

12.6 The Polynomial and Qualitative Predictor Models

The Polynomial Model

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

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

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i \quad (12.17)$$

This is still a linear regression model because the expression on the right in Equation 12.17 is a linear combination of the regression coefficients.



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

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



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:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \beta_4 x_{2i}^3 + E_i \quad (12.18)$$

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

2. There is a risk 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 only rarely. It is difficult to interpret the regression coefficients in terms of degree 4 or higher.
4. Suppose there are n observations, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. 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 12.58** shows a polynomial of degree 5 passing through six points.

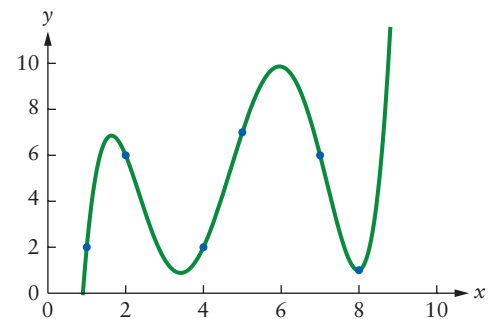


Figure 12.58 An illustration of an exact polynomial fit.

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 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 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

A recent medical study suggested that the cerebral blood flow velocity in certain patients is affected by age. 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 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

- (a) MFV (y) is the dependent variable and age (x) is the independent variable. Figure 12.59 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.

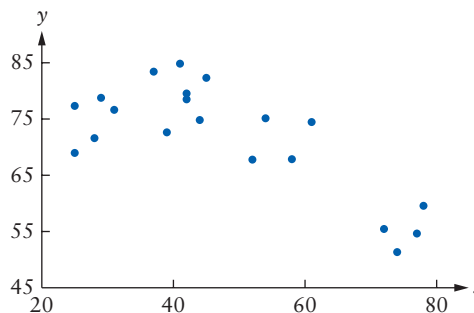


Figure 12.59 Scatter plot of MFV versus age.

The quadratic model is

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

Compute age^2 for each patient and use technology to find the estimated regression coefficients. The results are shown in Figure 12.60.

```

R
> results <- lm(y ~ x + x_sq)
> anova(results)
Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
x       1 1014.81  1014.81  42.502 5.247e-06 ***
x_sq    1   399.98   399.98  16.752 0.0007582 ***
Residuals 17  405.90    23.88

> summary(results)

Call:
lm(formula = y ~ x + x_sq)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.273356   10.420838   4.920 0.000130 ***
x            1.354821    0.438244   3.091 0.006624 **
x_sq        -0.017209    0.004205  -4.093 0.000758 ***

Residual standard error: 4.886 on 17 degrees of freedom
Multiple R-squared:  0.7771,    Adjusted R-squared:  0.7508
F-statistic: 29.63 on 2 and 17 DF,  p-value: 2.881e-06

```

Figure 12.60 The ANOVA table and estimated regression coefficients.

The estimated regression equation is $y = 51.2734 + 1.3548x - 0.0172x^2$. The ANOVA table is shown here:

ANOVA summary table

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 $r^2 = 0.7771$. 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. Here is the F test for a significant regression with $\alpha = 0.01$:

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

$$H_a: \beta_i \neq 0 \text{ for at least on } i$$

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

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

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

$$f = \frac{\text{MSR}}{\text{MSE}} = \frac{707.40}{23.88} = 29.63 (\geq 6.11)$$

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 .

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

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

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

Using the technology results in Figure 12.60,

$$t = \frac{\hat{\beta}_1 - 0}{s_{B_1}} = 3.091$$

The value of the test statistic lies in the rejection region, $|t| = |3.091| = 3.091 \geq 2.1098$ (or, equivalently, $p = 0.0066 \leq 0.05$). 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 .

$$H_0: \beta_2 = 0$$

$$H_a: \beta_2 \neq 0$$

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

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

Using the technology results in Figure 12.60, the value of the test statistic is

$$t = \frac{\hat{\beta}_2 - 0}{s_{B_2}} = -4.093$$

The value of the test statistic lies in the rejection region, $|t| = |-4.093| = 4.093 \geq 2.109$ (or, equivalently, $p = 0.0008 \leq 0.05$). There is evidence to suggest that the regression coefficient on the quadratic term, β_2 , is different from 0.

Figure 12.61 shows a graph of the estimated regression equation and a scatter plot of the data.

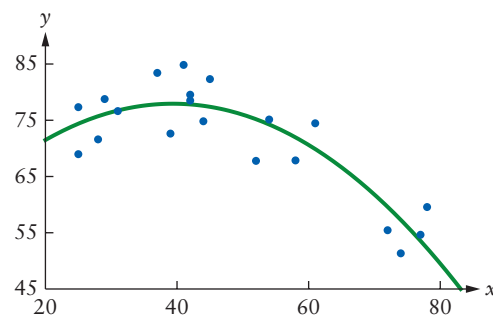


Figure 12.61 Scatter plot of MFV (y) versus age (x) and a graph of the estimated regression equation.

To check the regression assumptions, we might consider constructing a normal probability plot of the residuals along with the usual residual plots. ■

TRY IT NOW Go to Exercise 12.201

EXAMPLE 12.16 Citrus Pests

The Asian citrus psyllid, *Diaphorion 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 *Diaphorion citri* is affected by temperature. 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 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 12.62 shows a scatter plot of the data. This graph suggests a quadratic model of the form $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$.

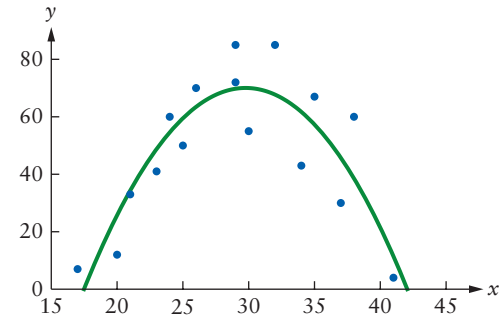


Figure 12.62 Scatter plot and graph of the estimated regression equation for the citrus pest data.

- (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°C 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 regression coefficients, the confidence interval, and the prediction interval. The results are shown in Figure 12.63.

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 12.62 also shows a graph of the regression equation.

Using the results in Figure 12.63, 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 (at 27°C).

- (b) Using the results in Figure 12.63, a 95% prediction interval for a single observation of the number of eggs when $x = 27$ (and $x^2 = 729$) is (34.59, 98.50). Notice that this prediction interval is wider than the corresponding confidence interval. ■



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

TRY IT NOW Go to Exercise 12.203



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:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + E_i \quad (12.19)$$

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

```

R
> results <- lm(y ~ x + x_sq)
> anova(results)
Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
x       1  324.4   324.4   1.6598  0.2201
x_sq    1 7168.3  7168.3  36.6735 4.06e-05 ***
Residuals 13 2541.0   195.5

> summary(results)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -341.15562    63.03799  -5.412 0.000119 ***
x              27.63192     4.48287   6.164 3.41e-05 ***
x_sq          -0.46414     0.07664  -6.056 4.06e-05 ***

Residual standard error: 13.98 on 13 degrees of freedom
Multiple R-squared:  0.7468,    Adjusted R-squared:  0.7078
F-statistic: 19.17 on 2 and 13 DF,  p-value: 0.0001328

> x_star <- data.frame(x=27,x_sq=27^2)
> predict(results,x_star,interval='confidence',level=0.95)
      fit      lwr      upr
1 66.54947 56.1144 76.98454

> predict(results,x_star,interval='prediction',level=0.95)
      fit      lwr      upr
1 66.54947 34.59398 98.50497

```

Figure 12.63 The ANOVA table, estimated regressions coefficients, confidence interval, and prediction interval.

if the first predictor variable increases by 1 (with the second predictor variable held constant). However, in Equation 12.19, if x_1 increases by 1 and x_2 is held constant, the response variable changes by $\beta_1 + \beta_3 x_2$.

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:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^3 + \beta_3 x_{2i} + \beta_4 x_{1i} x_{2i} + E_i \quad (12.20)$$

Qualitative Variables

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 El Niño is present. The regression model is

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + E_i \quad (12.21)$$

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

$$y = \beta_0 + \beta_1 x_1 \quad (12.22)$$

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

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 \quad (12.23)$$

The regression coefficient β_2 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 that a regression must account for c categories. The model is adjusted by adding $c - 1$ indicator variables. For example, in the hurricane maximum wind speed example, suppose instead that the diameter of the hurricane's 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 follows:

$$x_2 = \begin{cases} 1 & \text{if low sunspot activity} \\ 0 & \text{otherwise} \end{cases} \quad x_3 = \begin{cases} 1 & \text{if medium sunspot activity} \\ 0 & \text{otherwise} \end{cases} \quad (12.24)$$

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

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + E_i \quad (12.25)$$

EXAMPLE 12.17 Sleep Duration and Body Mass



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 variable x_2 is an indicator variable, defined later in this example. The data are given in the 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

Additional evidence suggests 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 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 coordinate axes with a different plot symbol for women and men, as in Figure 12.64.

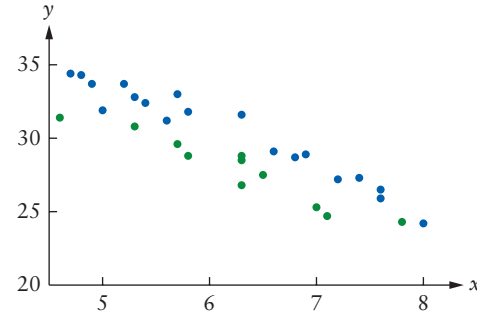


Figure 12.64 Scatter plot of BMI (y) versus sleep duration (x_1). Green points correspond to observations for women; blue points correspond to observations for men.

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.

- (b) Use technology to find the estimated regression coefficients. The results are shown in Figure 12.65.

```

R
> results <- lm(y ~ x_1 + x_2)
> anova(results)
Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
x_1    1 226.601  226.601  386.238 < 2.2e-16 ***
x_2    1  37.429   37.429   63.797 1.387e-08 ***
Residuals 27  15.841    0.587

> summary(results)

Call:
lm(formula = y ~ x_1 + x_2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.9566     0.9176  48.992 < 2e-16 ***
x_1          -2.7369     0.1422 -19.247 < 2e-16 ***
x_2           2.3205     0.2905   7.987 1.39e-08 ***

Residual standard error: 0.766 on 27 degrees of freedom
Multiple R-squared:  0.9434,    Adjusted R-squared:  0.9392
F-statistic: 225 on 2 and 27 DF,  p-value: < 2.2e-16
    
```

Figure 12.65 The ANOVA table and the estimated regression coefficients.

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

ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	F	p Value
Factor	264.03	2	132.01	225	< 0.0001
Error	15.84	27	0.59		
Total	279.87	29			

The coefficient of determination is $r^2 = 0.9434$. The model can be used to explain approximately 94% of the variation in the dependent variable.

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

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

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

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

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

Using the results in Figure 12.65, the value of the test statistic is

$$f = \frac{\text{MSR}}{\text{MSE}} = \frac{132.01}{0.59} = 225$$

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 results in Figure 12.65 suggest 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/m².

TRY IT NOW Go to Exercises 12.207 and 12.215

Section 12.6 EXERCISES

Concept Check

12.184 True or False A polynomial regression model may contain more than one predictor variable.

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

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

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

12.188 True or False All predictor variables in a multiple linear regression model must be continuous.

12.189 True or False If a polynomial regression model includes a predictor variable x of degree n ($n \geq 2$), then it must also include the predictor variable x of degree 1, 2, ..., $n - 1$.

12.190 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.

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

Practice

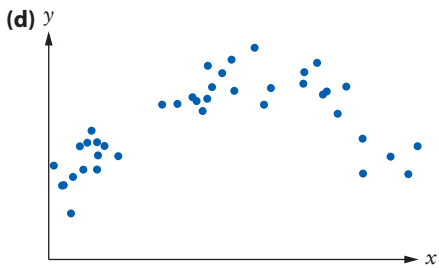
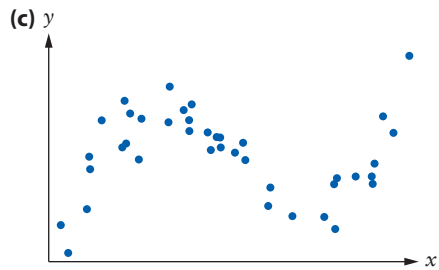
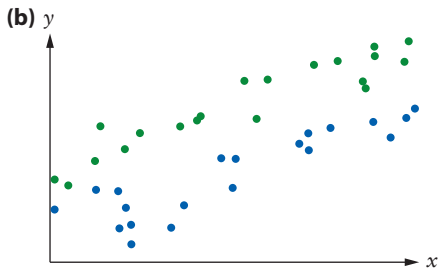
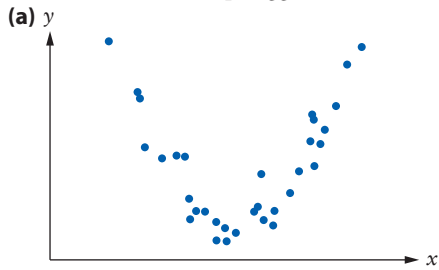
12.192 Write the regression model that includes predictor variables as described in each situation.

- (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)

12.193 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

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



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

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

12.196 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.

12.197 An experiment resulted in the observations on a dependent variable and a single independent variable given in the table. EX12.197

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.

12.198 An experiment resulted in observations on a single dependent variable and one independent variable for three categories. EX12.198

- (a) Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \beta_3x_{3i} + E_i$, where x_2 and x_3 are indicator variables. Find the coefficient of determination and interpret this value.
- (b) Conduct an F test for a significant regression. Use $\alpha = 0.01$. Use technology to find the exact p value.
- (c) Conduct the appropriate hypothesis tests to determine which predictor variables are significant.
- (d) Explain the meaning of the estimate 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 .


12.199 An experiment resulted in observations on a single dependent variable and two independent variables. EX12.199

- (a) Estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \beta_3x_{1i}x_{2i} + E_i$. 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 $Y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + E_i$. 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

12.200 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 with patients involved in traumatic accidents. Some evidence suggests that experience is related to stress, and that the relationship between these two variables is not necessarily linear. A random sample of EMTs was obtained and the number of years of experience (x) and systolic blood pressure (y , in mm Hg), an indicator of stress, were measured for each. The data are given in the table. 


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.

12.201 Public Health and Nutrition Some evidence suggests that the amount of fluoride in drinking water may affect a child's IQ level.¹ Although fluoride is recommended to strengthen teeth and fight tooth decay, evidence indicates 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. 

- (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.

12.202 Public Health and Nutrition Evidence suggests that diet soda is linked to diabetes and heart disease.² 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 consumed 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 $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$. R was used to estimate the regression coefficients, and the output is shown here. 

```

R
> results <- lm(y ~ x + x_sq)
> anova(results)
Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
x         1  6888.9   6888.9   62.001 2.007e-09 ***
x_sq      1  2117.1   2117.1   19.054 9.821e-05 ***
Residuals 37  4111.1    111.1
> summary(results)


Call:
lm(formula = y ~ x + x_sq)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.90920     6.38303   2.022  0.0504 .
x           -1.51236     0.67410  -2.244  0.0309 *
x_sq         0.07099     0.01626   4.365  9.82e-05 ***


Residual standard error: 10.54 on 37 degrees of freedom
Multiple R-squared:  0.6866, Adjusted R-squared:  0.6696
F-statistic: 40.53 on 2 and 37 DF, p-value: 4.766e-10

```


- (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.


12.203 Biology and Environmental Science The Jackson ratio is used to determine whether certain tortoises are ready for hibernation. This value is defined as the weight of the tortoise divided by the 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.  JACKSON

- (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.

12.204 Medicine and Clinical Studies Researchers have concluded that nicotine adversely affects certain regions of the brain that are responsible for attention, memory, and learning.³ 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.  NICOT

- (a) Consider a quadratic regression model for these 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.

12.205 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 8 oz 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 milligrams) 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 optimal time to brew one cup of Yerba Maté—that is, the time that yields the highest amount of manganese.

12.206 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. Consider the regression model, which includes an interaction term:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + E_i. \quad \text{RUNWAY}$$

- (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?

12.207 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). R was used to estimate the regression coefficients and the output is shown here.  DEPRESS


```

R
> results <- lm(y ~ x_1 + x_2)
> anova(results)
Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
x_1    1 1859.98  1859.98  60.2810 5.743e-10 ***
x_2    1  168.32   168.32   5.4551  0.02383 *
Residuals 47 1450.20    30.86

> summary(results)


Call:
lm(formula = y ~ x_1 + x_2)

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.7286     1.8605  18.129 < 2e-16 ***
x_1          -1.9673     0.2468  -7.970 2.82e-10 ***
x_2           3.8957     1.6680   2.336  0.0238 *

Residual standard error: 5.555 on 47 degrees of freedom
Multiple R-squared:  0.5831, Adjusted R-squared:  0.5654


```

- Construct an ANOVA summary table and conduct an F test for a significant regression. Find the exact p value associated with this test.
- Does gender affect the relationship between depression and sports involvement? Justify your answer.
- Estimate the true mean depression score for a female who participates in sports for 4 hr per week.
- 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 to avoid depression.


12.208 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 optimal distance between wax combs that promotes maximum honey production. A random sample of bee colonies was obtained, and the bee space (x , in centimeters) and the yearly honey production (y , in kilograms) were measured for each.  **HONEY**

- Construct a scatter plot for the data and write an appropriate regression model.
- Find the estimated regression coefficients.
- Conduct an F test for a significant regression. Find the exact p value.
- 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.

12.209 Biology and Environmental Science A researcher believes there is a linear relationship between the amounts of phosphorus (x , 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**

- Write the appropriate regression model and find the estimated regression coefficients.
- Is the overall regression significant? Justify your answer.
- Is there any evidence to suggest that the annual rainfall affects the relationship between phosphorus and nitrogen? Justify your answer.
- 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?

12.210 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. Oats were 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 $\mu\text{g/g}$).  **WEEDS**


- Construct a scatter plot of the data. Describe the relationship.
- Consider a simple linear regression model for these 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 these 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?

12.211 Biology and Environmental Science Grass carp are often used to control the aquatic growth in ponds. These fish are originally from Asia but the triploid grass carp was introduced in the United States in the 1960s. A random sample of grass carps was obtained and tagged. The percentage weight gain over six months (y) and the dietary lysine level (x , in g/kg diet) were measured for each.  **CARP**

- (a) Construct a scatter plot of the data. Describe the relationship.
- (b) Find the estimated regression coefficients. Is the overall regression significant? Justify your answer.
- (c) 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.
- (d) Find a 95% confidence interval for the mean percentage weight gain for $x = 14$ g/kg.
- (e) Use the regression equation to estimate the lysine level that produces the maximum weight gain.

Extended Applications

12.212 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 follows:  **QUAKE**


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

Consider the regression model


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

- (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?

12.213 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.


12.214 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.  **GRASS**

- (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?

12.215 Biology and Environmental Science 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 $\mu\text{g/L}$) were measured.  **ALUMIN**

- (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.

12.216 Education and Child Development Some evidence suggests that the diet of a woman while pregnant may affect the weight, or tendency toward obesity, of her child. One theory indicates that if a woman's diet is high in fat, then the child may be destined for adult obesity.⁴ 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 , in 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 gender 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 are defined in the table.

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

R was used to estimate the regression coefficients, and a portion of the results are shown here.

```

> results <- lm(y ~ x_1 + x_2 + x_3 + x_4 + x_5)
> anova(results)
Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
x_1     1   939.70      939.70  15.287 < 2e-16 ***
x_2     1    65.87      65.87   2.999  0.00369 **
x_3     1    23.72      23.72   1.006  0.31749
x_4     1    14.83      14.83   1.549  0.12561
x_5     1    25.68      25.68   2.480  0.01542 *
Residuals 74   309.01

> summary(results)

Call:
lm(formula = y ~ x_1 + x_2 + x_3 + x_4 + x_5)

Coefficients:
(Intercept)  9.77163  1.03128  9.475 2.12e-14 ***
x_1          0.17033  0.01114 15.287 < 2e-16 ***
x_2         -1.91001  0.63698  -2.999  0.00369 **
x_3         -0.64441  0.64029  -1.006  0.31749
x_4          0.99210  0.64041   1.549  0.12561
x_5          1.17486  0.47375   2.480  0.01542 *

Residual standard error: 2.043 on 74 degrees of freedom
Multiple R-squared:  0.7759, Adjusted R-squared:  0.7607

```

- (a) Complete the ANOVA table and conduct an F test for a significant regression with $\alpha = 0.01$.
- (b) Does gender 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 g.
- (d) How does the regression line shift from a White male to an African American female? To an Asian American female? Justify your answers.


12.217 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 decibels, dB), and the output power from the feed line (y , in watts) were measured for each. Consider the regression model (with an interaction term)

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + E_i. \quad \text{RADIO}$$

- (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?

12.218 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:  **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) Consider 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 Problems

12.219 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, including total number of people who have had the flu (y), and the days since the first reported case (x). On the first day of reported flu cases, day 0, 160 people had the flu.  **SPREAD**

- (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.

12.220 Biology and Environmental Science The median-median line is an alternative to the least-squares regression line. Given n pairs of observations, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 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 the points $(\tilde{x}_1, \tilde{y}_1)$ and $(\tilde{x}_3, \tilde{y}_3)$. (Remember how to find the equation of a line through two points?)
- (d) Adjust this line one-third of the distance to $(\tilde{x}_2, \tilde{y}_2)$. This is the median-median line.

Results from a recent study suggest 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.

- (a) Construct a scatter plot of the data. Describe the relationship between elevation and vessel diameter.  **MAPLE**
- (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.