a. Use backward elimination to find the best multiple linear regression model.
b. Use forward selection to find the best multiple linear regression model. Compare this model with the one in part (a). Which do you think is better? Why?
c. Consider a cup of Frosted Mini-Wheats with $x = (151.5, 1, 4, 36, 1.5, 15)$. Estimate the mean amount of fiber in one cup of this cereal using both models. If the true amount of fiber is 4 grams, which model is better?
d. Use the best subsets approach to find a multiple linear regression model. Construct a graph of r^2 versus various models.

12.242 Manufacturing and Product Development Aardvark, with stores in Santa Ana, California, and Las Vegas, Nevada, sells a wide variety of equipment and supplies used to make ceramics. Researchers at this company recently collected data in an attempt to predict the fired shrinkage of various types of clay. CLAY

The following measurements were recorded for each randomly selected clay sample:

y = percent fired shrinkage
x_1 = water absorption, percent
x_2 = pH
x_3 = indicator variable: dark clay = 0, light clay = 1

a. Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + E_i$.
b. Conduct a model utility test and the other appropriate tests to determine which variables are the most important predictors of fired shrinkage.
c. Consider a new, reduced regression model as a result of the hypothesis tests in part (b). Estimate the regression coefficients in this model.
d. Using the reduced model, estimate the mean fired shrinkage for $x_1 = 3.5$ in dark clay, and for $x_1 = 5.5$ in light clay.

12.243 Biology and Environmental Science When the last ice age ended, retreating glaciers in Norway created deep valleys that became filled with sea water. These fjords are often very deep and characterized by steep sides. In an article by Salvanes et al.,[62] the authors studied the density of *Calanus finmarchicus* (the dominant species in the zooplankton biomass) in fjords as a function of the distance from the coast. A random sample of west coast Norwegian fjords was obtained. The following measurements were recorded for each fjord. FJORDS

y = zooplankton density, in mg/m^2
x_1 = distance from the outer coast, in km
x_2 = indicator variable for season

a. Consider a regression model with the appropriate number of indicator variables. Conduct the model utility test ($\alpha = 0.05$).
b. Consider a regression model to predict $\ln(y)$ using $\ln(x_1)$ and the appropriate number of indicator variables. Find the estimated regression equation. Conduct the model utility test ($\alpha = 0.05$) and explain the meaning of each estimated regression coefficient.
c. Construct a normal probability plot of the residuals from the model in part (b). Is there any evidence to suggest that the residuals are not normally distributed?
d. Using the model in part (b), find an estimate of the mean density of zooplankton during the summer in a fjord that is 150 km from the coast.

12.244 Travel and Transportation As towns and municipalities add or improve streets, one of the biggest concerns is safety. There is an undocumented theory that narrow streets are safer than wider streets. Research was conducted to predict the safety of town streets as a function of several characteristics.[63] A random sample of streets from all across the country was obtained, and yearly accident reports were examined. To focus on street characteristics, accidents caused by road conditions (wet, icy, or snow covered), substance abuse, or traffic volume were eliminated from the study. The following variables were recorded for each street: SAFETY

y = accidents per mile per year
x_1 = degree of curvature of the street
x_2 = street width, in feet
x_3 = curb type: 0 = none, 1 = 6 inch vertical, 2 = modified
x_4 = tree density: trees per 1000 feet along the street
x_5 = number of traffic lights on the street
x_6 = mean number of vehicles per day, in thousands
x_7 = parking density: parking spaces per mile

a. Consider a regression model with six predictors and two indicator variables (for curb type). Find the estimated regression equation. Which variables do you think are significant? Justify your answer.
b. Use forward selection to find the best model ($\alpha = 0.10$). Compare the significant predictors with those identified in part (a).
c. Use backward elimination to find the best model ($\alpha = 0.10$).
d. Explain the relationship between each significant predictor variable and the accidents per mile per year.
e. Construct a normal probability plot of the residuals. Is there any evidence that the residuals are not normal? Justify your answer.

12.245 Psychology and Human Behavior Many factors affect how optimistic a person is about life in general. Family situation, relationships, and even the weather probably all have an effect. Recent research suggests that a person's optimism might be affected by the ceiling height in the home. Contractors and real estate agents frequently find it easier to sell homes with higher ceilings. A random sample of American adults living in the Northeast was obtained, and each was asked to complete a detailed survey that resulted in an Optimism Score (y), a number from 1 to 100 that measures the level of optimism people feel about themselves and the future. A large number suggests increased optimism. In addition to this score, the following variables were also recorded for each individual: POSVIEW

x_1 = height of home ceiling (ft)
x_2 = total living area of home (ft^2)
x_3 = temperature of the thermostat in the winter
x_4 = color of the walls in the main family room: 0, dark; 1, light

a. Use the techniques discussed in these two sections to find the best model to predict the optimism score.
b. Construct a normal probability plot of the residuals. Is there any evidence that the residuals are not normal?
c. Based on your model, how does an increase in ceiling height of one foot change the optimism score?

CHALLENGE

12.246 Biology and Environmental Science Consumption advisories all over the country warn people about eating certain species of fish because of elevated levels of mercury and other contaminants. Some state fishing regulations suggest that individuals should not have more than one or two meals per month of walleye or trout caught in certain lakes and rivers, because of high levels of mercury in these species. The U.S. Geological Survey is conducting a study to predict the mercury level in smallmouth bass in the Susquehanna River. Five locations (USGS stations) were used, and the following measurements were recorded for each bass: FISH

y = level of mercury, in ppm
x_1 = river flow rate, in thousands of cubic feet per second
x_2 = temperature of the water, in °C
x_3 = length of the fish, in inches
x_4 = weight of the fish, in pounds
x_5 = pH of the water
x_6 = USGS station number

a. Use the techniques discussed in this chapter to find the best model to predict the level of mercury in a smallmouth bass. Consider interaction terms and polynomial models, and be sure to use the appropriate indicator variables.
b. Use your model to suggest the *best* location to fish for smallmouth bass, that is, the spot on the river where fish tend to have the lowest levels of mercury.