When the levels of X are equispaced, as is the case with the reaction time data, solution of Equation (15.18) is simplified by using the coding of Equation (15.12). This gives

$$y'_u = b'_0 + b'_1 u + b'_2 u^2 \tag{15.19}$$

$$\left.\begin{aligned}
\sum_i \sum_j y_{ij} &= b'_0 N + b'_1 \sum_j n_j u_j + b'_2 \sum_j n_j u_j^2 \\
\sum_i \sum_j u_j y_{ij} &= b'_0 \sum_j n_j u_j + b'_1 \sum_j n_j u_j^2 + b'_2 \sum_j n_j u_j^3 \\
\sum_i \sum_j u_j^2 y_{ij} &= b'_0 \sum_j n_j u_j^2 + b'_1 \sum_j n_j u_j^3 + b_2 \sum_j n_j u_j^4
\end{aligned}\right\} \tag{15.20}$$

Because of the choice of the u_j's, each term in Equation (15.20) that involves odd powers of the u_j's equals zero. So, the equations become

$$\left.\begin{aligned}
\sum_i \sum_j y_{ij} &= b'_0 N + b'_2 \sum_j n_j u_j^2 \\
\sum_i \sum_j u_j y_{ij} &= b'_1 \sum_j n_j u_j^2 \\
\sum_i \sum_j u_j^2 y_{ij} &= b'_0 \sum_j n_j u_j^2 + b'_2 \sum_j n_j u_j^4
\end{aligned}\right\} \tag{15.21}$$

Example 15.4 (Least Squares Parabola for Reaction Time Data)

Consider again the reaction time data of Table 15.1. There are three observations at each of the five dosage levels so $N = 3 \times 5 = 15$ and $n_j = 3$ for $j = 1, 2, 3, 4$, and 5. From Table 15.3, $T_{.1} = 83$, $T_{.2} = 84$, $T_{.3} = 89$, $T_{.4} = 96$, and $T_{.5} = 115$. So,

$$\sum_i \sum_j y_{ij} = \sum_j T_{.j} = 83 + 84 + 89 + 96 + 115 = 467$$

From Example 15.3, $u_1 = -2$, $u_2 = -1$, $u_3 = 0$, $u_4 = 1$, and $u_5 = 2$. Thus,

$$\begin{aligned}
\sum_i \sum_j u_j y_{ij} &= \sum_j u_j T_{.j} = (-2)(83) + (-1)(84) + (0)(89) + (1)(96) \\
&\quad + (2)(115) = 76 \\
\sum_i \sum_j u_j^2 y_{ij} &= \sum_j u_j^2 T_{.j} = (-2)^2(83) + (-1)^2(84) + (0)^2(89) + (1)^2(96) \\
&\quad + (2)^2(115) = 972 \\
\sum_j n_j u_j^2 &= (3)[(-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2] = 3(10) = 30 \\
\sum_j n_j u_j^4 &= (3)[(-2)^4 + (-1)^4 + (0)^4 + (1)^4 + (2)^4] = 3(34) = 102
\end{aligned}$$

A SAS Program

A SAS command file that produces a least squares line for the reaction time data was given in Table 15.4. We can have SAS calculate a least squares polynomial for these data by defining a new variable, the square of the dosage, and including that variable in the model statement. This is illustrated in Table 15.7. Except for rounding, the resulting model (see Figure 15.5) is the same as that obtained manually and using JMP.

15.3.1 Departure from the Quadratic Model: A Lack of Fit Test

The parabola of Figure 15.4 fits the reaction time data better than the line of Figure 15.1. This is evidenced by our visual perception as well as by the observation that the coefficient of determination for the quadratic model ($R^2 = 0.908626$) is markedly larger than that for the

TABLE 15.7
SAS Command File for Example 15.4

```
OPTIONS LINESIZE=80;
DATA REACTION;
INPUT DOSE TIME @@;
CARDS;
0.5 26 0.5 28 0.5 29 1.0 28 1.0 26 1.0 30 1.5 28 1.5 30 1.5 31
2.0 32 2.0 33 2.0 31 2.5 38 2.5 39 2.5 38
;
DOSE2 = DOSE*DOSE;
PROC GLM;
MODEL TIME = DOSE DOSE2;
```

```
                    GENERAL LINEAR MODELS PROCEDURE
DEPENDENT VARIABLE: TIME

                                    SUM OF          MEAN
SOURCE                      DF      SQUARES        SQUARE      F VALUE
MODEL                        2  226.91428571  113.45714286        59.66
ERROR                       12   22.81904762    1.90158730      PR > F
CORRECTED TOTAL             14  249.73333333                    0.0001

R-SQUARE                 C.V.      ROOT MSE       TIME MEAN
0.908626               4.4293    1.37898053     31.13333333

SOURCE                      DF      TYPE I SS      F VALUE      PR > F
DOSE                         1  192.53333333       101.25      0.0001
DOSE2                        1   34.38095238        18.08      0.0011

SOURCE                      DF     TYPE III SS     F VALUE      PR > F
DOSE                         1    9.41339445         4.95      0.0460
DOSE2                        1   34.38095238        18.08      0.0011

                                  T FOR H0:      PR > |T|   STD ERROR OF
PARAMETER              ESTIMATE   PARAMETER=0                   ESTIMATE
INTERCEPT          29.86666667         17.49      0.0001    1.70756177
DOSE               -5.79047619         -2.22      0.0460    2.60255133
DOSE2               3.61904762          4.25      0.0011    0.85112526
```

Figure 15.5 SAS output for Example 15.4.

linear model ($r^2 = 0.770956$). We now turn our attention to a formal test of the "goodness" of the fit of the quadratic model. Before proceeding, recall that eta squared is 0.9199 for these data, indicating that the variabilty in reaction times accounted for by our quadratic model is near the maximum amount that can be accounted for by a curve passing through the five sample means.

From Example 15.3, the quadratic model for the reaction time data is

$$y'_u = \frac{3079}{105} + \frac{38}{15}u + \frac{19}{21}u^2 = \frac{1}{105}(3079 + 266u + 95u^2)$$

when expressed in terms of U values or

$$y'_x = \frac{1}{105}(3136 - 608x + 380x^2)$$

when expressed in terms of X values. Either can be used to see how well the model predicts the Y's from the X's. Table 15.8 shows the predicted Y's and the departure of the means from the quadratic model.

Since each departure of a mean must be weighted with three observed values for each value of X, we can write the sum of the squares of the departure from the quadratic from Table 15.8 as follows:

$$\text{SS}_{\text{departures}} = \sum_j n_j(\bar{y}_{\cdot j} - y'_{x_j})^2 = 3\left(\frac{10{,}360}{11{,}025}\right) = \frac{296}{105} \approx 2.82$$

This step can be added to refine Table 15.6 further and give Table 15.9. The departure from the quadratic is not significant at the 5% level ($p = 0.519$), so it is appropriate to stop with the quadratic for predicting reaction time from dosage.

Note that the error and nonsignificant departure term could be pooled, giving the error term in the SAS summary of Figure 15.5. Using this error term, the *standard error of estimate* for the quadratic model is

$$s_{Y.X,X^2} = \sqrt{\frac{20.00 + 2.82}{10 + 2}} = 1.38$$

which might be used to set confidence limits on the Y's around the quadratic curve.

TABLE 15.8
Departures from Quadratic

x_j	u_j	$\bar{y}_{\cdot j}$	y'_x	$\bar{y}_{\cdot j} - y'_x$	$(\bar{y}_{\cdot j} - y'_x)^2$
0.5	−2	83/3	2927/105	−22/105	484/11,025
1.0	−1	84/3	2908/105	32/105	1,024/11,025
1.5	0	89/3	3079/105	36/105	1,296/11,025
2.0	1	96/3	3440/105	−80/105	6,400/11,025
2.5	2	115/3	3991/105	34/105	1,156/11,025
Totals				0	10,360/11,025

TABLE 15.9
ANOVA on Reaction Time with Quadratic Regression

Source	df	SS	MS	*F*	*p* value
Between dosages	4	229.73			
Linear	1	192.53	192.53	96.3	0.000
Quadratic	1	34.38	34.38	17.2	0.002
Departure from model	2	2.82	1.41	0.7	0.519
Error	10	20.00	2.00		
Totals	14	249.73			

15.3.2 Two Factors: One Qualitative and One Quantitative

We now consider how to determine the effects of depth (D) and position (P) in a tank on the concentration (Y) of a cleaning solution. Concentrations are measured at three depths from the surface of the tank: 0, 15, and 30 inches. At each depth, measurements are taken at five different lateral positions in the tank, as depicted in Figure 15.6. These are considered as five qualitative positions, although some orientation measure probably might be made on them. At each depth and position, two observations are taken. This is then a 5×3 factorial with two replications per cell (total of 30 observations). The data, collected in random order, are summarized in Table 15.10.

Analyzing the data as if both factors are qualitative, we have the model

$$Y_{ijk} = \mu + D_i + P_j + DP_{ij} + \varepsilon_{k(ij)}$$

with $i = 1, 2, 3$; $j = 1, 2, 3, 4, 5$; and $k = 1, 2$. As Figure 15.7 indicates, only the depth effect is significant.

Even though the interaction between depth and position is not significant, there may be an interaction between the linear effect of depth and positions or between the quadratic effect of depth and positions. To check for such interactions, we will fit a full model using the **GLM** procedure in SAS with the **MODEL** statement

```
MODEL CONC=DEPTH DEPTH*DEPTH POS DEPTH*POS DEPTH*DEPTH*POS;
```

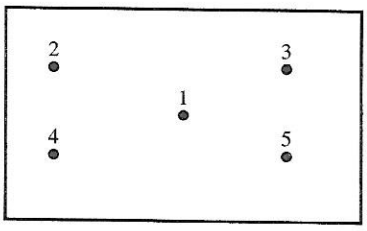


Figure 15.6 Positions in tank at each depth.

TABLE 15.10
Cleaning Solution Concentration Data

	Depth from Top of Tank D_i (in.)		
Position P_j	**0**	**15**	**30**
1	5.90	5.90	5.94
	5.91	5.89	5.80
2	5.90	5.89	5.75
	5.91	5.89	5.83
3	5.94	5.91	5.86
	5.90	5.91	5.83
4	5.93	5.94	5.83
	5.91	5.90	5.89
5	5.90	5.94	5.83
	5.87	5.90	5.86

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	0.0390	0.0028	2.2135
Error	15	0.0189	0.0013	Prob > F
C Total	29	0.0579		0.0694
Source	**DF**	**Sum of Squares**	**F Ratio**	**Prob > F**
depth	2	0.0282	11.1772	0.0011
pos	4	0.0050	1.0013	0.4374
depth*pos	8	0.0058	0.5787	0.7802

Figure 15.7 JMP ANOVA for Table 15.10 data.

Here POS is a classification variable. Coding the data by subtracting 5.90 and multiplying by 100 gives Figure 15.8.

Only the depth factor (linear and quadratic) from the SAS analysis of Figure 15.8 needs to be considered, since the other factors are not significant. [**Note:** The Type I sums of squares are used.] The plot of the sample means versus depth (Figure 15.9) confirms the appropriateness of a quadratic model.

Using JMP with the Table 15.10 data to fit a second-degree polynomial gives Figure 15.10. From that figure, the least squares equation is

$$y' = 5.9070 + 0.0022d - 0.0001d^2 \tag{15.22}$$

which accounts for about 49% of the variation in concentration. The large error variance leads us to believe that the experimenters failed to consider at least one factor that has a significant effect on that variability. Further, the scattergram seems to indicate that the concentration readings at a depth of 30 inches vary more than those at the other depths.

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: CONC

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE
MODEL	14	390.46666667	27.89047619	2.21
ERROR	15	189.00000000	12.60000000	PR > F
CORRECTED TOTAL	29	579.46666667		0.0694

R-SQUARE	C.V.	ROOT MSE	CONC MEAN
0.673838	242.0214	3.54964787	-1.46666667

SOURCE	DF	TYPE I SS	F VALUE	PR > F
DEPTH	1	211.25000000	16.77	0.0010
DEPTH*DEPTH	1	70.41666667	5.59	0.0320
POS	4	50.46666667	1.00	0.4374
DEPTH*POS	4	41.50000000	0.82	0.5303
DEPTH*DEPTH*POS	4	16.83333333	0.33	0.8508

SOURCE	DF	TYPE III SS	F VALUE	PR > F
DEPTH	1	16.25000000	1.29	0.2739
DEPTH*DEPTH	1	70.41666667	5.59	0.0320
POS	4	16.60000000	0.33	0.8539
DEPTH*POS	4	14.73076923	0.29	0.8794
DEPTH*DEPTH*POS	4	16.83333333	0.33	0.8508

Figure 15.8 SAS output for concentration data.

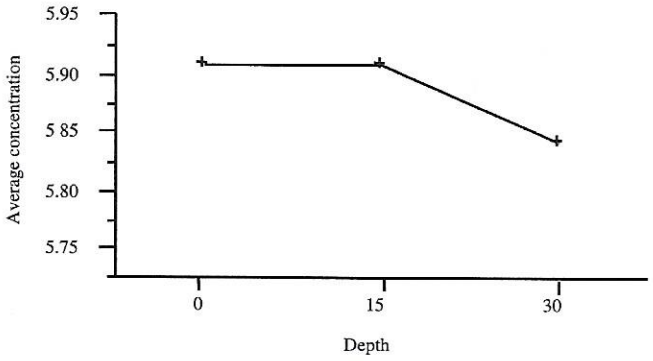
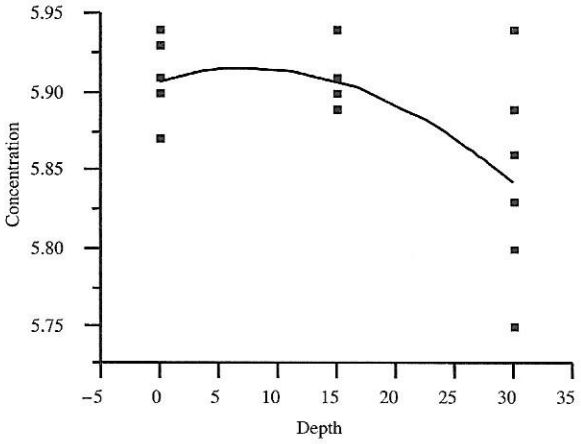


Figure 15.9 Means plot for concentration data.

Two Additional Comments

In the concentration study, the interaction effect was not significant. Had there been a significant interaction, we would have had to find a least squares equation in depth for each position. Had position been significant with no interaction, Equation (15.22) could have been used, with a constant added for each position.



Analysis of Variance

Source	DF	SS	MS	F Ratio
Model	2	0.0282	0.0141	12.7686
Error	27	0.0298	0.0011	Prob>F
C Total	29	0.0579		0.0001

RSquare	0.4861
RSquare Adj	0.4480

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.9070	0.0105	562.45	0.0000
depth	0.0022	0.0018	1.21	0.2353
depth^2	−0.0001	0.0001	−2.53	0.0177

Figure 15.10 JMP outputs for concentration study.

15.4 ORTHOGONAL POLYNOMIALS

In Section 10.1 it was shown that a linear and quadratic effect can be extracted by proper use of coefficients in orthogonal contrasts when the levels of X are equispaced. This concept can be extended to cubic, quartic, and so on, contrasts as the number of levels (k) of X increases. Statistical Table F gives the proper coefficients for such contrasts for $k = 3$ through 10. The use of these coefficients will provide orthogonal contrasts that can be treated independently of each other and from which a polynomial of the highest order whose coefficients are significant will provide an adequate equation to use for predicting the average value of Y at each level of X.

Begin by obtaining the treatment and error sums of squares for the analysis of variance model. Then calculate the numerical value of a contrast in treatment totals, starting with the linear contrast. Next determine the sum of squares associated with that contrast and divide that sum of squares by the error mean square. Under the usual ANOVA assumptions, this ratio is the observed value of an F with $\nu_1 = 1$ and $\nu_2 = N - (k+1)$. If the corresponding F test indicates significance, repeat the procedure for a second-degree polynomial. Continue this procedure to determine the highest order polynomial to consider as an adequate prediction equation. Finally, use appropriate computer software to find such a least squares polynomial.

We illustrate with the reaction time data of Example 15.1. From Table 15.3, the treatment totals are $T_{.1} = 83$, $T_{.2} = 84$, $T_{.3} = 89$, $T_{.4} = 96$, and $T_{.5} = 115$. Also, the error mean square for a one-way analysis of variance is 2.00 and $\nu_2 = 10$ from Table 15.2. Because five levels of dosage were considered, we use the $k = 5$ block of Statistical Table F to determine the contrast coefficients. Each contrast value is found by means of

$$C = \xi_1 T_{.1} + \xi_2 T_{.2} + \xi_3 T_{.3} + \xi_4 T_{.4} + \xi_5 T_{.5}$$

with ξ_j the value in the j column of the appropriate row. The associated sum of squares is found by means of

$$\text{SS} = \frac{c^2}{3\sum_{j=1}^{5}(\xi_j)^2}$$

since $n = 3$ observations were obtained at each treatment level. Notice that the value of $\sum_{j=1}^{k}(\xi_j)^2$ is also included in Statistical Table F.

For the linear effect, we use the linear row of the $k = 5$ block to find

$$\begin{aligned} c_{\text{linear}} &= (-2)(T_{.1}) + (-1)(T_{.2}) + (0)(T_{.3}) + (1)(T_{.4}) + (2)(T_{.5}) \\ &= (-2)(83) + (-1)(84) + (0)(89) + (1)(96) + (2)(115) = 76 \end{aligned}$$

and

$$\text{SS}_{\text{linear}} = \frac{(c_{\text{linear}})^2}{3\sum_{j=1}^{5}(\xi_j)^2} = \frac{(76)^2}{3(10)} = \frac{2888}{15} = 192.53$$

Thus, $f = 192.53/2 = 96.3$ and $P(F_{1,10} \geq 96.3) = 0.000$, indicating that the linear effect is significant for any reasonable value of α. This agrees with Table 15.9.

Proceeding in the same manner, we use the quadratic row of the $k = 5$ block to find

$$c_{\text{quadratic}} = (2)(83) + (-1)(84) + (-2)(89) + (-1)(96) + (2)(115) = 38$$

and

$$\text{SS}_{\text{quadratic}} = \frac{(38)^2}{3(14)} = \frac{722}{21} = 34.38$$

Thus, $f = 34.38/2 = 17.2$ and $P(F_{1,10} \geq 17.2) = 0.002$, indicating that the quadratic effect is significant for any reasonable value of α, which also agrees with Table 15.9.

Likewise,

$$c_{\text{cubic}} = (-1)(83) + (2)(84) + (0)(89) + (-2)(96) + (1)(115) = 8$$

and

$$\text{SS}_{\text{cubic}} = \frac{(8)^2}{3(10)} = \frac{32}{15} = 2.13$$

Thus, $f = 2.13/2 = 1.1$ and $P(F_{1,10} \geq 1.1) = 0.325$, indicating that the cubic effect may be negligible.

Since the test for a cubic effect has such a large p value ($p = 0.325$) but the linear and quadratic effects test highly significant (p values of 0.000 and 0.002, respectively), we would use a quadratic as an adequate equation. This conclusion agrees with that in Section 15.3.1.

We are now prepared to use a statistical computing program to determine the least squares quadratic associated with our data. Recall that, in Section 15.3, such a polynomial was obtained manually, using JMP, and using SAS.

15.5 MULTIPLE REGRESSION

We now consider a more general situation in which Y may be a function of several independent variables $X_1, X_2, \ldots, X_k$ with no restrictions on the settings of these k independent variables. In fact, in most such multiple regression situations the X variables have already acted and we simply record their values along with those of the dependent variable Y. This is, of course, *ex-post-facto* research, as opposed to experimental research in which one manipulates the X's and observes the effect on Y.

In practice there are many studies of this type. For example, one may wish to predict the surface finish of steel from dropout temperature and back-zone temperature. Here Y is a function of two recorded temperatures X_1 and X_2. Or, it may be of interest to predict college grade-point average (GPA) in the freshman year for students whose input data include rank in high school, high school Regents' average, SAT (Scholastic Aptitude Test) verbal score, and SAT mathematics score. Here Y, the freshman year GPA, is to be predicted from four independent variables: X_1, high school rank; X_2, high school Regent's average; X_3, SAT verbal; and X_4, SAT mathematical.

As in the preceding sections a mathematical model is written and the coefficients in the model are determined from the observed sample data by the method of least squares, making the sum of squares of deviations from this model a minimum.

To predict the value of a dependent variable (Y) from its regression on several independent variables ($X_1, X_2, \ldots, X_k$), the linear population model is given as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{15.23}$$

where the β's are the true coefficients to be used to weight the observed X's. In practice, a random sample of size N is chosen and the values of all variables ($Y, X_1, X_2, \ldots, X_k$) are recorded for each item in the sample. The corresponding sample model is

$$Y' = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_k X_k \tag{15.24}$$

where the B's are estimators of the β's. When $k = 1$, this gives the straight-line model of Section 15.2.

After sampling, sample estimates of the β's are calculated, giving

$$y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \tag{15.25}$$

Since each observed y can be expressed as

$$y_i = y_i' + e_i$$

with

$$i = 1, 2, \ldots, N$$

where the e's are called *residuals*, the b_i's are determined by minimizing $\sum e_i^2$. Differentiating and setting each partial derivative to zero gives the following least squares equations:

$$\left.\begin{aligned}
\sum y &= b_0 N + b_1 \sum x_1 + b_2 \sum x_2 + \cdots + b_k \sum x_k \\
\sum x_1 y &= b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \cdots + b_k \sum x_1 x_k \\
\sum x_2 y &= b_0 \sum x_2 + b_1 \sum x_2 x_1 + b_2 \sum x_2^2 + \cdots + b_k \sum x_2 x_k \\
&\vdots \\
\sum x_k y &= b_0 \sum x_k + b_1 \sum x_k x_1 + b_2 \sum x_k x_2 + \cdots + b_k \sum x_k^2
\end{aligned}\right\} \quad (15.26)$$

This set of $k + 1$ equations in $k + 1$ unknowns (b_0, b_2, . . . , b_k) can be solved for the b's. Solving these equations is tedious if more than two or three independent variables are involved, so a specialized computer program is usually used to obtain a solution.

Even though many independent variables may be used, simpler models are more appealing and easier to interpret. Thus, we try to identify the smallest subset of the independent variables that will provide an adequate model. To this end, many practitioners compare the values of the adjusted R-squares of the different models. If N observations are made and $k + 1$ parameters are estimated, *adjusted R-square* is given by

$$R^2_{\text{adj}} = \frac{(N-1)R^2 - k}{N - 1 - k} \quad (15.27)$$

with $R^2 = \text{SS}_{\text{model}}/\text{SS}_{\text{total}}$ the coefficient of multiple determination for the regression model. If there is little difference between the adjusted R-squares for two models, we have evidence that supports using the simpler model. Before any model is recommended, of course, its adequacy should be thoroughly assessed, along with the adequacy of the regression assumptions.

To illustrate the procedure, we consider an experiment involving 24 samples of a steel alloy. For each sample, a chemical analysis is made and the percentages of five specific chemical elements (X_1, X_2, X_3, X_4, X_5) are recorded. Placing each sample under stress, the percent elongation (Y) is then determined. The resulting data are given in Table 15.11.

In Figure 15.11 we give Minitab output obtained using the **Best Subsets** module. Minitab searched through the $2^5 = 32$ models that could be developed from our five independent variables and reported the adjusted R-squares for the best two models in one, two, three, and four variables. That statistic is also reported for the full model in five variables. The best single predictor of Y is X_2, since the symbol $\times$ is below x_2 in the line that contains the first 1 in the variables column (Vars). Likewise, the best two-variable model contains X_2 and X_5, since the symbol $\times$ is below both x_2 and x_5 in the line that contains the first 2 in the variables column. Proceeding in this fashion, we find that the best three-variable model contains X_2, X_3, and X_5, whereas the best four-variable model contains X_2, X_3, X_4, and X_5.

Suppose we decide that an appropriate (tentative) model should contain only X_2 and X_5. For this analysis we would use the **Regression** module in Minitab, obtaining the outputs in Figure 15.12. Notice that an unusually large residual has been detected. Problem 15.38 asks the reader to conduct a residual analysis for this situation.

TABLE 15.11
Percent Elongation Data

Item	y	x_1	x_2	x_3	x_4	x_5
1	11.3	0.50	1.3	0.4	3.4	0.010
2	10.0	0.47	1.2	0.3	3.6	0.012
3	9.8	0.48	3.1	0.7	4.3	0.000
4	8.8	0.54	2.6	0.7	4.0	0.022
5	7.8	0.45	2.8	0.7	4.2	0.000
6	7.4	0.41	3.2	0.7	4.7	0.000
7	6.7	0.62	3.0	0.6	4.7	0.026
8	6.3	0.53	4.1	0.9	4.6	0.035
9	6.3	0.57	3.7	0.8	4.6	0.000
10	6.3	0.67	2.7	0.6	4.8	0.013
11	6.0	0.54	3.1	0.7	4.2	0.000
12	6.0	0.42	3.1	0.7	4.4	0.000
13	5.8	0.33	2.6	0.6	4.7	0.008
14	5.5	0.51	3.9	0.9	4.4	0.000
15	5.5	0.54	3.1	0.7	4.2	0.000
16	4.7	0.48	4.0	1.1	3.7	0.024
17	4.1	0.38	3.3	0.8	4.1	0.000
18	4.1	0.39	3.2	0.7	4.6	0.016
19	3.9	0.60	2.9	0.7	4.3	0.025
20	3.5	0.54	3.2	0.7	4.9	0.022
21	3.1	0.33	2.9	2.9	1.0	0.063
22	1.6	0.40	3.2	3.2	1.0	0.059
23	1.1	0.64	2.5	0.7	3.8	0.018
24	0.6	0.34	5.0	1.3	3.9	0.044

Best Subsets Regression

Response is y

Vars	R-sq	Adj. R-sq	C-p	s	x 1	x 2	x 3	x 4	x 5
1	31.4	28.3	5.4	2.3077		X			
1	28.3	25.1	6.5	2.3590			X		
2	49.2	44.4	0.8	2.0328		X			X
2	47.3	42.3	1.5	2.0703		X	X		
3	50.6	43.1	2.3	2.0552		X	X		X
3	49.7	42.2	2.6	2.0727		X		X	X
4	51.1	40.8	4.1	2.0975		X	X	X	X
4	50.9	40.6	4.1	2.1007	X	X	X		X
5	51.3	37.8	6.0	2.1496	X	X	X	X	X

Figure 15.11 Minitab summary for elongation models.

It is interesting to note that observation 23 yields an unusually large residual for every model listed in Figure 15.11. In Problem 15.39, the reader is asked to conduct an analysis of the data in Table 15.11 after observation 23 has been removed.

The procedure illustrated here is descriptive and exploratory. Once a short list of possible models has been developed using a program such as Minitab, further analysis may indicate that one or more of these *may* serve as an adequate model. However, we

```
Regression Analysis

The regression equation is
y = 11.7 - 1.62 x2 - 63.2 x5

Predictor       Coef     Stdev   t-ratio          p
Constant      11.695     1.690      6.92      0.000
x2           -1.6200    0.5448     -2.97      0.007
x5            -63.18     23.30     -2.71      0.013

s = 2.033    R-sq = 49.2%    R-sq(adj) = 44.4%

Analysis of Variance
SOURCE           DF        SS        MS          F          p
Regression        2    84.071    42.035      10.17      0.001
Error            21    86.774     4.132
Total            23   170.845

SOURCE           DF    SEQ SS
x2                1    53.680
x5                1    30.391

Unusual Observations
Obs.             x2         y       Fit   Stdev.Fit   Residual   St.Resid
 23            2.50     1.100     6.508       0.524     -5.408     -2.75R

R denotes an obs. with a large st. resid.
```

Figure 15.12 Minitab output for the tentative two-factor elongation model.

cannot formally test or estimate using such tentative models. Rather, a new study should be conducted to assess the appropriateness of the proposed model.

Further Comments Concerning the Elongation Study

Residuals for a regression analysis are easily obtained when statistical computing programs are used for the analysis. From the Minitab output of Figure 15.12, we observed that one residual was suspiciously large. Further analysis of the residuals is requested in Problem 15.38. For now, however, consider the values of the residuals (obtained using Minitab) as given in Table 15.12. Note that the residuals (or errors of estimate) seem to be nonrandom inasmuch as, in general, they are positive for large values of Y and negative for small values of Y. This may indicate a need to consider some higher order terms such as X_2^2, X_5^2, or X_2X_5. Only modest changes in the model statement of the computer program will effect such a refinement.

Alternatives to All-Subsets Regression

The procedure illustrated by means of the elongation data is an all-subsets regression analysis. In such an analysis, each possible subset model involving the variables under consideration is fitted to the sample data and assessed based on a given criterion (e.g., the coefficient of determination, the adjusted R-square value). Other commonly used procedures include stepwise regression and two special cases—forward selection and backward elimination. However, these procedures may miss some models considered by the

TABLE 15.12
Deviations from Regression on Elongation Data

y	y'	e	y	y'	e
11.3	8.95704	2.34296	5.8	6.97740	−1.17740
10.0	8.99268	1.00732	5.5	5.37684	0.12316
9.8	6.67284	3.12716	5.5	6.67284	−1.17284
8.8	6.09290	2.70710	4.7	3.69854	1.00146
7.8	7.15884	0.64116	4.1	6.34884	−2.24884
7.4	6.51084	0.88916	4.1	5.49997	−1.39997
6.7	5.19218	1.50782	3.9	5.41736	−1.51736
6.3	2.84157	3.45843	3.5	5.12090	−1.62090
6.3	5.70084	0.59916	3.1	3.01655	0.08345
6.3	6.49951	−0.19951	1.6	2.78327	−1.18327
6.0	6.67284	−0.67284	1.1	6.50761	−5.40761
6.0	6.67284	−0.67284	0.6	0.81496	−0.21496

all-subsets procedure. Thus, use of all-subsets regression is recommended when adequate computing facilities are available.

15.5.1 A SAS Program

The **RSQUARE** procedure in SAS is an all-subsets regression that can be used like the Minitab **Best Subsets** module. Since, however, stepwise regression is often used in determinations of the variables to be included in a regression model, we now illustrate the use of the **STEPWISE** procedure in SAS with the **MAXR** option. This technique is considered to be almost as good as an all-subsets regression.

A command file for the analysis of the elongation data of Table 15.11 is given in Table 15.13. Notice that the **MAXR** option is requested by placing the symbol / at the end of the model statement and following that symbol with the name of the option to be used. Other options include **F** for forward selection, **B** for backward elimination, and **STEPWISE** for stepwise regression.

The outputs from the command file of Table 15.13 are given in Figure 15.13. Notice that the "best" two-factor model

TABLE 15.13
SAS Command File for Elongation Study

```
DATA ELONG;
INPUT Y X1 X2 X3 X4 X5;
CARDS;
11.3 0.50 1.3 0.4 3.4 0.010
10.0 0.47 1.2 0.3 3.6 0.012
                ⋮
 0.6 0.34 5.0 1.3 3.9 0.044
;
PROC STEPWISE;
MODEL Y = X1 X2 X3 X4 X5/MAXR;
```

```
MAXIMUM R-SQUARE IMPROVEMENT FOR DEPENDENT VARIABLE Y

STEP 1   VARIABLE X2 ENTERED   R SQUARE = 0.31420181   C(P) = 5.35523227
                        DF   SUM OF SQUARES   MEAN SQUARE        F   PROB>F
REGRESSION               1     53.67980887   53.67980887    10.08   0.0044
ERROR                   22    117.16519113    5.32569051
TOTAL                   23    170.84500000

                   B VALUE        STD ERROR   TYPE II SS        F   PROB>F
INTERCEPT      11.57748846
X2             -1.92211293       0.60542644   53.67980887    10.08   0.0044
THE ABOVE MODEL IS THE BEST 1 VARIABLE MODEL FOUND.

STEP 2   VARIABLE X5 ENTERED   R SQUARE = 0.49208798   C(P) = 0.77845055
                        DF   SUM OF SQUARES   MEAN SQUARE        F   PROB>F
REGRESSION               2     84.07077035   42.03538517    10.17   0.0008
ERROR                   21     86.77422965    4.13210617
TOTAL                   23    170.84500000

                   B VALUE        STD ERROR   TYPE II SS        F   PROB>F
INTERCEPT      11.69482586
X2             -1.61999580       0.54479609   36.53686509     8.84   0.0073
X5             -63.17916865      23.29632535   30.39096148     7.35   0.0131
THE ABOVE MODEL IS THE BEST 2 VARIABLE MODEL FOUND.

STEP 3   VARIABLE X3 ENTERED   R SQUARE = 0.50554119   C(P) = 2.28106031
DF           SUM OF SQUARES      MEAN SQUARE             F   PROB>F
REGRESSION               3     86.36918442   28.78972814     6.82   0.0024
ERROR                   20     84.47581558    4.22379078
TOTAL                   23    170.84500000

                   B VALUE        STD ERROR   TYPE II SS        F   PROB>F
INTERCEPT      11.82239196
X2             -1.54994707       0.55893256   32.48009675     7.69   0.0117
X3             -0.74853877       1.01473201    2.29841407     0.54   0.4693
X5             -42.22569494      36.89984445    5.53104177     1.31   0.2660
THE ABOVE MODEL IS THE BEST 3 VARIABLE MODEL FOUND.

STEP 4   VARIABLE X4 ENTERED   R SQUARE = 0.51071485   C(P) = 4.08978045
                        DF   SUM OF SQUARES   MEAN SQUARE        F   PROB>F
REGRESSION               4     87.25307854   21.81326963     4.96   0.0066
ERROR                   19     83.59192146    4.39957481
TOTAL                   23    170.84500000

                   B VALUE        STD ERROR   TYPE II SS        F   PROB>F
INTERCEPT      14.31358314
X2             -1.22427578       0.92375797    7.72774578     1.76   0.2008
X3             -1.66044527       2.28291111    2.32745790     0.53   0.4759
X4             -0.65610814       1.46379635    0.88389412     0.20   0.6591
X5             -41.70034653      37.80329379    5.87923996     1.34   0.2620
THE ABOVE MODEL IS THE BEST 4 VARIABLE MODEL FOUND.

STEP 5   VARIABLE X1 ENTERED   R SQUARE = 0.51314320   C(P) = 6.0000000
                        DF   SUM OF SQUARES   MEAN SQUARE        F   PROB>F
REGRESSION               5     87.66794928   17.53358986     3.79   0.0160
ERROR                   18     83.17705072    4.62094726
TOTAL                   23    170.84500000

                   B VALUE        STD ERROR   TYPE II SS        F   PROB>F
INTERCEPT      14.91722299
X1             -1.57654684       5.26157760    0.41487075     0.09   0.7679
X2             -1.27616946       0.96242424    8.12483590     1.76   0.2014
X3             -1.69396517       2.34231361    2.41685102     0.52   0.4788
X4             -0.57829883       1.52248084    0.66670232     0.14   0.7085
X5             -43.14470468      39.67043862    4.97078416     1.08   0.3134
THE ABOVE MODEL IS THE BEST 5 VARIABLE MODEL FOUND.
```

Figure 15.13 SAS analysis of the elongation data.