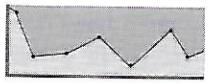
Chapter 15



REGRESSION

15.1 INTRODUCTION

In Figure 3.1 (After ANOVA, what?), we noted several procedures that might follow an analysis of variance when the effects of factors were significant. However, we have not yet treated cases in which the factor levels are quantitative. Figure 3.1 suggests two procedures to follow when significant effects are present: the method of orthogonal polynomials when the factor levels are equispaced, and general methods of curve fitting or regression.

Some experimenters prefer to treat all analyses as regression models because even the ANOVA procedures considered to this point can be cast into a regression model. This is evidenced by the term GLM (general linear model) of most computer packages. If any one of the independent variables (X's) is quantitative and its effect is statistically significant, one probably should use regression analysis in an attempt to determine a mathematical model suitable for predicting the magnitude of the response variable from values of the independent variables. This regression procedure will be examined for Y as a function of one X, where the relationship may be linear, quadratic, or a higher order polynomial. The case of multiple regression, where Y is a function of several X's, will also be considered. Our search for an adequate prediction equation relies on computer software because most real problems are too difficult to handle otherwise.

Some Notation

Suppose an experiment is conducted for which n_j observations are made when $X = x_j$. Throughout this chapter we will let y_{ij} denote the ith value of Y at x_j. The average of the sample of n_j observations is denoted $\bar{y}_{\cdot j}$, and the predicted value of Y is denoted y'_{x_j} with the subscripts omitted as the context permits.

Before data are collected, the ith value of Y at $X = x_j$ is a random variable and, as such, will be denoted Y_{ij}. Likewise, $\bar{Y}_{\cdot j}$ denotes the sample mean that will result once the sample data have become available, and $Y'_{X=x_j}$ denotes the value of Y that will result.

15.2 LINEAR REGRESSION

To review linear regression and become familiar with the notation used, consider an experiment designed to determine the effect of the amount of drug dosage X (in milligrams)

on a person's reaction time Y (in milliseconds) to a given stimulus. When 15 subjects were randomly assigned one of five dosages of drug—0.5, 1.0, 1.5, 2.0, or 2.5 mg—with three subjects assigned to a given drug dosage, the reaction times of Table 15.1 were recorded.

From Table 15.2, the ANOVA results for the data in Table 15.1, it is obvious that dosage has a highly significant effect on reaction time. But since reaction time is a quantitative factor, the next question might well be: How does reaction time vary with drug dosage? Can one find a functional relationship between these two variables that might allow the prediction of reaction time from drug dosage?

As a first step, we construct Figure 15.1, a scattergram of all 15 (x, y) pairs of points, where the $\times$'s denote the observed average, $\bar{y}_{\cdot j}$, at each x_j. The plot shows that Y increases with increases in X.

A reasonable second step is to try a straight-line "fit" as a first approximation for predicting Y from X. Such a line is shown in Figure 15.1. One can see how close the line comes to $\bar{y}_{\cdot j}$ for each x_j. Note that for a particular value (or level) of X, the ith value of Y can be partitioned into three parts:

1. y'_{x_j}, which is the predicted value of Y at $X = x_j$.
2. $\bar{y}_{\cdot j} - y'_{x_j}$, which is the amount by which the sample average for Y at $X = x_j$ deviates from its predicted value (referred to as the *departure from linear regression*).

TABLE 15.1
Reaction Time Data

	Dosage				
	0.5	**1.0**	**1.5**	**2.0**	**2.5**
Time	26	28	28	32	38
	28	26	30	33	39
	29	30	31	31	38

TABLE 15.2
Minitab ANOVA for Reaction Time

```
One-Way Analysis of Variance

Analysis of Variance on Time
Source      DF         SS         MS         F         p
Dosage       4     229.73      57.43     28.72     0.000
Error       10      20.00       2.00
Total       14     249.73

                                        Individual 95% CIs For Mean
                                        Based on Pooled StDev
Level       N       Mean      StDev     -------+---------+---------+---------+
    1       3     27.667      1.528       (---*----)
    2       3     28.000      2.000       (----*----)
    3       3     29.667      1.528            (---*----)
    4       3     32.000      1.000                (----*----)
    5       3     38.333      0.577                          (----*---)
                                        -------+---------+---------+---------+
Pooled StDev =      1.414                 28.0      32.0      36.0      40.0
```

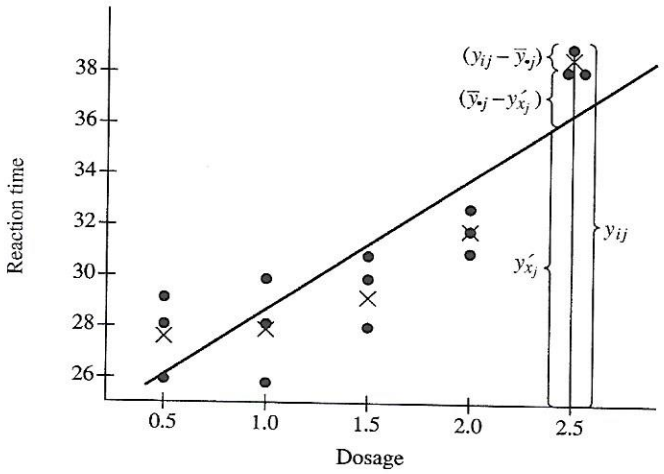


Figure 15.1 Reaction time versus drug dosage.

3. $y_{ij} - \bar{y}_{\cdot j}$, which is the amount by which the ith value of Y at $X = x_j$ varies from the mean of the sample of n_j observations.

15.2.1 The Simple Linear Model

In the analysis of variance model we have

$$Y_{ij} = \mu + (\mu_j - \mu) + (Y_{ij} - \mu_j) \tag{15.1}$$

and now we try to predict μ_j from some function based on X_j. If this predicted mean is labeled $\mu_{Y|X}$, the model can be expanded to read

$$Y_{ij} = \mu + (\mu_{Y|X} - \mu) + (\mu_j - \mu_{Y|X}) + (Y_{ij} - \mu_j) \tag{15.2}$$

where $\mu_{Y|X}$ is the true predicted mean based on X and $\mu_j - \mu_{Y|X}$ is the amount by which the true mean μ_j departs from its predicted value.

For the sample model, Equation (15.2) becomes

$$Y_{ij} = \bar{Y}_{\cdot\cdot} + (Y'_X - \bar{Y}_{\cdot\cdot}) + (\bar{Y}_{\cdot j} - Y'_X) + (Y_{ij} - \bar{Y}_{\cdot j}) \tag{15.3}$$

If the second and third terms on the right of this equation are combined, we have the ANOVA sample model. Thus, we attempt here to partition the mean of a given treatment into two parts: that which can be predicted by regression on X, and the amount by which the mean shows a departure from such a regression model. The purpose, of course, is to find a model for predicting the response means for which the departures from these means are very small. By choosing a polynomial of high enough degree, one can find a model that actually goes through all the Y means. However, it is hoped that an adequate lower degree model can be found whose departures are small and insignificant compared with the error of individual Y's around their means.

The Straight-Line Model

In Equation (15.3) Y'_X represents the predicted Y (or average Y) for a given X for any assumed model. The straight-line model is given by

$$Y'_X = B_0 + B_1 X \tag{15.4}$$

where B_0 is the Y intercept and B_1 is the slope. When the true average Y and the corresponding level of X are linearly related, B_0 and B_1 are estimators of the intercept and slope, respectively, of the true line of the means. In many cases, of course, such a linear relation is only approximate but adequate.

15.2.2 The Least Squares Line

The usual method for determining estimators of parameters in a simple linear model is the method of *least squares*. The estimators in Equation (15.4) are determined from a random sample in a way that minimizes the sum of squares of the deviations of each Y_{ij} from its predicted value Y'_X. [One could also find estimators such that the sum of the squares of the departures from regression (i.e., $\bar{Y}_{\cdot j} - Y'_X$) is minimized, but this leads to the same results.] In this case,

$$Y_{ij} - Y'_X = Y_{ij} - B_0 - B_1 X_j$$

and

$$\text{SS}_{\text{deviations}} = \sum_i \sum_j (Y_{ij} - Y'_X)^2 = \sum_i \sum_j (Y_{ij} - B_0 - B_1 X_j)^2$$

Because the summations of X's and Y's can be found from a given set of data, $\text{SS}_{\text{deviations}}$ is a function of B_0 and B_1. To minimize $\text{SS}_{\text{deviations}}$ then, one differentiates $\text{SS}_{\text{deviations}}$ and sets the two partial derivatives to zero. This gives

$$\frac{\partial(\text{SS}_{\text{deviations}})}{\partial B_0} = 2\sum_i \sum_j (Y_{ij} - B_0 - B_1 X_j)(-1) = 0$$

$$\frac{\partial(\text{SS}_{\text{deviations}})}{\partial B_1} = 2\sum_i \sum_j (Y_{ij} - B_0 - B_1 X_j)(-X_j) = 0$$

Dividing through by -2 and simplifying gives the system of equations

$$\left.\begin{aligned} \sum_i \sum_j Y_{ij} &= B_0 N + B_1 \sum_j n_j X_j \\ \sum_i \sum_j X_j Y_{ij} &= B_0 \sum_j n_j X_j + B_1 \sum_j n_j X_j^2 \end{aligned}\right\} \tag{15.5}$$

with N the total sample size and n_j the sample size at the jth level of X. These equations are often called the *least squares normal equations*, which can now be solved to obtain expressions for the random variables B_0 and B_1. The sample values obtained for a specific sample of (x, y) pairs are denoted b_0 and b_1. These are obtained by solving the system

$$\left.\begin{aligned}\sum_i \sum_j y_{ij} &= b_0 N + b_1 \sum_j n_j x_j \\ \sum_i \sum_j x_j y_{ij} &= b_0 \sum_j n_j x_j + b_1 \sum_j n_j x_j^2\end{aligned}\right\} \tag{15.6}$$

and the resulting sample equation is

$$y'_x = b_0 + b_1 x \tag{15.7}$$

Example 15.1 (Least Squares Line for Reaction Time Data)

The analysis of variance for the reaction time data of Table 15.1 revealed a significant time effect. We will now begin a search for an adequate model of reaction time as a function of dosage. For convenience, the data have been reproduced with totals in Table 15.3.

From Table 15.3, we write $\sum_i \sum_j y_{ij} = \sum_j T_{\cdot j} = 83 + 84 + 89 + 96 + 115 = 467$; $n_j = 3$ for $j = 1, 2, 3, 4, 5$; and $N = 3 \times 5 = 15$. Further,

$$\sum_j x_j = 0.5 + 1.0 + 1.5 + 2.0 + 2.5 = 7.5$$

and

$$\sum_j x_j^2 = (0.5)^2 + (1.0)^2 + (1.5)^2 + (2.0)^2 + (2.5)^2 = 13.75$$

To find the cross product, note that since $\sum_i y_{ij} = T_{\cdot j}$, $\sum_i \sum_j x_i y_{ij} = \sum_j x_j T_{\cdot j}$. Thus,

$$\sum_i \sum_j x_i y_{ij} = (0.5)(83) + (1.0)(84) + (1.5)(89) + (2.0)(96) + (2.5)(115) = 738.5$$

Substituting in Equation (15.6) gives

$$467.0 = 15.0b_0 + 22.50b_1$$
$$738.5 = 22.5b_0 + 41.25b_1$$

TABLE 15.3
Reaction Time Data with Totals

	Dosage				
	0.5	**1.0**	**1.5**	**2.0**	**2.5**
Time	26	28	28	32	38
	28	26	30	33	39
	29	30	31	31	38
$T_{\cdot j}$	83	84	89	96	115

To solve the system given here, one might multiply the first equation by 1.5, giving

$$700.5 = 22.5b_0 + 33.75b_1$$
$$738.5 = 22.5b_0 + 41.25b_1$$

Subtracting and solving for b_1, we have

$$b_1 = \frac{700.5 - 738.5}{33.75 - 41.25} = \frac{-38.0}{-7.5} = \frac{76}{15}$$

and from the original first equation,

$$b_0 = \frac{467 - (22.5)(76/15)}{15} = \frac{353}{15}$$

So, the linear model for this set of sample data is

$$y'_x = \frac{353}{15} + \frac{76}{15}x \approx 23.53 + 5.07x$$

The sample equation for a least squares line is seldom obtained by solving the system of normal equations directly. Even moderately priced calculators often contain a simple linear regression routine. Regression modules are also included in statistical computing programs. For example, when the Table 15.3 data are analyzed using the **Regression** module in Minitab we obtain Figure 15.2. The test for a linear effect (a t test on milligrams or an F test on regression) indicates that inclusion of an x term in our model is reasonable ($p = 0.000$).

A SAS Program

The least squares line obtained manually in Example 15.1 and included in the Minitab summary of Figure 15.2 can also be obtained using the **GLM** (or **REG**) procedure in SAS.

```
Regression Analysis

The regression equation is
Time = 23.5 + 5.07 dose

Predictor         Coef       Stdev      t-ratio          p
Constant        23.533       1.270        18.53      0.000
dose             5.067       0.766         6.61      0.000

s = 2.098      R-sq = 77.1%      R-sq(adj) = 75.3%

Analysis of Variance

SOURCE              DF          SS          MS          F          p
Regression           1      192.53      192.53      43.76      0.000
Error               13       57.20        4.40
Total               14      249.73
```

Figure 15.2 Minitab regression analysis for reaction time study.

An appropriate command file is given in Table 15.4. Since we are estimating a functional relationship between the two variables, a class statement is not used. Except for rounding, the slope and intercept included in the outputs of Figure 15.3 are the same as those obtained manually and using Minitab.

A Formula for B_0 *and* B_1

If the normal equations of Equation (15.5) are solved simultaneously, we have

$$B_0 = \bar{Y} - B_1\bar{X} \tag{15.8}$$

and

$$B_1 = \frac{\sum_i \sum_j X_i Y_{ij} - \frac{\left(\sum_i \sum_j X_i\right)\left(\sum_i \sum_j Y_{ij}\right)}{N}}{\sum_i \sum_j X_i^2 - \frac{\left(\sum_i \sum_j X_j\right)^2}{N}} = \frac{\sum_i \sum_j (X_i - \bar{X})(Y_{ij} - \bar{Y})}{\sum_i \sum_j (X_i - \bar{X})^2} \tag{15.9}$$

TABLE 15.4
SAS Command File for Example 15.1

```
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
;
PROC GLM;
MODEL TIME = DOSE;
```

```
                    GENERAL LINEAR MODELS PROCEDURE
DEPENDENT VARIABLE: TIME

                                      SUM OF          MEAN
SOURCE                      DF        SQUARES        SQUARE       F VALUE
MODEL                        1   192.53333333   192.53333333         43.76
ERROR                       13    57.20000000     4.40000000       PR > F
CORRECTED TOTAL             14   249.73333333                       0.0001

R-SQUARE                 C.V.        ROOT MSE       TIME MEAN
0.770956               6.7375      2.09761770     31.13333333

SOURCE                      DF       TYPE I SS       F VALUE       PR > F
DOSE                         1   192.53333333         43.76       0.0001

SOURCE                      DF      TYPE III SS       F VALUE       PR > F
DOSE                         1   192.53333333         43.76       0.0001

                                    T FOR H0:                   STD ERROR OF
PARAMETER             ESTIMATE   PARAMETER=0      PR > |T|        ESTIMATE
INTERCEPT         23.53333333         18.53        0.0001      1.27017059
DOSE               5.06666667          6.61        0.0001      0.76594169
```

Figure 15.3 SAS output for Example 15.1.

Denoting the numerator and denominator of the last fraction in Equation (15.9) by SP_{XY} and SS_X, respectively, we have

$$B_1 = \frac{SP_{XY}}{SS_X} \tag{15.10}$$

Example 15.2 [Using Equations (15.8) and (15.9) with Reaction Time Data]

From Example 15.1, we know that $\sum_i \sum_j y_{ij} = 467$, $\sum_i \sum_j x_j = 3(7.5) = 22.5$, $\sum_i \sum_j x_j^2 = 3(13.75) = 41.25$, $N = 15$, and $\sum_i \sum_j x_i y_{ij} = 738.5$ for the reaction time data of Table 15.1. Using these with the sample equivalents of Equations (15.8) and (15.9), we find

$$b_1 = \frac{\sum_i \sum_j x_i y_{ij} - \frac{\left(\sum_i \sum_j x_i\right)\left(\sum_i \sum_j y_{ij}\right)}{N}}{\sum_i \sum_j x_i^2 - \frac{\left(\sum_i \sum_j x_j\right)^2}{N}} = \frac{738.5 - \frac{(22.5)(467)}{15}}{41.25 - \frac{(22.5)^2}{15}} = \frac{76}{15}$$

and

$$b_0 = \bar{y} - b_1\bar{x} = \frac{467}{15} - \frac{76}{15} \times \frac{22.5}{15} = \frac{353}{15}$$

as before.

15.2.3 Departure from the Linear Model: A Lack of Fit Test

We noted earlier that the line of Figure 15.1 does not provide a perfect fit to the data. In fact, that figure indicates that a nonlinear effect is present. To see how well the empirical model

$$y'_x = \frac{353}{15} + \frac{76}{15}x$$

predicts the observed value of Y from a given value of X, we construct Table 15.5. This table shows the departures from linear regression. For example, when the observed value of X is 1.0, the predicted value of Y is $y' = 353/15 + (76/15)(1) = 143/5$, or 28.6. But the average value of Y when X is 1.0 is $84/3 = 28$. In this instance, the departure from linear regression is

$$\bar{y} - y' = 28.0 - 28.6 = -0.6 = -\tfrac{3}{5}$$

Note that the sum of the departures adds to zero and the observed sum of squares of departures from linear regression is

$$SS_{\text{departures}} = \sum_j n_j(\bar{y}_{\cdot j} - y'_{x_j})^2 = 3\left(\frac{2791}{225}\right) = \frac{2791}{75} \approx 37.21$$

as each departure of a mean must be weighted with three observed values for each value of X.

TABLE 15.5
Departures from Linear Regression

x_j	$\bar{y}_{.j}$	y'_x	$\bar{y}_{.j} - y'_x$	$(\bar{y}_{.j} - y'_x)^2$
0.5	83/3	391/15	24/15	576/225
1.0	84/3	429/15	−9/15	81/225
1.5	89/3	467/15	−22/15	484/225
2.0	96/3	505/15	−25/15	625/225
2.5	115/3	543/15	32/15	1025/225
Totals			0	2791/225

From Equation (15.3), we write

$$(Y'_X - \bar{Y}_{..}) + (\bar{Y}_{.j} - Y'_X) = \bar{Y}_{.j} - \bar{Y}_{..}$$

where the right-most term is the deviation of a treatment mean from the overall mean in an ANOVA model. Squaring both sides and summing over i and j gives

$$\sum_i \sum_j (Y'_X - \bar{Y}_{..})^2 + \sum_i \sum_j (\bar{Y}_{.j} - Y'_X)^2 = \sum_i \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2 \qquad (15.11)$$

with the third sum of squares the treatment sum of squares for the ANOVA model. When the sum of squares of departures from linear regression is known, the ANOVA Table 15.2 can be expanded to give Table 15.6. This table shows a highly significant linear effect but also a significant departure from the linear model (p value = 0.012). This indicates that a higher order model may be required to adequately predict the reaction time from drug dosage.

Notice that the sum of $SS_{departure}$ and SS_{error} in Table 15.6 is the error sum of squares in Figures 15.2 and 15.3. Thus, the regression error has been decomposed into the error you get from ANOVA (often called *pure error*) and a portion due to (nonlinear) treatment effects.

Many statistical computing programs include such an analysis, called a lack of fit test, as an option but to use this option requires multiple observations on at least one x. For example, Minitab includes

TABLE 15.6
ANOVA on Reaction Time with Linear Regression

Source	df	SS	MS	F	p value
Between dosages*	4	229.73			
Linear	1	192.53	192.53	96.3	0.000
Departure from linear	3	37.21	12.40	6.2	0.012
Error	10	20.00	2.00		
Totals	14	249.73			

*$SS_{linear} + SS_{departure} \neq SS_{between}$ due to round-off error.

```
Pure error test − F = 6.20   P  = 0.0119   DF(pure error) = 10
```

with the outputs of Figure 15.2 when a test for departure from the linear model is requested.

Because there is a strong linear effect in this example and, in some problems, the linear model is adequate to explain most of the variation in the Y variable, we now consider some useful statistics that can be determined from Table 15.6.

Coefficient of Determination

The proportion of the total sum of squares that can be accounted for by linear regression is sometimes called the *coefficient of determination* and denoted r^2. From Table 15.6, the coefficient of determination for the reaction time data is found to be

$$r^2 = \frac{SS_{linear}}{SS_{total}} = \frac{192.53}{249.73} = 0.7712$$

Thus, linear regression will account for about 77% of the variation seen in the reaction time Y.

Recall that the positive square root of the coefficient of determination with the sign of b_1, affixed, denoted r, is the *Pearson product–moment correlation coefficient*. In this case, $r = (0.7712)^{0.5} = 0.88$.

Eta Squared

The ratio of the sum of squares between means to the total sum of squares is called *eta squared* and is denoted η^2. This ratio gives the maximum amount of the total variation that could be accounted for by a curve or model that passes through all the mean Y's for each X. Here

$$\eta^2 = \frac{SS_{between}}{SS_{total}} = \frac{229.73}{249.73} = 0.9199$$

Thus, approximately 92% of the variation in reaction time can be accounted for by a model through all the means for each of the five dosage levels.

Standard Error of Estimate

Another statistic used with a linear model is the standard error of estimate, denoted $S_{Y.X}$. It is the standard deviation of the deviations of the Y_{ij}'s from their predicted values

$$Y_{ij} - Y'_X = (Y_{ij} - \bar{Y}_{.j}) + (\bar{Y}_{.j} - Y'_X)$$

This expression adds the error about the mean and the departure of the mean from the predicted curve to obtain

$$S_{Y.X} = \sqrt{\frac{SS_{departure} + SS_{error}}{N - 2}}$$

and is appropriate only if the departure is nonsignificant. This statistic is often used to set confidence limits around a line of best fit when linear regression is appropriate.

15.2.4 The Use of Equispaced Levels

To this point, no use has been made of the fact that the dosages of the reaction time study are equispaced. In such cases, one can code the X_j's by considering their overall mean $\bar{X}$ and the width c of the interval between two consecutive values. If we choose

$$U_j = \frac{X_j - \bar{X}}{c} \tag{15.12}$$

the simple linear model becomes

$$Y'_U = B'_0 + B'_1 U \tag{15.13}$$

whose least squares normal equations are given by

$$\left.\begin{aligned} \sum_i \sum_j Y_{ij} &= B'_0 N + B'_1 N \sum_j U_j \\ \sum_i \sum_j U_j Y_{ij} &= B'_0 \sum_j n_j U_j + B'_1 \sum_j n_j U_j^2 \end{aligned}\right\} \tag{15.14}$$

Upon equispacing the values of X, we find that the corresponding values of U sum to zero. Knowing this, we can give the intercept and slope of the least squares line for a particular sample by

$$\left.\begin{aligned} b'_0 &= \frac{\sum_i \sum_j y_{ij}}{N} = \bar{y} \\ b'_1 &= \frac{\sum_i \sum_j u_j y_{ij}}{\sum_j n_j u_j^2} = \frac{\sum_j u_j T_{.j}}{\sum_j n_j u_j^2} \end{aligned}\right\} \tag{15.15}$$

Equation (15.15) is relatively easy to solve. It also has the advantage that the numerator of the expression for the slope is a contrast in the treatment totals. Thus, the observed sum of squares due to linear regression can be found by means of

$$\mathrm{SS}_{\text{linear}} = \frac{\left(\sum_j u_j T_{.j}\right)^2}{\sum_j n_j u_j^2} \tag{15.16}$$

Example 15.3 (Another Look at the Least Squares Line for Reaction Time Data)

For the reaction time data of Table 15.1, the levels of X are 0.5, 1.0, 1.5, 2.0, and 2.5, with three observations per level. Thus, $\bar{x} = (0.5 + 1.0 + 1.5 + 2.0 + 2.5)/5 = 1.5$. Since the width of the interval between any two consecutive levels is 0.5, the lowest level of U is $u_1 = (0.5 - 1.5)/0.5 = -2$. Likewise, $u_2 = -1$, $u_3 = 0$, $u_4 = 1$, and $u_5 = 2$. From Table 15.3, the sample totals are $T_{.1} = 83$, $T_{.2} = 84$, $T_{.3} = 89$, $T_{.4} = 96$, and $T_{.5} = 115$. Summing these totals gives $\bar{y} = 467/15$. From Equation (15.15),

$$b'_0 = \bar{y} = \frac{467}{15}$$

and

$$b_1' = \frac{\sum_j u_j T_{.j}}{\sum_j n_j u_j^2} = \frac{(-2)(83) + (-1)(84) + (0)(89) + (1)(96) + (2)(115)}{3[(-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2]}$$
$$= \frac{76}{30} = \frac{38}{15}$$

So, the straight-line model is

$$y_u' = \frac{467}{15} + \frac{38}{15}u$$

To see that this is the same model, substitute $(x - 1.5)/0.5$ for u to obtain

$$y_x' = \frac{467}{15} + \frac{38}{15} \times \frac{x - 1.5}{0.5} = \frac{467}{15} - \frac{114}{15} + \frac{76}{15}x = \frac{353}{15} + \frac{76}{15}x$$

as before. Note also that we can use Equation (15.16) to obtain

$$\text{SS}_{\text{linear}} = \frac{(76)^2}{30} = \frac{2888}{15} = 192.53$$

as shown in Table 15.6.

15.3 CURVILINEAR REGRESSION

When a least squares line is not sufficient to explain all the significant variation in the means, the next logical step is to consider the second-degree (or quadratic) least squares polynomial:

$$y_x' = b_0 + b_1 x + b_2 x^2 \tag{15.17}$$

In fact, careful consideration of the scattergram in Figure 15.1 combined with the significant departure from linear for the reaction time data leads us to believe that such a model will provide a better description of the relationship between dosage and reaction time than that of the least squares straight line.

Formulas for the estimates b_0, b_1, and b_2 are obtained by setting the three partial derivatives for Equation (15.17) to zero and solving the resulting system of normal equations. In Problem 15.37, the reader is asked to show that these equations are

$$\left.\begin{aligned}
\sum_i \sum_j y_{ij} &= b_0 N + b_1 \sum_j n_j x_j + b_2 \sum_j n_j x_j^2 \\
\sum_i \sum_j x_j y_{ij} &= b_0 \sum_j n_j x_j + b_1 \sum_j n_j x_j^2 + b_2 \sum_j n_j x_j^3 \\
\sum_i \sum_j x_j^2 y_{ij} &= b_0 \sum_j n_j x_j^2 + b_1 \sum_j n_j x_j^3 + b_2 \sum_j n_j x_j^4
\end{aligned}\right\} \tag{15.18}$$

The standard Gaussian elimination procedures can be used to solve these equations.