$$\hat{y} \approx 11.695 - 1.620x_2 - 63.179x_5$$

is the same as that obtained using Minitab.

15.5.2 Testing for Significance of the Coefficient of Determination

The proper test for checking the significance of a given R^2 when the model contains k factors is

$$F_{k,N-k-1} = \frac{R^2/k}{(1-R^2)/(N-k-1)} \tag{15.28}$$

This is the statistic for the test of H_0: The coefficients of the predictor variables in the model are all zero. The test is valid when data are collected to check the appropriateness of a specific model.

When exploratory procedures are applied in the selection of a "best" model, F tests are often used as guides in deciding the next step. To decide whether any more variables are needed after a subset of r variables has been found significant, for example, one can use

$$F_{k-r,N-k-1} = \frac{(R_k^2 - R_r^2)/(k-r)}{(1-R_k^2)/(N-k-1)} \tag{15.29}$$

where R_k^2 is the coefficient of multiple determination for the model containing the entire set of k independent variables and R_r^2 is that coefficient for a subset of r of the k independent variables. Usually, no more variables are added to the model when the p value of the test is large.

Example 15.5 (Model Tests for the Elongation Study)

From Figure 15.13, the coefficients of multiple determination are $R_5^2 = 0.51314320$ for the five-variable model and $R_2^2 = 0.49208798$ for the model containing only X_2 and X_5.

To test the five-variable model, we use Equation (15.28) and find

$$F_{5,18} = \frac{0.51314320/5}{(1-0.51314320)/18} = 3.79$$

as in ANOVA. If the original intent was to use the five-variable model, we conclude that at least one of the five independent variables in the model has a nonzero coefficient.

When deciding whether to add variables other than X_2 and X_5 to the linear model, we use Equation (15.29) and find

$$F_{3,18} = \frac{(0.51314320 - 0.49208798)/3}{(1-0.51314320)/18} = 0.2695$$

Since $P(F_{3,18} > 0.2695) = 0.8465$, one may decide not to go beyond X_2 and X_5.

15.6 SUMMARY

The examples of this chapter give the basic adequate equations. The methods may be augmented to include terms in multiple variables that are more complex. For example, with three variables (X_1, X_2, and X_3) one might create new variables such as $X_4 = X_1^2$, $X_5 = X_1X_2^2$, and $X_6 = X_3^3$, allowing for prediction surfaces that go beyond the rectilinear model of Equation (15.24). So there is much more that can be done when the X's are quantitative.

The following practical steps summarize some of the procedures discussed in this chapter.

CASE I. Polynomial regression when X is set at equispaced levels.

1. Plot a rough scattergram of y versus x to get an idea of the underlying polynomial.
2. Determine the treatment totals for each level of X.
3. Use the table of orthogonal polynomials to test for the highest degree polynomial for an adequate prediction equation. This step may not be needed with certain computer programs.
4. Use appropriate computer software to find this adequate equation.
5. Note the size of R^2 to see how much of the variation in Y is associated with the proposed linear model and whether this is worthwhile (even though statistically significant) from a practical point of view. Include a residual analysis to determine the appropriateness of the model.
6. If the model is deemed appropriate, note the standard error of estimate (square root of MS_{error}) and place confidence limits around the predictions, if desirable.

CASE II. Polynomial regression when the levels of X are not equispaced—usually an ex-post-facto situation.

1. Plot a rough scattergram to help in estimating the highest degree that might be used.
2. Run a computer program to find this highest degree equation and note if all B's are statistically significant at some desired level.
3. If some B's are not significant, try a polynomial of one degree less and check the B's.
4. Continue as in step 3 until an adequate prediction model is found.
5. Based on the value of R^2 and a thorough residual analysis, decide whether the tentative regression equation should be recommended for use. A second experiment may be necessary to check the adequacy of the proposed model.
   Caution: When writing the model for a given polynomial, be sure to include as variables all lower powers of X as well.

	x				
	3	4	5	6	7
y	7	8	10	11	10
	8	8	9	9	10
	9	9	9	10	9

15.2 For Problem 15.1 present your results in an ANOVA table and comment on these results.

15.3 From your table in Problem 15.2 find r^2, η^2, and the standard error of estimate.

15.4 Use the method of orthogonal polynomials on the data of Problem 15.1 and find the best fitting polynomial.

15.5 An experiment to determine the effect of planting rate (X) in thousands of plants per acre on yield (Y) in bushels per acre of corn gave the following results.

Planting Rate	Yield
12	130.5, 129.6, 129.9
16	142.5, 140.3, 143.4
20	145.1, 144.8, 144.1
24	147.8, 146.6, 148.4
28	134.8, 135.1, 136.7

Plot a scattergram of these data and comment on the degree polynomial that might prove appropriate for predicting yield from planting rate.

15.6 From the data of Problem 15.5, use orthogonal polynomials and determine what degree polynomial is appropriate here.

15.7 Find the equation of the polynomial in Problem 15.6.

15.8 From an ANOVA table of the results of Problem 15.6, find r^2, R^2, and η^2, and comment on what these statistics tell you.

15.9 Seasonal indexes of hog prices taken in two different years for five months gave the following results.

	Month				
Year	1	2	3	4	5
1	93.9	94.3	101.2	96.7	107.5
2	95.4	96.5	94.7	97.1	107.2

a. Sketch a scatterplot of these data and comment on what type of curve might be used to predict hog prices when the month is known.
b. Do a complete analysis of these data and find the equation of the best fitting curve for predicting hog price from month.

15.10 A bakery is interested in the effect of six baking temperatures on the quality of its cakes. The bakers agree on a measure of cake quality, Y. A large batch of cake mix is prepared and 36 cakes are poured. The baking temperature is assigned at random to the cakes such that six cakes are baked at each temperature. Results are as follows:

Temperature (°F)					
175	**185**	**195**	**205**	**215**	**225**
22	4	10	22	32	20
10	21	22	14	27	36
18	18	18	21	22	33
26	28	32	25	37	33
21	21	28	26	27	20
21	28	25	25	31	25

Do a complete analysis of these data and write the "best" fitting model for predicting cake quality from temperature.

15.11 Compute statistics that will indicate how good your fit is in Problem 15.10 and comment. Include a residual analysis.

15.12 Here are some data on the effect of age on retention of information.

Ages:	6–14	15–23	24–32	33–41
Total score of 10 people, $T_{\cdot j}$:	80	92	103	90

Source	**df**	**SS**	**MS**
Between ages	3	26.675	8.892
Linear	1	8.405	8.405
Quadratic	1	15.625	15.625
Cubic	1	2.645	2.645
Error	36	72.000	2.000
Total	39	98.675	

a. Determine the significance of each prediction equation and state which degree equation (linear, quadratic, or cubic) will adequately predict retention score from age.
b. Write the equation for part a.
c. Find the departure from regresson of the mean retention score of a 37-year-old based on your answer to part b.

15.13 A study of the effect of chronological age on history achievement scores gave the following results.

History Achievement Score, Coded by $(Y - 18)/15$	Chronological Age				
	8	**9**	**10**	**11**	**12**
5			1	6	5
4		1	7	4	5
3		6	2		
2	3	2			
1	6	1			
0	1				

Do a complete analysis of justifying your curve of best fit.

15.14 For a study of the effect of the distance a road sign stands from the edge of the road on the amount by which a driver swerves from the edge of the road, five distances were chosen. Then the five distances were set at random, and four cars were observed at each set distance. The results gave

$$\eta^2 = 0.70$$

$$r^2 = 0.61$$

a. Test whether a linear function will "fit" these data. (You may assume a total SS of 10,000 if you wish, but this is not necessary.)
b. In fitting the proper curve, what are you assuming about the four readings at each specified distance?

15.15 Thickness of a film is studied to determine the effect of two types of resin and three gate settings on the thickness.

	Gate Setting (mm)		
Resin Type	**2**	**4**	**6**
1	1.5	2.5	3.6
	1.3	2.5	3.8
2	1.4	2.6	3.7
	1.3	2.4	3.6

Do an analysis of variance of the results above, following the methods of Chapter 5.

15.16 For Problem 15.15, extract a linear and a quadratic effect of gate setting and test for significance. Also partition the interaction and test its components.

15.17 Write any appropriate prediction equation or equations from the data of Problem 15.15.

15.18 The experiment of Problem 15.15 is extended to include a third factor, weight fraction at three levels, as indicated by the accompanying tabulation. Compile an ANOVA table for this three-factor experiment as in Chapter 5.

Gate Setting (mm)	Resin Type 1 Weight Fraction 0.20	0.25	0.30	Resin Type 2 Weight Fraction 0.20	0.25	0.30
2	1.6	1.5	1.5	1.5	1.4	1.6
	1.5	1.3	1.3	1.4	1.3	1.4
4	2.7	2.5	2.4	2.4	2.6	2.2
	2.7	2.5	2.3	2.3	2.4	2.1
6	3.9	3.6	3.5	4.0	3.7	3.4
	4.0	3.8	3.4	4.0	3.6	3.3

15.19 Outline a further breakdown of the ANOVA table in Problem 15.18 based on how the independent variables are set. Show the proper degrees of freedom.

15.20 For the significant ($\alpha = 0.05$) effect in Problem 15.18 complete the ANOVA breakdown suggested in Problem 15.19.

15.21 Graph any significant effects found in Problem 15.20 and discuss.

15.22 Data on the effect of knife-edge radius R in inches and feedroll force F in pounds per inch on the energy necessary to cut 1-inch lengths of alfalfa are given as follows:

Feedroll Force F_j (lb/in.)	Knife-Edge Radius, R_i (in.) 0.000	0.005	0.010	0.015	$T_{j.}$
5	29	98	44	84	
	30	128	81	100	
	20	67	77	63	
	79	293	202	247	821
10	22	35	53	103	
	26	80	93	90	
	16	29	59	98	
	64	144	205	291	704
15	18	49	58	80	
	17	68	103	91	
	11	61	128	77	
	46	178	289	248	761

20	38		68		87		86	
	31		74		116		113	
	21		47		90		81	
		90		189		293	280	852
$T_{i..}$		279		804		989	1066	$T_{...} = 3138$

Write the mathematical model for this experiment and do an ANOVA on the two factors and their interaction.

15.23 Outline a further ANOVA based on Problem 15.22 with single-degree-of-freedom terms.

15.24 Do an analysis for each of the terms in Problem 15.23.

15.25 Plot some results of Problem 15.22 and try to argue that they confirm some results found in Problem 15.24.

15.26 The sales volume Y of a product in thousands of dollars, price X_1 per unit in dollars, and advertising expense X_2 in hundreds of dollars were recorded for $n = 8$ cases. The results are as follows.

	Case							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
y	10.1	6.5	5.1	11.0	9.9	14.7	4.8	12.2
x_1	1.3	1.9	1.7	1.5	1.6	1.2	1.6	1.4
x_2	8.8	7.1	5.5	13.8	18.5	9.8	6.4	10.2

a. Find the means and standard deviations of each variable and the intercorrelation matrix.
b. On the basis of the discussion in Section 15.2, find the regression equation Y from X_1 alone.

15.27 In studying the percent conversion (Y) in a chemical process as a function of time in hours (X_1) and average fusion temperature in degrees Celsius (X_2) the following data were collected.

y	x_1	x_2
62.7	3	297.5
76.2	3	322.5
80.8	3	347.5
80.8	6	297.5
89.2	6	322.5
78.6	6	347.5
90.1	9	297.5
88.0	9	322.5
76.1	9	347.5

Analyze these data using multiple linear regression. Reexamine the data and comment on any peculiarities noted.

15.28 A management consulting firm attempted to predict annual salary of executives from the executives' years of experience (X_1), years of education (X_2), sex (X_3), and number of employees supervised (X_4). A sample of 25 executives gave an average salary of \$79,700 with a standard deviation of \$1300. From a computer program the following statistics were recorded.

$$r^2_{Y4} = 0.42$$

$$R^2_{Y.24} = 0.78$$

$$R^2_{Y.124} = 0.90$$

$$R^2_{Y.1234} = 0.95$$

a. Explain how you would interpret these statistics.
b. Test whether one could stop after variables 2 and 4 have been entered into the equation.
c. If variables 1, 2, and 4 were used in the prediction equation, what would the limits on the salaries expected for a predicted salary have to be in order to be correct about 95% of the time?

15.29 Consider a prediction situation in which some dependent variable Y is to be predicted from four independent variables X_1, X_2, X_3, and X_4 and you have a stepwise regression printout on this problem involving 25 observations. Explain briefly each of the following with respect to this problem and its printout.
a. How does the computer decide which independent variable to enter first into a regression equation?
b. How does it decide which variable to enter next in the equation?
c. How do you decide when to stop adding variables based on the printout?
d. Do you need to add any more variables if $R^2_{Y.1234} = 0.7$ and $R^2_{Y.24} = 0.6$?

15.30 A study was made to determine the effect of four variables on the grade-point average of 55 students who attended a junior college after graduation from high school. These variables were high school rank (X_1), SAT score (X_2), IQ (X_3), and age in months (X_4). From the computer printout the following statistics were recorded.

$$R^2_{Y.1234} = 0.45 \qquad R^2_{Y.12} = 0.33$$

$$R^2_{Y.123} = 0.42 \qquad r^2_{Y1} = 0.19$$

Use this information to test for the significance of each added variable and determine whether an ordered subset of these four will make a satisfactory prediction. If you wish, you may assume a total of sum of squares in this study of 10,000 units. Also explain how you would predict with approximate 2 : 1 odds the maximum GPA expected for an individual student based on his or her scores of the variables in the appropriate equation.

15.31 In a study of several variables that might affect freshman GPA (Y) a sample of 55 students reported their scores on the college entrance arithmetic test (X_1), and the analogies test on an Ohio battery (X_2), as well as their high school average (X_3) and their interest score on

an interest inventory (X_4). Results of this study are given in the following (oversimplified) ANOVA table.

Source*	df	SS	MS
Due to X_2	1	1360	1360
Due to X_3/X_2	1	480	480
Due to $X_1/X_2, X_3$	1	80	80
Due to $X_4/X_1, X_2, X_3$	1	80	80
Error	50	2000	40
Total	54	4000	

* The symbol/means "given."

Assuming a stepwise procedure was used so that variables were entered in their order of importance, answer the following.

a. Determine $R_{Y.1234}$ and explain its meaning in this problem.
b. Determine which of the four variables are sufficient to predict GPA (Y) and show that your choice is sufficient.
c. If a regression equation based on your answer to part b is used to predict student A's GPA based on test scores, how close do you think this prediction will be to the actual GPA? Justify your answer and note any assumptions you are making.

15.32 In an attempt to predict the average value per acre of farm land in Iowa in 1920, the following data were collected.

Y = average value in dollars per acre
X_1 = average corn yield in bushels per acre for 10 preceding years
X_2 = percentage of farmland in small grain
X_3 = percentage of farmland in corn

y	x_1	x_2	x_3	y	x_1	x_2	x_3
87	40	11	14	193	41	13	28
133	36	13	30	203	38	24	31
174	34	19	30	279	38	31	35
385	41	33	39	179	24	16	26
363	39	25	33	244	45	19	34
274	42	23	34	165	34	20	30
235	40	22	37	257	40	30	38
104	31	9	20	252	41	22	35
141	36	13	27	280	42	21	41
208	34	17	40	167	35	16	23
115	30	18	19	168	33	18	24
271	40	23	31	115	36	18	21
163	37	14	25				

Use multiple linear regression to analyze these data, and comment on the results.

15.33 In a study involving handicapped students researchers attempted to predict GPA from two demographic variables, sex and ethnicity, and two independent measured variables, interview score and contact hours used in counseling and/or tutoring. A dummy variable for sex was taken as 1 = male, 2 = female. For ethnicity: 1 = African-American, 2 = Hispanic, and 3 = white, non-Hispanic. Do a complete analysis of the data below and justify the regression equation that will adequately "fit" the data.

Students	Ethnic	Sex	Interview	Hours	GPA
1	1	2	11.0	4.0	5.50
2	1	2	10.0	5.0	4.10
3	1	2	12.0	73.0	5.00
4	1	2	11.5	68.0	4.22
5	1	2	10.8	82.0	5.00
6	1	1	12.5	72.5	5.00
7	1	1	9.5	64.0	4.60
8	1	1	9.5	78.0	4.25
9	1	1	8.0	64.0	4.00
10	1	1	7.5	13.0	2.00
11	2	2	9.0	37.0	4.25
12	2	2	8.2	4.0	4.00
13	2	2	10.7	38.5	4.61
14	2	2	8.5	3.0	2.93
15	2	2	12.5	10.5	5.50
16	2	1	12.0	80.0	4.77
17	2	1	12.2	6.0	5.00
18	2	1	7.0	6.5	3.25
19	2	1	8.6	22.0	2.66
20	2	1	8.3	28.5	3.37
21	3	2	10.9	12.0	5.00
22	3	2	9.0	9.0	4.00
23	3	2	10.0	5.0	5.00
24	3	2	7.2	12.0	3.87
25	3	2	8.5	4.0	3.00
26	3	1	10.0	8.0	4.77
27	3	1	8.5	8.0	5.00
28	3	1	10.0	22.0	5.08
29	3	1	11.4	61.5	5.57
30	3	1	11.9	37.0	6.00

15.34 In a study on the effect of several variables on posttest achievement scores of 48 biology students in a rural high school, the following table was presented. It included variables that had a significant effect on achievement at the 10% level of significance.

TABLE F
Coefficients of Orthogonal Polynomials

		j										
k	**Polynomial**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	$\sum \xi_j^2$
3	Linear	−1	0	1								2
	Quadratic	1	−2	1								6
	Linear	−3	−1	1	3							20
4	Quadratic	1	−1	−1	1							4
	Cubic	−1	3	−3	1							20
	Linear	−2	−1	0	1	2						10
5	Quadratic	2	−1	−2	−1	2						14
	Cubic	−1	2	0	−2	1						10
	Quartic	1	−4	6	−4	1						70
	Linear	−5	−3	−1	1	3	5					70
6	Quadratic	5	−1	−4	−4	−1	5					84
	Cubic	−5	7	4	−4	−7	5					180
	Quartic	1	−3	2	2	−3	1					28
	Linear	−3	−2	−1	0	1	2	3				28
7	Quadratic	5	0	−3	−4	−3	0	5				84
	Cubic	−1	1	1	0	−1	−1	1				6
	Quartic	3	−7	1	6	1	−7	3				154
	Linear	−7	−5	−3	−1	1	3	5	7			168
	Quadratic	7	1	−3	−5	−5	−3	1	7			168
8	Cubic	−7	5	7	3	−3	−7	−5	7			264
	Quartic	7	−13	−3	9	9	−3	−13	7			616
	Quintic	−7	23	−17	−15	15	17	−23	7			2184
	Linear	−4	−3	−2	−1	0	1	2	3	4		60
	Quadratic	28	7	−8	−17	−20	−17	−8	7	28		2772
9	Cubic	−14	7	13	9	0	−9	−13	−7	14		990
	Quartic	14	−21	−11	9	18	9	−11	−21	14		2002
	Quintic	−4	11	−4	−9	0	9	4	−11	4		468
	Linear	−9	−7	−5	−3	−1	1	3	5	7	9	330
	Quadratic	6	2	−1	−3	−4	−4	−3	−1	2	6	132
10	Cubic	−42	14	35	31	12	−12	−31	−35	−14	42	8580
	Quartic	18	−22	−17	3	18	18	3	−17	−22	18	2860
	Quintic	−6	14	−1	−11	−6	6	11	1	−14	6	780