Regression Concepts

1.1   What Is Regression?

Multiple linear regression is a means to express the idea that a response variable, y, varies with a set of independent variables, x_1, x_2, ..., x_m. The variability that y exhibits has two components: a systematic part and a random part. The systematic variation of y is modeled as a function of the x variables. This model relating y to x_1, x_2, ..., x_m is called the *regression equation*. The random part takes into account the fact that the model does not exactly describe the behavior of the response.

1.1.1 Seeing Regression

It is easiest to see the process of regression by looking at a single x variable. In this example, the independent (x) variable is a person's height and the dependent (y) variable is a person's weight. The data are shown in Figure 1.1 and stored in the Small Class.jmp data file.

First, we explore the variables to get a sense of their distribution. Then, we illustrate regression by fitting a model of height to weight.

Figure 1.1
Small
Class.jmp
Data Set

height	weight
44	75
52	95
59	90
65	100
78	99
80	113

An excellent first step in any data analysis (one that is not expressly stated in each analysis of this book) is to examine the distribution of the *x* and *y* variables. Note that this step is not required in a regression analysis, but nevertheless is good practice. To duplicate the display shown here,

� Select **Analyze** → **Distribution**.

� Click on height in the list of columns.

� Click the **Y, Columns** button.

� Repeat these steps for the weight variable.

� Click **OK**.
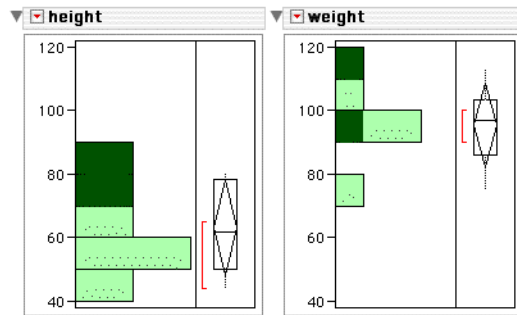

It is clear that the height values are generally less than the weight values. The box plots tell us that there are no outliers or unusual points. However, these unidimensional plots do not tell us much about the relationship between height and weight.

Remember that JMP is an interactive, exploratory tool and that in general, most plots and data are connected to each other. In the case of the two histograms, clicking on a histogram bar in one plot also highlights bars and points in the other plot. The example here shows the result when the two higher height bars are selected, revealing that these high values of height are associated with high values of weight. This fact is a good start in our exploration of the two variables.

It is now appropriate to get a bivariate plot of the two variables. In general, this is accomplished through the Fit Y by X platform, resulting in a graph similar to Figure 1.2.

Figure 1.2
Graph of
weight *by*
height

Platform
menu
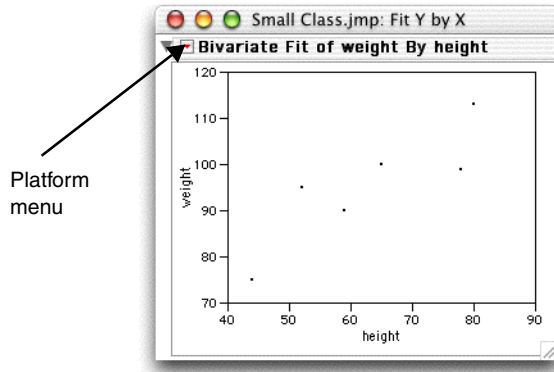
🖰 Select **Analyze → Fit Y by X**.

🖰 Select weight in the list of columns and click the **Y, Response** button.

🖰 Select height in the list of columns and click the **X, Factor** button.

🖰 Click **OK**.

We now see the relationship between weight and height. It appears that weight values increase with corresponding increases in height, and it appears that they increase in a linear way. Therefore, we fit a line to the data. The objective is to find a single line that best summarizes the data—to find a linear equation that predicts values of weight given values of height. To compute and draw this line,

🖰 Select **Fit Line** from the platform menu on the title of the plot (see Figure 1.2).

JMP fits the regression and displays the output shown in Figure 1.3.

Most cases of simple linear regression are accomplished by repeating these steps.

For this introduction, the scatterplot has been coded into a script named demoLeastSquares.jsl. It is used as an illustration of the inner workings of the least-squares method. In practice, you complete the analysis as shown above. The script and accompanying activity are purely pedagogical.

Figure 1.3
Regression
Output
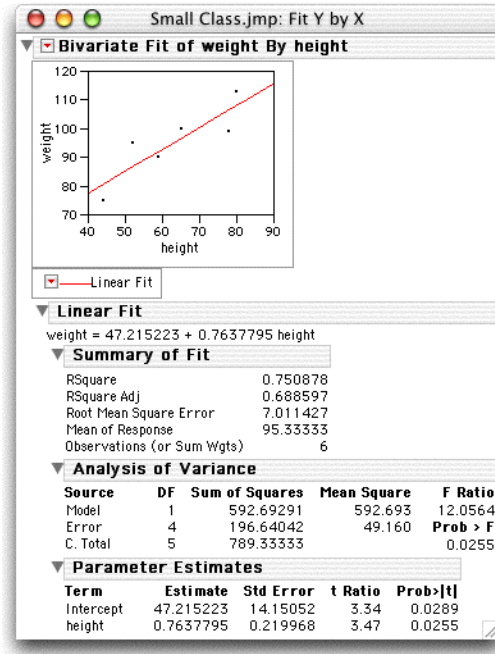
〇 Open the script demoLeastSquares.jsl, found in the Sample Scripts folder. Note that the Open dialog must show files of type JSL, so adjust the dialog to show Data Files, JMP Files, or All Readable documents.

〇 Select **Edit → Run Script**.

〇 When the script output appears, click the **Your Residuals** button (shown in Figure 1.5).

This button turns on a measurement for the accuracy of the linear model. The model itself is represented by an orange line. The distance from this line to each data point is called the *residual*, and it is represented by the dark blue vertical lines that connect the orange line to each data point. See Figure 1.4 for an explanation of the graph produced by the script.

The handles are used to move the model around so that it approximates the shape of the data. As you drag one handle, the other stays fixed. Obviously, many lines can be drawn through the data, each resulting in different residual values. The question is, of all the possible lines that can be drawn through the data, which is the best?
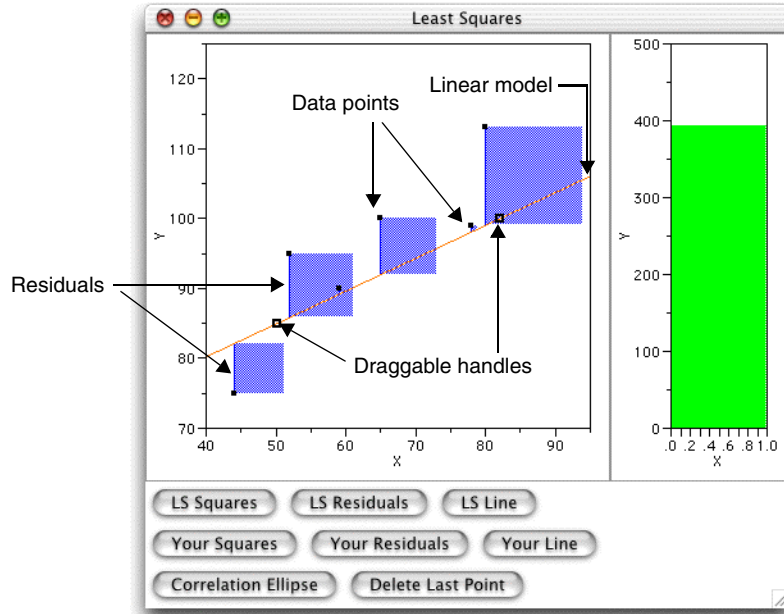
One logical criterion for a "best-fitting" line could be that it is the one where the sum of the residuals is the smallest number. Or, one could claim the best-fitting line is the one that has equal numbers of points above and below the line. Or, one could choose the line that minimizes the perpendicular distances from the points to the line.

In the least-squares method, the squares of the residuals are measured. Geometrically, this value of (residual)^2 can be represented as a square with the length of a side equal to the value of the residual.

⌒ Click the **Your Squares** button to activate the squares of the residuals.

It is this quantity that is minimized. The line with the smallest sum of the squared residuals is used as the best fitting. Geometrically, this means that we seek the line that has the smallest area covered by the squares superimposed on the plot. As an aid to show the sum of the squared residuals, a "thermometer" is appended to the right of the graph (Figure 1.5). This thermometer displays the total area covered by the squared residuals.

⌒ Drag the handles of the model to try to minimize the sum of the squared residuals.

Figure 1.5
Squared
Residuals
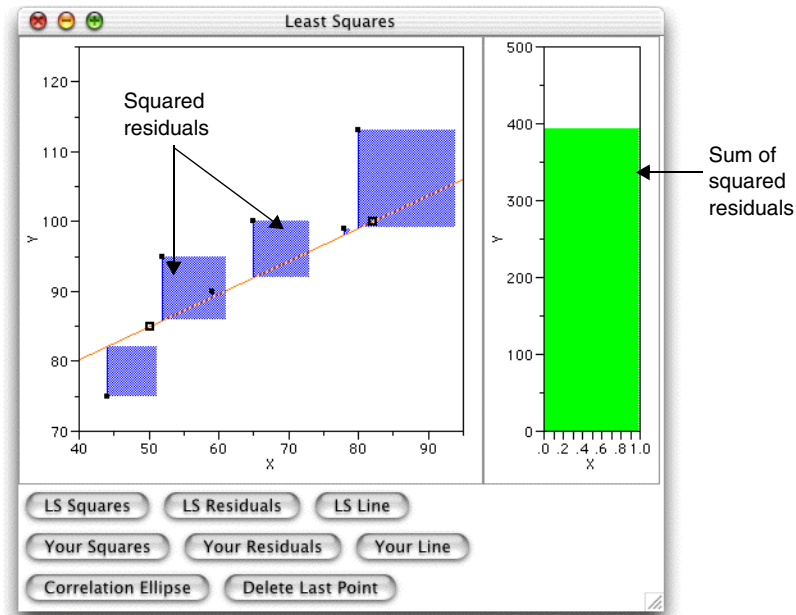
When you think the sum is at a minimum,

[⊕] Push the **LS Line** and **LS Squares** buttons.

The actual (calculated) least-squares line is superimposed upon the graph, along with the squares of its residuals. Additionally, a line is drawn on the thermometer representing the sum of the squared residuals from this line.

[⊕] Drag the handles so that the movable line corresponds exactly to the least-squares line.

Note that the sum of squares now corresponds to the least-squares sum (Figure 1.6). You should notice that as you dragged the handles onto the least-squares line, the sum of the squared residuals decreased until it matched the minimum sum of squares.

[⊕] Drag the handles so that the movable line moves off the least-squares line. Experiment with the line moving away from the least-squares line in all directions.

This illustrates the central concept of least-squares regression. The *least-squares line*, often called the *line of best fit*, is the one that *minimizes the sum of squares of the residuals*.

Figure 1.6
Least-Squares
Line
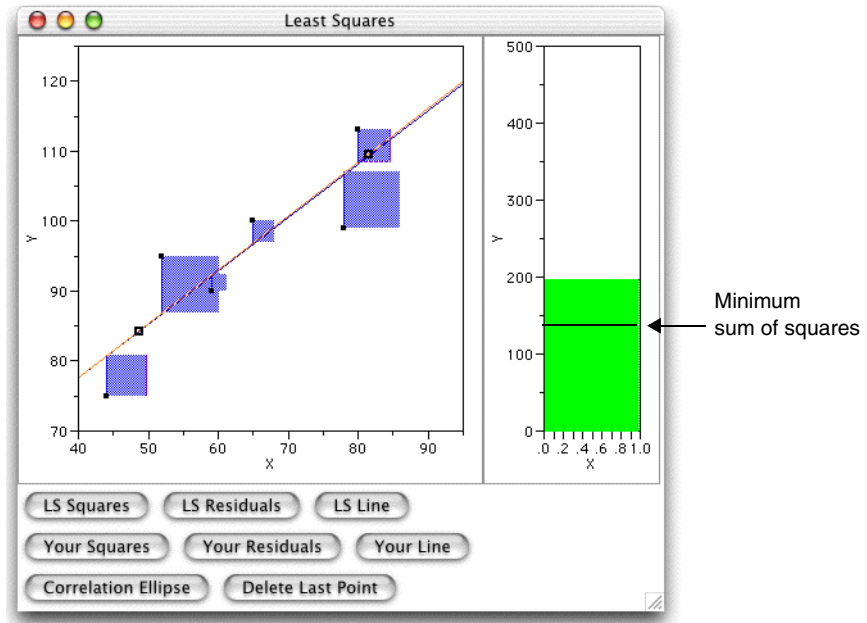
1.1.2 A Physical Model of Regression

The following explanation shows how fitting a least-squares line can be thought of in terms of a physical process. Specifically, it relates regression to properties of springs.

A well-known property of springs is that the force they exert is proportional to the distance they are stretched. Additionally, the energy in the spring is the integral of this force. Specifically, Hooke's law tells us that

$F=kx$

where

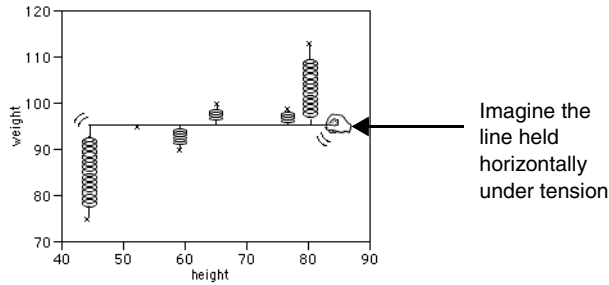F = force exerted by the spring

k = spring constant measuring stiffness of the spring

x = distance the spring is stretched.

For reasons that will become evident later, let the spring constant *k* be represented by $1/\sigma$. Then, the energy in a spring with this value of *k* is

$$E = \int \frac{1}{\sigma} x\, dx = \frac{x^2}{2\sigma}$$

Again consider the height/weight data shown in Figure 1.1.

A graph of weight vs. height is shown here with a line drawn at the mean value of weight. Now, imagine hooking up springs, with one end on a data point and the other end on the mean line. If x_i represents a data point and μ represents the value of the overall mean, then each spring is stretched a distance of $x_i - \mu$.

If the whole system were released, the mean line would rotate and shift to the point of minimum total energy in the system. Mathematically, where μ is the mean of *x* and σ is the standard deviation of *x*,

$$\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma} \sum_{i=1}^{n} (x_i - \mu)^2$$

would be minimized. The constant on the outside of the sum would not come into play with a minimization, so in essence we would minimize

$$\sum_{i=1}^{n} (x_i - \mu)^2$$

This is exactly the expression that is minimized in least-squares fitting. So the physical system of springs is equivalent to fitting a least-squares line. The point where the line is balanced, rotationally and vertically, is the least-squares line.
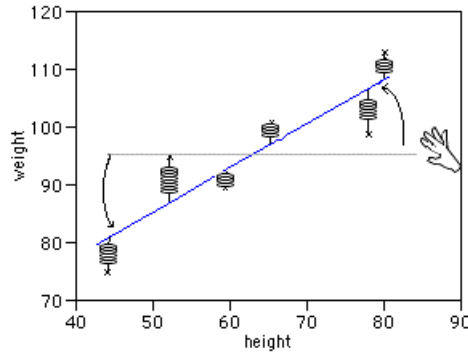
The reason $1/\sigma$ was chosen as the spring constant is that when $k = \sigma$, the expression to be minimized is also equivalent to the log-likelihood of the density of the normal distribution. In other words, the line given by maximizing the log-likelihood is also modeled by this same system.

1.1.3 Multiple Linear Regression

These ideas can be extended to regressions involving several x variables. For two independent variables, the line becomes a plane, and the minimized springs and squares balance the plane in three dimensions.

Although not easy to visualize, the same idea extends to n independent variables, with the same quantities being minimized.

1.1.4 Correlation

It is useful to measure the strength of the relationship between two (or more) variables. Two variables x and y are related if they tend to vary together—if they co-relate. So any measure of correlation should reflect whether they vary together or not. The measure should have a high value when this correlation is strong and a low value when the correlation is weak.

The rules of correlation mimic those of integer multiplication. Factors that are the same (either both positive or both negative) yield a positive product. Similarly, factors that are different yield a negative product. This is similar to our desired measure of correlation, so multiplication of the values of x and y is used in the correlation measure. Of course, x and y might be measured on different scales, so standardized values are used instead of actual values.

The corresponding (standardized) values for each pair are multiplied together, and a weighted average of these products is computed. Stated mathematically,

$$r = \frac{\left(\frac{x_1 - \bar{x}}{s_x}\right)\left(\frac{y_1 - \bar{y}}{s_y}\right) + \left(\frac{x_2 - \bar{x}}{s_x}\right)\left(\frac{y_2 - \bar{y}}{s_y}\right) + \ldots + \left(\frac{x_n - \bar{x}}{s_x}\right)\left(\frac{y_n - \bar{y}}{s_y}\right)}{n - 1}$$

This quantity, termed the *correlation coefficient* and abbreviated r, measures linear correlation in two variables. A similar quantity, known as the coefficient of determination and abbreviated R^2, is used in measuring correlation in multiple regression. In fact, $r^2 = R^2$ for the case of simple regression.
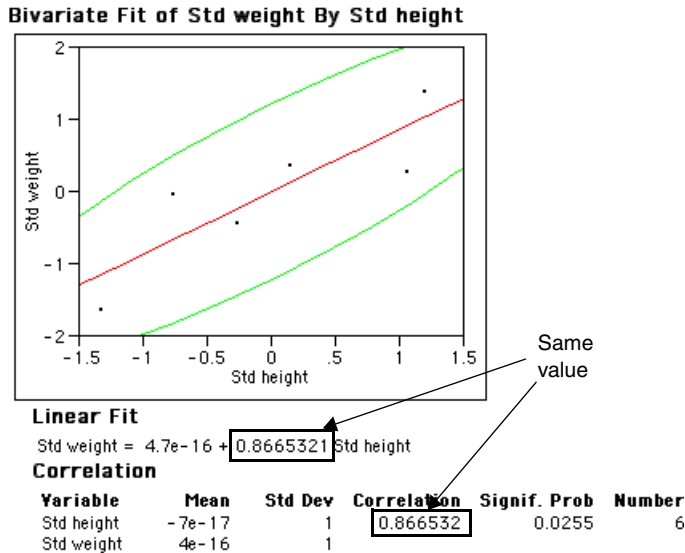
Thinking about this formula reveals some properties of r.

❏   Switching the roles of x and y in the regression does not change the value of r, since multiplication is commutative. This does *not* mean that one can do something similar with the coefficients in regression. Although switching x and y values does not influence correlation, it does affect the regression equation.

❏   Changing the units of x or y does not change the correlation, since standardized scores are used.

❏   The correlation coefficient is sensitive to outliers, since it is based on the mean and standard deviation.

An interesting property of r is that it can be calculated through regression of the standardized variables. Figure 1.7 shows the result of a least-squares regression of standardized values of weight on standardized values of height. When the variables are standardized, note that the slope of the regression line and the correlation coefficient are equal.

Finally, it should be noted that there are other (equivalent) formulas for computing the correlation coefficient r.

Figure 1.7
Standardized
Regression

Bivariate Fit of Std weight By Std height

Linear Fit

Std weight = 4.7e-16 + 0.8665321 Std height

Correlation

Variable	Mean	Std Dev	Correlation	Signif. Prob	Number
Std height	-7e-17	1	0.866532	0.0255	6
Std weight	4e-16	1			

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{(x - \bar{x})^2}\sqrt{(y - \bar{y})^2}}$$

shows the relationship among the sum of the cross-products (in the numerator) and a standardization measure (in the denominator) in a similar way to the discussion above.

$$r = \hat{\beta}\frac{s_x}{s_y}$$

shows the relationship between *r* and the slope of the regression of *y* on *x*. If the standard deviations of *y* and *x* are equivalent (as they are when both variables were standardized above), then *r* is equal to the slope of the regression line.

1.1.5 Seeing Correlation

The demoLeastSquares script used above can also demonstrate correlation.

If there are lines and squares still showing on the plot,

🖰 Click the buttons at the bottom of the screen to toggle all lines and squares off.

🖰 Click the **LS Line** button to turn on the least-squares line.

Now, overlay a correlation ellipse and a label of the correlation coefficient.

🖱 Click the **Correlation Ellipse** button.

The display now appears as in Figure 1.8.

Figure 1.8
Correlation
Ellipse
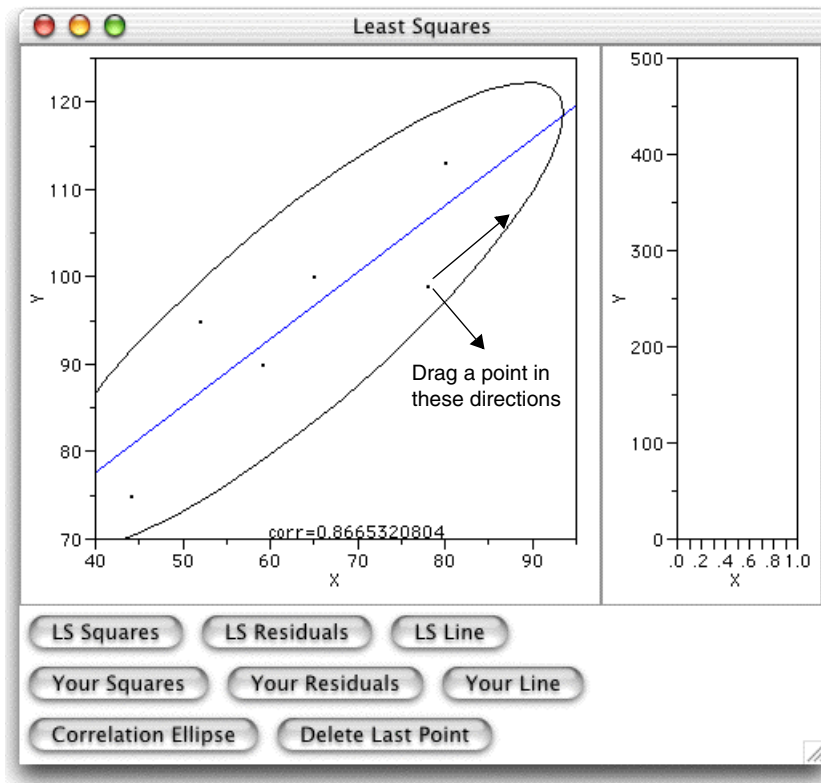
🖱 Drag one of the points in directions parallel to the least-squares line.

Notice that the line does not move much and that the correlation does not change much.

🖱 Drag one of the points in directions perpendicular to the least-squares line.

In this case, both the correlation and least-squares line change drastically. This demonstrates the effect of outliers on the correlation—and why the correlation should always be interpreted in conjunction with a graphical display of the data.

Similarly, points can be added to the plot close to the existing least-squares line with little effect, but single outliers can be added with great effect.

⤷ Click near and away from the regression line to add points and see their effect.

1.2   Statistical Background

This section presents the rigorous definitions of many concepts involved in regression. A background in linear algebra is necessary to follow the discussion. However, it is not necessary to know all of this material to be able to use and interpret the results from JMP. Return to this section as a reference when regression methods are studied in detail.

Formally, multiple linear regression fits a response variable *y* to a function of regressor variables and parameters. The general linear regression model has the form

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \varepsilon$$

where

y is the response, or dependent, variable

$\beta_0, \beta_1, ..., \beta_m$ are unknown parameters

$x_1, x_2, ..., x_m$ are the regressor, or independent, variables

ε is a random error term.

Least squares is a technique used to estimate the parameters. The goal is to find estimates of the parameters $\beta_0, \beta_1, ..., \beta_m$ that minimize the sum of the squared differences between the actual *y* values and the values of *y* predicted by the equation. These estimates are called the *least-squares estimates* and the quantity minimized is called the *error sum of squares*.

Typically, you use regression analysis to do the following:

❏   obtain the least-squares estimates of the parameters

❏   estimate the variance of the error term

❏   estimate the standard error of the parameter estimates

❏   test hypotheses about the parameters

❏   calculate predicted values using the estimated equation

❏   evaluate the fit or lack of fit of the model.

The classical linear model assumes that the responses, *y*, are sampled from several populations. These populations are determined by the corresponding values of $x_1, x_2, ..., x_m$. As the investigator, you select the values of the *x*'s; they are not random. However, the response values are random. You select the values of the *x*'s to meet your experimental needs, carry out the experiment with the set values of the *x*'s, and measure the responses. Often, though, you cannot control the actual values of the independent variables. In these cases, you should at least be able to assume that they are fixed with respect to the response variable.

In addition, you must assume the following:

1.   The form of the model is correct; that is, all important independent variables are included and the functional form is appropriate.

2.   The expected values of the errors (from the mean) are zero.

3.   The variances of the errors (and thus the response variable) are constant across observations.

4.   The errors are uncorrelated.

5.   For hypothesis testing, the errors are normally distributed.

Not all regression models are necessarily linear in their parameters. For example, the model

$$y = \beta_1 e^{\beta_2 x} + \varepsilon$$

is not linear in the parameter β_2. Specifically, the term $e^{\beta_2 x}$ is not a linear function of β_2. This particular nonlinear model, called the *exponential growth* or *decay* model, is used to represent increase (growth) or decrease (decay) over time (*t*) of many types of responses such as population size or radiation counts. Chapter 7, "Nonlinear Models," is devoted to analyses appropriate for this type of model.

Additionally, the random error may not be normally distributed. If this is the case, the least squares technique is not necessarily the appropriate method for estimating the parameters. One such model, the logistic regression model, is presented in the section "Binary Response Variable: Logistic Regression" on page 203.

1.3   Terminology and Notation

The principle of least squares is applied to a set of n observed values of y and the associated x_j to obtain estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_m$ of the respective parameters $\beta_0, \beta_1, ..., \beta_m$. These estimates are then used to construct the fitted model, or estimating equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_m x_m$$

Many regression computations are illustrated conveniently in matrix notation. Let y_i, x_{ij}, and ε_i denote the values of y, x_j, and ε, respectively, in the ith observation. The **Y** vector, the **X** matrix, and the $\boldsymbol{\varepsilon}$ vector can be defined as follows:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_m \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & . & . & x_{1m} \\ . & . & . & . & . \\ . & . & & & . \\ . & . & & & . \\ 1 & x_{n1} & . & . & x_{nm} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_m \end{bmatrix}$$

Then the model in matrix notation is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta}' = (\beta_0\ \beta_1 ... \beta_m)$ is the parameter vector.

The vector of least-squares estimates is

$$\hat{\boldsymbol{\beta}}' = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ . \\ . \\ . \\ \hat{\beta}_m \end{bmatrix}$$

and is obtained by solving the set of normal equations

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

Assuming that **X'X** is of full rank (nonsingular), there is a unique solution to the normal equations given by

$$\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

The matrix $(\mathbf{X'X})^{-1}$ is very useful in regression analysis and is often denoted as follows:

$$(\mathbf{X'X})^{-1} = \mathbf{C} = \begin{bmatrix} c_{00} & c_{01} & \cdot & \cdot & \cdot & c_{0m} \\ c_{10} & c_{11} & \cdot & \cdot & \cdot & c_{1m} \\ \cdot & \cdot & & & \cdot & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ c_{m1} & c_{m2} & \cdot & \cdot & \cdot & c_{mn} \end{bmatrix}$$

1.3.1 Partitioning the Sums of Squares

A basic identity results from least squares, specifically,

$$\sum(y - \bar{y})^2 = \sum(\bar{y} - \hat{y})^2 + \sum(y - \hat{y})^2$$

This identity shows that the total sum of squared deviations from the mean,

$$\sum(y - \bar{y})^2$$

is equal to the sum of squared differences between the mean and the predicted values,

$$\sum(\bar{y} - \hat{y})^2$$

plus the sum of squared deviations from the observed y's to the regression line,

$$\sum(y - \hat{y})^2$$

These two parts are called the regression (or model) sum of squares and the residual (or error) sum of squares. Thus,

Corrected Total SS = Model SS + Residual SS

Corrected total SS always has the same value for a given set of data, regardless of the model that is used; however, partitioning into model SS and residual SS depends on the model. Generally, the addition of a new x variable to a model increases the model SS and, correspondingly, reduces the residual SS. The residual, or error, sum of squares is computed as follows:

$$\text{Residual SS} = \mathbf{Y'}(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'})\mathbf{Y}$$

$$= \mathbf{Y'Y} - \mathbf{Y'X}(\mathbf{X'X})^{-1}\mathbf{X'Y}$$

$$= \mathbf{Y'Y} - \hat{\boldsymbol{\beta}}'\mathbf{X'Y}$$

The error, or residual, mean square

$$s^2 = \text{MSE} = (\text{Residual SS}) / (n - m - 1)$$

is an unbiased estimate of σ^2, the variance of the ε's.

Sums of squares, including the various sums of squares computed by any regression procedure such as the **Fit Y by X** and **Fit Model** commands, can be expressed conceptually as the difference between the regression sums of squares for two models, called *complete* (unrestricted) and *reduced* (restricted) models, respectively. This approach relates a given SS to the comparison of two regression models.

For example, denote as SS_1 the regression sum of squares for a complete model with $m = 5$ variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

Denote the regression sum of squares for a reduced model not containing x_4 and x_5 as SS_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Reduction notation can be used to represent the difference between regression sums of squares for the two models:

$$R(\beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2, \beta_3) = \text{Model SS}_1 - \text{Model SS}_2$$

The difference or reduction in error $R(\beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2, \beta_3)$ indicates the increase in regression sums of squares due to the addition of β_4 and β_5 to the reduced model. It follows that

$$R(\beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2, \beta_3) = \text{Residual SS}_2 - \text{Residual SS}_1$$

which is the decrease in error sum of squares due to the addition of β_4 and β_5 to the reduced model. The expression

$$R(\beta_4, \beta_5 \mid \beta_0, \beta_1, \beta_2, \beta_3)$$

is also commonly referred to in the following ways:

❏ the sums of squares due to β_4 and β_5 (or x_4 and x_5) adjusted for β_0, β_1, β_2, β_3 (or the intercept and x_1, x_2, x_3)

❏ the sums of squares due to fitting x_4 and x_5 after fitting the intercept and x_1, x_2, x_3

❏ the effects of x_4 and x_5 above and beyond (or partialing) the effects of the intercept and x_1, x_2, x_3.

1.3.2 Hypothesis Testing

Inferences about model parameters are highly dependent on the other parameters in the model under consideration. Therefore, in hypothesis testing, it is important to emphasize the parameters for which inferences have been adjusted. For example, $R(\beta_3 \mid \beta_0, \beta_1, \beta_2)$ and $R(\beta_3 \mid \beta_0, \beta_1)$ may measure entirely different concepts. Similarly, a test of H_0: $\beta_3 = 0$ may have one result for the model

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$$

and another for the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Differences reflect actual dependencies among variables in the model rather than inconsistencies in statistical methodology.

Statistical inferences can also be made in terms of linear functions of the parameters of the form

$$H_0: \mathbf{L}\boldsymbol{\beta}_0: l_0\beta_0 + l_1\beta_1 + ... + l_m\beta_m = 0$$

where the l_i are arbitrary constants chosen to correspond to a specified hypothesis. Such functions are estimated by the corresponding linear function

$$\mathbf{L}\boldsymbol{\beta} = l_0\beta_0 + l_1\beta_1 + ... + l_m\beta_m$$

of the least-squares estimates $\hat{\boldsymbol{\beta}}$. The variance of $\mathbf{L}\hat{\boldsymbol{\beta}}$ is

$$V(\mathbf{L}\hat{\boldsymbol{\beta}}) = (\mathbf{L}(\mathbf{X'X})^{-1}\mathbf{L'})\sigma^2$$

A t test or F test is used to test H_0: $(\mathbf{L}\boldsymbol{\beta}) = 0$. The denominator is usually the residual mean square (MSE). Because the variance of the estimated function is

based on statistics computed for the entire model, the test of the hypothesis is made in the presence of all model parameters. Confidence intervals can be constructed to correspond to these tests, which can be generalized to simultaneous tests of several linear functions.

Simultaneous inference about a set of linear functions $\mathbf{L}_1\boldsymbol{\beta}$, ..., $\mathbf{L}_k\boldsymbol{\beta}$ is performed in a related manner. For notational convenience, let $\mathbf{L}$ denote the matrix whose rows are $\mathbf{L}_1$, ..., $\mathbf{L}_k$:

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_1 \\ . \\ . \\ . \\ \mathbf{L}_k \end{bmatrix}$$

Then the sum of squares

$$\mathrm{SS}(\mathbf{L}\boldsymbol{\beta} = 0) = (\mathbf{L}\boldsymbol{\beta})'(\mathbf{L}(\mathbf{X'X})^{-1}\mathbf{L}')(\mathbf{L}\boldsymbol{\beta})$$

is associated with the null hypothesis

$$H_0: \mathbf{L}_1\boldsymbol{\beta} = ... = \mathbf{L}_k\boldsymbol{\beta} = 0$$

A test of H_0 is provided by the F statistic

$$F = (\mathrm{SS}(\mathbf{L}\boldsymbol{\beta} = 0) / k) / \mathrm{MSE}$$

Three common types of statistical inferences follow:

1. A test that all parameters ($\beta_1, \beta_2, ..., \beta_m$) are zero. The test compares the fit of the complete model to that using only the mean:

   $$F = (\text{Model SS} / m) / \mathrm{MSE}$$

   where

   $$\text{Model SS} = R(\beta_0, \beta_1, ..., \beta_m \mid \beta_0)$$

   The F statistic has $(m, n - m - 1)$ degrees of freedom.

   Note that $R(\beta_0, \beta_1, ..., \beta_m)$ is rarely used. For more information on no-intercept models, see the section "Regression through the Origin" on page 44.

2. A test that the parameters in a subset are zero. The problem is to compare the fit of the complete model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_g x_g + \beta_{g+1} x_{g+1} + ... + \beta_m x_m + \varepsilon$$

to the fit of the reduced model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_g x_g + \varepsilon$$

An F statistic is used to perform the test

$$F = (R(\beta_{g+1}, ..., \beta_m \mid \beta_0, \beta_1, ..., \beta_g) / (m - g))\text{MSE}$$

Note that an arbitrary reordering of variables produces a test for any desired subset of parameters. If the subset contains only one parameter, β_m, the test is

$$F = (R(\beta_m \mid \beta_0, \beta_1, ..., \beta_{m-1}) / 1) / \text{MSE} = (\text{partial SS due to } \beta_m) / \text{MSE}$$

which is equivalent to the t test

$$t = \frac{\hat{\beta}_m}{s_{\hat{\beta}_m}} = \frac{\hat{\beta}_m}{\sqrt{c_{mm}\text{MSE}}}$$

The corresponding $(1 - \alpha)$ confidence interval about β_m is

$$\beta_m \pm t_{\alpha/2}\sqrt{c_{mm}\text{MSE}}$$

3. Estimation of a subpopulation mean corresponding to a specific x. For a given set of x values described by a vector $\mathbf{x}$, denote the population mean by μ_x. The estimated population mean is

$$\mu_x = \beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_m \mathbf{x}_m = \mathbf{x}\beta$$

The vector $\mathbf{x}$ is constant; hence, the variance of the estimate $\hat{\mu}_{\mathbf{x}}$ is

$$V(\hat{\mu}_x) = \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\sigma^2$$

This equation is useful for computing confidence intervals. A related inference concerns a future single value of y corresponding to a specified x. The relevant variance estimate is

$$V(\hat{y}_x) = (1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x})\sigma^2$$

1.3.3 Using the Generalized Inverse

Many applications of regression procedures involve an **X'X** matrix that is not of full rank and has no unique inverse. Fit Model and Fit Y by X compute a generalized inverse $(\mathbf{X'X})^-$ and use it to compute a regression estimate

$$\mathbf{b} = (\mathbf{X'X})^- \, \mathbf{X'Y}$$

A generalized inverse of a matrix **A** is any matrix **G** such that **AGA** = **A**. Note that this also identifies the inverse of a full-rank matrix.

If **X'X** is not of full rank, then an infinite number of generalized inverses exist. Different generalized inverses lead to different solutions to the normal equations that have different expected values; that is, $E(\mathbf{b}) = (\mathbf{X'X})^- \, \mathbf{X'Y}\beta$ depends on the particular generalized inverse used to obtain **b**. Therefore, it is important to understand what is being estimated by the solution.

Fortunately, not all computations in regression analysis depend on the particular solution obtained. For example, the error sum of squares is invariant with respect to $(\mathbf{X'X})^-$ and is given by

$$\text{SSE} = \mathbf{Y'(1 - X(X'X)^- X')Y}$$

Hence, the model sum of squares also does not depend on the particular generalized inverse obtained.

The generalized inverse has played a major role in the presentation of the theory of linear statistical models, notably in the work of Graybill (1976) and Searle (1971). In a theoretical setting, it is often possible, and even desirable, to avoid specifying a particular generalized inverse. To apply the generalized inverse to statistical data using computer programs, a generalized inverse must actually be calculated. Therefore, it is necessary to declare the specific generalized inverse being computed. For example, consider an **X'X** matrix of rank k that can be partitioned as

$$\mathbf{X'X} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where $\mathbf{A}_{11}$ is $k \times k$ and of rank k. Then A_{11}^{-1} exists, and a generalized inverse of **X'X** is

$$(\mathbf{X'X})^- = \begin{bmatrix} A_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \end{bmatrix}$$

where each φ_{ij} is a matrix of zeros of the same dimension as $\mathbf{A}_{ij}$.

This approach to obtaining a generalized inverse, the method used by Fit Model and Fit Y by X, can be extended indefinitely by partitioning a singular matrix into several sets of matrices as illustrated above. Note that the resulting solution to the normal equations, $\mathbf{b} = (\mathbf{X'X})^-\mathbf{X'Y}$, has zeros in the positions corresponding to the rows filled with zeros in $(\mathbf{X'X})^-$. This is the solution printed by these procedures, and it is regarded as providing a biased estimate of $\mathbf{b}$.

However, because $\mathbf{b}$ is not unique, a linear function, $\mathbf{Lb}$, and its variance are generally not unique either. However, a class of linear functions called *estimable functions* exists, and they have the following properties:

❏   The vector $\mathbf{L}$ is a linear combination of rows of $\mathbf{X}$.

❏ $\mathbf{Lb}$ and its variance are invariant through all possible generalized inverses. In other words, $\mathbf{Lb}$ is unique and is an unbiased estimate of $\mathbf{L\beta}$.

Analogous to the full-rank case, the variance of an estimable function $\mathbf{Lb}$ is given by

$$V(\mathbf{Lb}) = (\mathbf{L}(\mathbf{X'X})^-\mathbf{L'})\sigma^2$$

This expression is used for statistical inference. For example, a test of H_0: $\mathbf{L\beta} = 0$ is given by the *t* test

$$t = \frac{\mathbf{Lb}}{\sqrt{\mathbf{L}(\mathbf{X'X})^-\mathbf{L'}}\,\text{MSE}}$$

Simultaneous inferences on a set of estimable functions are performed in an analogous manner.

1.4 Regression with JMP

This section reviews the following JMP platforms that are useful in regression analysis.

Note   The two platforms for most regression analyses are Fit Y by X and Fit Model.

Fit Y by X
fits general linear models of two variables. Its capabilities include simple regression, analysis of variance, contingency table analysis, and logistic regression. In addition, the regression portion allows for orthogonal fits that take variation in x variables into account and for the fitting of cubic splines.

Fit Model
fits general linear models. In addition to many other analyses, Fit Model can perform simple, multiple, polynomial, and weighted regression, as well as analysis of variance and analysis of covariance. It is also used for logistic regression and stepwise regression, and for the analysis of response surface models.

Distribution
provides univariate statistics. Although it does not compute regressions itself, it is an essential platform for in-depth regression analyses. It is useful for preliminary analyses of variables, as well as for plotting residuals to assess their distribution (shape).

Multivariate
analyzes multivariate data, providing several types of correlations and three different methods of principal components.

Survival
fits univariate and regression models to failure-time data that may be right-, left-, or interval-censored. These types of models are commonly used in survival analysis.

Nonlinear
fits nonlinear regression models. The iterative method is a modified Gauss-Newton.

PLS
analyzes data with Partial Least Squares.

Neural Net
is used for prediction when the form of the model is not important.

Finally, if a regression method cannot be performed by any of the platforms above, JMP provides a scripting language with extensive matrix abilities that can be used.