2 Simple Models

Definitions of Error and Parameter Estimates

In this chapter we consider the very simplest models—models with one or even no parameters. These simple models make the same prediction for all data observations; there are no differential predictions conditioned on whatever else we might know about each observation. Such a simple model may not seem very realistic or useful. However, this simple model provides a useful baseline against which we can compare more complicated models, and it will turn out to be more useful than it might appear at first. For many of the questions we will want to ask about our data, the appropriate Model C will be the simple model with no parameter or one parameter; this will be compared to the more complex Model A. Also, the simple model provides a useful first-cut description of data.

OVERVIEW OF THE SIMPLE MODEL

Formally, the simplest model is:

$$Y_i = B_0 + \varepsilon_i$$

where B_0 is some specified value not based on this particular batch of data (i.e., it is a specific a priori numeric value), and ε_i is the true error or the amount by which Y_i differs from B_0. This simple model in which no parameters are estimated from the data is not frequently used in the social sciences because we seldom have theories sufficiently powerful to make an explicit prediction for a parameter. Such models are much more common in fields such as biology. For example, a medical study measuring human body temperature might reasonably consider the model: except for error, all temperatures are 37°C. In this case, $B_0 = 37$ so the formal model would be:

$$Y_i = 37 + \varepsilon_i$$

Although we will sometimes want to consider a specified or hypothesized value of B_0 in order to ask an explicit question about data, it is much more common to consider the equation:

$$Y_i = \beta_0 + \varepsilon_i$$

where β_0 is a true parameter that is estimated from the data. Continuing with the medical example, suppose that the body temperatures were all from people who had taken a certain drug. We might suspect that, except for error, they all have the same body temperature, but it is not the usual body temperature of 37°C. We use β_0 to represent whatever the

body temperature might be for those who have taken the drug. It is important to realize that β_0 is unknowable; we can only estimate it from the data. In terms of these true values, ε_i is the amount by which Y_i differs from β_0 if we were ever to know β_0 exactly. We use b_0 to indicate the estimate of β_0 that we derive from the data. Then the predicted value for the ith observation is:

$$\hat{Y}_i = b_0$$

and

$$\text{DATA= MODEL + ERROR}$$

becomes

$$Y_i = b_0 + e_i$$

where e_i is the amount by which our prediction misses the actual observation. Thus, e_i is the estimate of ε_i. The goal of tailoring the model to provide the best fit to the data is equivalent to making the errors:

$$e_i = Y_i - b_0$$

as small as possible. We have only one parameter, so this means that we want to find the estimate b_0 for that one parameter β_0 that will minimize the errors. However, we are really interested not in each e_i but in some aggregation of all the individual e_i values. There are many different ways to perform this aggregation. In this chapter, we consider some of the different ways of aggregating the separate e_i into a summary measure of the error. Then we show how each choice of a summary measure of the error leads to a different method of calculating b_0 to estimate β_0 so as to provide the best fit of the data to the model. Finally, we consider expressions that describe the "typical" error.

Measures of location or *measures of central tendency* are the traditional names for the parameter estimates of β_0 developed from the different definitions of error. These names are appropriate because the parameter estimate in the simple model tells us about the location of a typical observation or about the center of a batch of data. Specific instances include the mode, median, and mean. *Measures of variability* or *measures of spread* are the traditional names for the expressions for typical errors. These names are appropriate because expressions for typical errors tell us how variable the observations are in a batch of data or, equivalently, how far the data spread out from the center. Specific instances include the median absolute deviation and standard deviation. Together, measures of central tendency and spread are known as *descriptive statistics*. However, this suggests a false distinction between these statistics and those to come later. We want to emphasize that the parameter estimates for β_0 in the simple model are no more nor less descriptive than the parameter estimates we will develop for more complicated models. Models and their parameter estimates always provide descriptions of data. Hence, we will generally avoid the phrase "descriptive statistics" and just refer to parameter estimates. The reader should be aware, however, that when other textbooks refer to descriptive statistics they are generally referring to the material in this chapter.

CONCEPTUAL EXAMPLE

Before considering simple models and measures of error more formally, we consider some conceptual examples that will help to build useful intuitions so that the subsequent mathematical formulas will be less abstract. Suppose that the data consisted of the five observations 1, 3, 5, 9, 14, representing the number of books read over the summer by each of five elementary school students. These observations are plotted in Figure 2.1.

The simple model makes the same prediction for all five observations. The horizontal line represents the value of that constant prediction. The vertical lines drawn from each observation to the prediction line represent the amount by which the prediction misses the actual data value. In other words, the length of the line is e_i. One way to find the best value for $\hat{Y}$ and, equivalently, the best estimate b_0 for β_0 is to adjust the $\hat{Y}$ line up or down so that the length of the lines is a minimum. (Note that we have dropped the subscript i from $\hat{Y}_i$ because all the predictions are the same for the simple model.) In other words, we can use trial and error to find the best estimate. For example, we might want to try 7 as our estimate. The data, the prediction line for $\hat{Y} = b_0 = 7$, and the errors are graphed in Figure 2.2. Note that the five line lengths are now 6, 4, 2, 2, 7, with a sum of 21.

For an estimate of 5, the line lengths were 4, 2, 0, 4, 9, with a sum of 19. The estimate $b_0 = 5$ thus produces less total error than $b_0 = 7$, so we can eliminate 7, in favor of 5, as an estimate if our goal is to minimize the total error. We can continue to try other estimates until we find the best one. Figure 2.3 shows the sum of the line lengths for different choices of b_0 between 0 and 10. The sum of the line lengths reaches a minimum of 19 when $b_0 = 5$, so that is our best estimate of β_0. We would get more total error, a larger sum of line lengths, if we used a value of b_0 that was either lower or higher than 5. Hence, $b_0 = 5$ is the optimum estimate. Note that 5 is the middle of our five observations; the middle observation in a batch of data that have been sorted from smallest to largest is often called the *median*.

It is interesting to ask how we would have to adjust the estimate b_0 if one of the observations were dramatically changed. For example, what if the 14 were replaced by 140 so that the five observations were 1, 3, 5, 9, and 140? Before reading on, test your intuitions by guessing what the new value for b_0 will be. Figure 2.4 shows the sum of the line lengths for different possible values of b_0. Although all of the sums are much larger than before, the minimum of 145 still occurs when $b_0 = 5$! The middle or median

FIGURE 2.1 Error as the sum of line lengths (estimate is $\hat{Y} = b_0 = 5$)

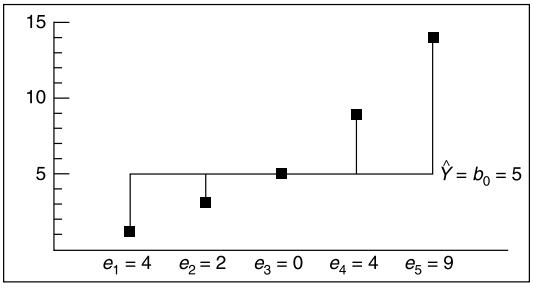


FIGURE 2.2 Error as the sum of line lengths (estimate is $\hat{Y} = b_0 = 7$)

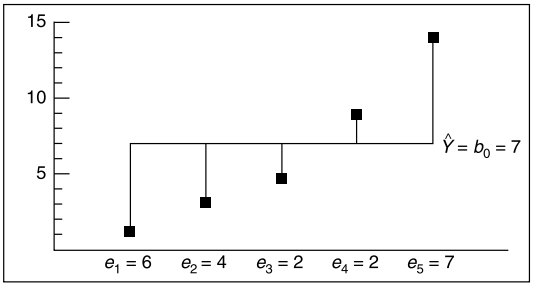


FIGURE 2.3 Sum of absolute error (SAE) as a function of $\hat{Y}_i$

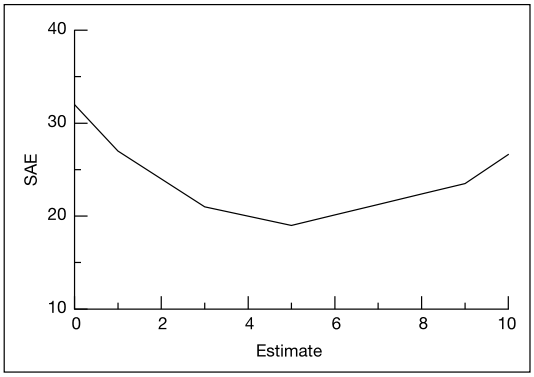
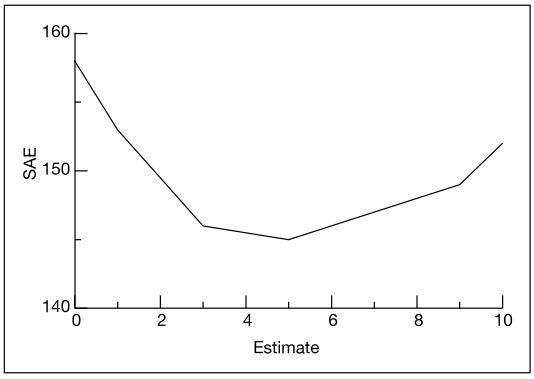


FIGURE 2.4 Sum of absolute error (SAE) as a function of $\hat{Y}_i$ with extreme observation (140)



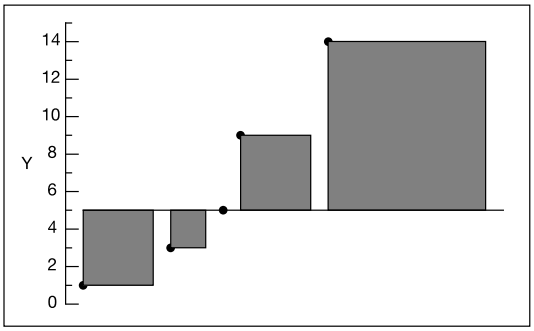
observation is still the best estimate even though one of the observations has been increased by a factor of 10.

Above we used the sum of the error line lengths as an aggregate summary index of error. This simple sum may not always be reasonable. For example, the simple sum implies that several small errors (e.g., four errors of length 1) are equivalent to one large error (e.g., one error of length 4). Instead, we may want to charge a higher penalty in the error index for big errors so that an error of 4 counts more than four errors of 1. One way to accomplish this is to square the line lengths before summing them. For example, $4^2 = 16$ adds a lot more to the error sum than $1^2 + 1^2 + 1^2 + 1^2 + 4$. Figure 2.5 depicts the original set of five observations with this new definition of error; each e_i is now represented by a square, the length of whose side is determined by the distance between the observation and the horizontal line representing the constant prediction $\hat{Y}_i$ of the simple model. The aggregate error is simply the sum of those squares.

Again, we can use brute force to find a value of $\hat{Y}_i$ and b_0 that will make that sum of squares as small as possible. Let us consider the possible estimates of 5 and 7, which we evaluated when we were using the sum of the line lengths as the error measure. For $b_0 = 5$, the five areas of the squares are:

$$4^2 = 16, \quad 2^2 = 4, \quad 0^2 = 0, \quad 4^2 = 16, \quad 9^2 = 81$$

FIGURE 2.5 Error as the sum of squares (estimate is $\hat{Y}_i = b_0 = 5$)



and the sum of those squares is 117. For $b_0 = 7$, the five areas are 36, 16, 4, 4, 49, and the sum of squares is 109. So $b_0 = 5$, which was the best estimate when using line lengths, is no longer the best estimate when we use squares, because $b_0 = 7$ produces a smaller sum of squares or smaller error. Figure 2.6 shows the sum of squares for different possible values of b_0 between 0 and 10. The best value for b_0 is about 6.4 with a minimum sum of squares of about 107. The estimates of 5 and 7 are not bad—sums of squares of 117 and 109, respectively—but clearly inferior to the optimum estimate of 6.4. Although not obvious, we will prove later that the best estimate when using squared errors is simply the arithmetic average or *mean* of the observations. For the five observations:

$$\frac{1 + 3 + 5 + 9 + 14}{5} = \frac{32}{5} = 6.4$$

which produces b_0, the best estimate of β_0.

It is interesting to ask again what would happen to the estimate b_0 if one of the observations were dramatically changed: say, the 14 were replaced by 140. Before reading on, again check your intuition by guessing the new estimate b_0. Figure 2.7 shows the sum of squares for the revised set of observations. The minimum sum of squares no longer occurs when $b_0 = 6.4$; instead, the minimum now occurs when b_0 is a whopping 31.6, which is again the average of the five observations. But note that although that estimate is the best, it is not very good, with a total sum of squares of about 14,723.

Before formalizing these examples in equations, it is useful to summarize the concepts introduced. First, the best estimate b_0 for a simple, one-parameter model is the constant prediction $\hat{Y}$, which minimizes the sum of the errors. Although we will generally have better ways to estimate the parameter than brute force, it is important to realize that the best estimate could be found by trial and error until we could no longer make the error any smaller. In fact, computer programs sometimes use precisely this strategy. Second, the choice of a method for summarizing or expressing the error—the lengths of the error lines or their squares in the above examples—affects the best estimate b_0. We will soon see that there are many other plausible choices for error terms. Third, if total error is the sum of the line lengths, then the median of the observations provides the best estimate b_0, where the median is simply the middle observation. Fourth, if total error is the sum of the squared line lengths, then the best estimate b_0 is the arithmetic

FIGURE 2.6 Sum of squared errors (SSE) as a function of $\hat{Y}_i$

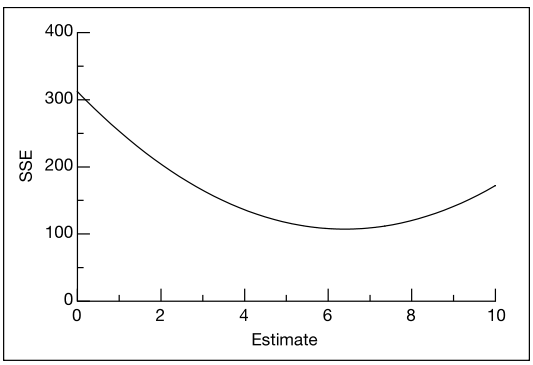
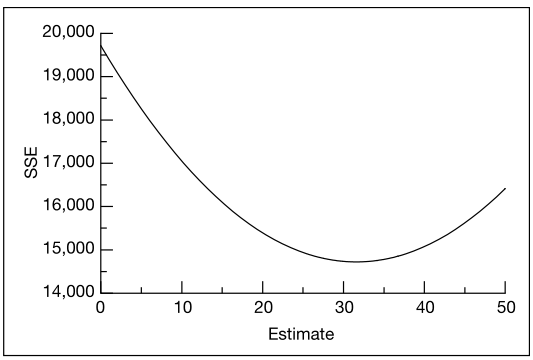


FIGURE 2.7 Sum of squared errors (SSE) as a function of $\hat{Y}_i$ with extreme observation (140)



average or mean of the observations. Fifth, the median does not change when an extreme observation is made more extreme, but the mean can change dramatically. Sixth, many times in this book we will encounter the phrase "sum of squares"; it is useful to realize that it can indeed be represented geometrically as a literal summation of squares.

FORMALITIES FOR SIMPLE MODELS

As always, we begin with the basic equation for data analysis:

DATA = MODEL + ERROR

For simple models, MODEL states that all observations in DATA are essentially the same, so:

$$Y_i = \beta_0 + \varepsilon_i$$

where β_0 represents the true, but unknown, parameter and ε_i represents the individual error disturbances around the unknown parameter. Then b_0 is the estimate of that parameter based on the data at hand. So the actual model used for predicting each Y_i becomes:

MODEL: $\hat{Y}_i = b_0$

The basic data analysis equation can then be written as:

$$Y_i = \hat{Y}_i + e_i \quad \text{or} \quad Y_i = b_0 + e_i$$

The error or residual associated with each observation is then simply the difference between the data and the model prediction, or:

$$e_i = Y_i - \hat{Y}_i$$

Our problem is how to select a single value for b_0 to represent all of the data. Clearly, we want b_0 to be in the "center" of the data so that it will be more or less close to all of the observations. Hence, estimates of b_0 are often called *measures of central tendency*. But we need to be much more precise about defining the center and what we mean by "close." As always, the key is to make e_i as small as possible. Instead of looking at each individual e_i, we need to consider ways of aggregating the separate e_i values into a summary measure of the total error. Once we have done that, it should be a simple procedure to choose a b_0 that will make the summary measure of error as small as possible. We now turn to a consideration of possible summary measures.

Count of errors (CE)

One possibility is simply to count the number of times Y_i does not equal $\hat{Y}_i$. This ignores the size of the individual errors and only counts whether an error occurred. Formally:

$$\text{ERROR} = \sum_{i=1}^{n} I(e_i) = \sum_{i=1}^{n} I(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} I(Y_i - b_0)$$

where $I(e_i) = 1$ if $e_i = 0$ and $I(e_i) = 0$ if $e_i = 0$. Functions such as I () are often called *indicator* functions and are simply a fancy way of representing whether or not something is to be included in an equation.

Sum of errors (SE)

In order not to lose information about the size of the e_i, we might add all the e_i values so that:

$$\text{ERROR} = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = \sum_{i=1}^{n} (Y_i - b_0)$$

But this is clearly unsatisfactory because it allows positive and negative errors to cancel one another. For example, suppose that b_0 underestimated one observation by 1000 and overestimated another observation by the same amount so that $e_1 = 1000$ and $e_2 = -1000$. Adding those two errors would produce zero, incorrectly implying that there was no error. We have no reason to be more interested in overestimates than underestimates, so one solution is to ignore the positive and negative signs of the e_i.

Sum of absolute errors (SAE)

One way to remove the signs is to sum the absolute values of the errors:

$$\text{ERROR} = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |Y_i - \hat{Y}_i| = \sum_{i=1}^{n} |Y_i - b_0|$$

The sum of absolute errors is the formal equivalent of summing the line lengths in Figure 2.1. As noted in the conceptual example above, it may not always be desirable to count one big error (e.g., an error of 4) as being the same as the equivalent amount of small errors (e.g., four errors of 1). The conceptual example therefore suggests the next measure of error.

Sum of squared errors (SSE)

Another way to remove the signs from the e_i is to square each one before summing them:

$$\text{ERROR} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0)^2$$

The sum of squared errors is the formal equivalent of adding up the error squares in Figure 2.5. Besides removing the signs, the squaring has the additional effect of making large errors more important.

Weighted sum of squared errors (WSSE)

So far, all of the suggested error measures have given equal weight to each observation. For a variety of reasons we may want to give more weight to some observations and less weight to others when calculating the aggregate error. For example, we may have reason to believe that certain observations are questionable or suspect because they were collected with less precision or less reliability than the other observations. Or we may not want to count an error of 10 as being the same when $\hat{Y}_i = 1000$ as when $\hat{Y}_i = 5$. In the former instance, the error of 10 amounts to only a 1% error, while in the latter

instance it is an error of 200%. Or, finally, we might be suspicious of a couple of extreme observations just because they are "outliers" with respect to other observations. Whatever our reasons, it is easy to incorporate a weight w_i for each observation into the formal definition:

$$\text{ERROR} = \sum_{i=1}^{n} w_i e_i^2 = \sum_{i=1}^{n} w_i (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} w_i (Y_i - b_0)^2$$

The weights w_i might be assigned a priori based on judgments of data quality or some formal index of each observation's reliability. One possibility is to weight all observations equally, in which case $w_i = 1$ for all i. This weighted sum of squared errors becomes simply the sum of squared errors above.

Statisticians have created some very clever ways of defining weights to solve a variety of complicated problems. We will encounter examples of those weights later in the context of specific problems. For now, just be aware that the use of weights gives us a great deal of flexibility in defining the aggregate measure of error.

ESTIMATORS OF β_0

As demonstrated by the conceptual examples presented earlier in this chapter, the choice of a method for aggregating the e_i influences the estimate b_0. It should not be surprising, therefore, that for each definition of aggregate error presented above there is a different way of calculating b_0 from the data. We could use the brute-force method for each definition of error by trying different values of b_0 until we found the one that gave the minimum value for error. However, it turns out that for each of the above definitions of error we can define a way of calculating b_0 from the data so that b_0 is guaranteed to produce the minimum possible value for error. While we will almost always use the calculation method, it is important to remember that the definition of the best estimate of β_0 is the value of b_0 that produces the least error. We list in Figure 2.8 the definition of b_0 for each definition of error considered so far.

Proof that the Mean Minimizes SSE

In this section, we present a formal proof that the mean $\bar{Y}$ does indeed produce the smallest possible value of SSE. Although we generally avoid proofs, we think it is important to understand that the choice of the mean as an estimator is not arbitrary; instead, the choice of SSE as an aggregate measure of ERROR also dictates the choice of the mean as the best estimator. Similar proofs can be given to show that the indicated estimator minimizes the corresponding definition of error in the list in Figure 2.8.

FIGURE 2.8 Estimators for each definition of error

Error definition	b_0, *estimator of* β_0
Count of errors	Mode = most frequent value of Y_i
Sum of absolute errors	Median = middle observation of all the Y_i
Sum of squared errors	Mean = average of all the Y_i
Weighted sum of squared errors	Weighted mean = weighted average of all the Y_i

Let us begin by assuming that $\hat{Y}$, the best estimator for reducing the sum of squared errors, is something other than

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

(Note that here $\hat{Y}$ does not have an *i* subscript because we make the same prediction for all observations in the simple model.) Our strategy is to proceed with that assumption until we reach a point where we can see that it is a bad assumption. At that point we will know that the only reasonable assumption is $\hat{Y} = \bar{Y}$.

The sum of squared errors is:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

Obviously, $\bar{Y} - \bar{Y} = 0$; we can add zero within the parentheses without changing the sum of squared errors. That is:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

Rearranging the terms slightly, we get:

$$\text{SSE} = \sum_{i=1}^{n} [(Y_i - \bar{Y}) + (\bar{Y} - \hat{Y})]^2$$

Squaring gives:

$$\text{SSE} = \sum_{i=1}^{n} [(Y_i - \bar{Y})^2 + 2(Y_i - \bar{Y})(\bar{Y} - \hat{Y}) + (\bar{Y} - \hat{Y})^2]$$

Breaking the sums apart yields:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + \sum_{i=1}^{n} [2(Y_i - \bar{Y})(\bar{Y} - \hat{Y})] + \sum_{i=1}^{n} (\bar{Y} - \hat{Y})^2$$

The last term contains no subscripts so the summation is equivalent to adding up the same quantity *n* times; hence, the summation sign can be replaced by multiplication by *n*. Similarly, the quantities without subscripts in the middle term can be taken outside the summation sign to give:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + 2(\bar{Y} - \hat{Y}) \sum_{i=1}^{n} (Y_i - \bar{Y}) + n(\bar{Y} - \hat{Y})^2$$

Now let us concentrate on the middle term. Note that:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \bar{Y}) &= \sum_{i=1}^{n} Y_i - n\bar{Y} \\ &= \sum_{i=1}^{n} Y_i - n\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) \\ &= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} Y_i \\ &= 0\end{aligned}$$

Hence, the middle term of SSE includes a multiplication by zero, which eliminates that term. We are left with:

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + n(\bar{Y} - \hat{Y})^2$$

We want SSE to be as small as possible. We have no freedom in the first term: Y_i and $\bar{Y}$ are whatever they happen to be. We do, however, have freedom to choose $\hat{Y}_i$ in the second term to make SSE as small as possible. Clearly, $n(\bar{Y} - \hat{Y})^2$ is positive, so it is making SSE larger. But if we let $\hat{Y} = \bar{Y}$, then $n(\bar{Y} - \hat{Y})^2 = 0$ and we are no longer adding anything extra to SSE. For any estimate of $\hat{Y}$ other than $\bar{Y}$, we will be making SSE larger. Hence, SSE is as small as possible when $\hat{Y} = \bar{Y}$, and the minimum is:

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Any other choice for $\hat{Y}$ would produce a larger SSE; thus, the mean is the best estimator for reducing SSE in the simple model.

Describing Error

If the goal is simply to describe a batch of data, then we can apply the simple model and use one or all of the measures of central tendency—mode, median, and mean—as descriptive statistics. When doing so, it is also useful to present a description of the typical error. Reporting the total error (e.g., SAE or SSE) is not desirable because the total depends so heavily on the number of observations. For example, aggregate error based on 50 observations is likely to be larger than aggregate error based on 15 observations, even if the typical errors in the former case are smaller. For each measure of central tendency there is a corresponding customary index of the typical error. We consider each below.

Modal error

When we use the count of the errors (CE) as the aggregate index of error, there can only be two values for e_i: either $e_i = 0$ when $Y_i = \hat{Y}$ or $e_i = 1$ when $Y_i \neq \hat{Y}$. The typical error is simply the more frequent or modal error. Modal error is seldom used, but we have presented it for completeness.

Median absolute deviation

When we use the sum of absolute errors (SAE) as the aggregate index of error, it is customary to use the median absolute error or deviation from the prediction to represent the typical error. To find the median absolute deviation, simply sort the $|e_i|$ into ascending order and find the middle one.

Standard deviation

When we use the sum of squared errors (SSE) as the aggregate index of error, the index is somewhat more complex. We will be making extensive use of SSE throughout this book. To avoid having to introduce more general formulas later, we present the more general formula now and then show how it applies in the case of simple models. In a general model with p parameters, those p parameters have been used to reduce the error. In principle, the maximum number of parameters we could have is n—one parameter for each observation—in which case the error would equal zero. Thus, there are $n - p$ potential parameters remaining that could be used to reduce the remaining error. A useful index of error is then the remaining error per remaining potential parameter. This index has the name *mean squared error* (MSE) and is given by:

$$\text{MSE} = \frac{\text{SSE}}{n-p} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p}$$

For the simple model considered in this chapter there is only the one parameter β_0 to be estimated, so $p = 1$ and the estimate of Y_i is $b_0 = \bar{Y}$, the mean value. For the simple model, MSE has the special name *variance* and is commonly represented by s^2, that is:

$$\text{Variance} = s^2 = \text{MSE} = \frac{\text{SSE}}{n-1} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

MSE represents the typical *squared* error; to express the typical error in the units in which the original data were recorded, it is useful to take the square root of MSE, which is often referred to, especially on computer printouts, as *ROOT MSE*. For the simple model, the square root of the variance or the MSE has the special name *standard deviation* and is given by:

$$\text{Standard deviation} = s = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

Another index sometimes used when SSE is used as the aggregate index of error is the *coefficient of variation*. It is common for the size of the standard deviation to be proportional to the size of the mean. For example, if $\overline{Y} =$ 10,000, we would expect the typical error or standard deviation to be much larger than when $\overline{Y} = 10$. Although this need not be the case, it usually is true. To remove the effect of the overall magnitude of the data from the description of the error, the coefficient of variation expresses the size of the standard deviation as a proportion of the mean, that is:

$$\text{Coefficient of variation} = \text{CV} = \frac{s}{\overline{Y}}$$

An example

We will use the percentage of households that had internet access in the year 2013 by US state, which are listed in Figure 1.1, as an example to illustrate the simple model and the descriptors of central tendency and error. To facilitate finding the mode and the median, we have rearranged the data of Figure 1.1 in Figure 2.9 in order of increasing percentages. Our goal is to fit to these data the simple model that has just one parameter. Thus, the basic data analysis equation is:

$$Y_i = \beta_0 + \varepsilon_i$$

and we want to fit the model to the data by finding the estimate b_0 for β_0 that minimizes error—the e_i in the equation:

$$Y_i = b_0 + e_i$$

How we find the estimate b_0 depends on which definition of aggregate error we adopt.

If CE is adopted as the criterion, then the best estimate for β_0 is the mode. To find the mode, we simply observe which percentage is the most frequent. For these data, there are eight values that occur twice (i.e., 72.2, 72.9, 73.0, 75.3, 76.5, 77.5, 78.9, and 79.6) and none that occurs more than twice. Thus, there are really eight modes. It is frequently the case with continuous variables that there is no mode, or at least not a single mode, so the mode is usually not as useful as either the median or the mean for such data. If we were to round the data to the nearest whole number, there would be only two modes of 73 and 79. Using a mode of 73 (or 79) to predict the rounded data, the prediction would be accurate 6 times and incorrect 44 times.

If SAE is adopted as the criterion, then the best estimate for β_0 is the median. There are 50 observations, so there are two middle values—the 25th and 26th. (If there is an odd number of observations then there will be only one middle observation.) The 25th and 26th values are both 73.0; so, in this case, the best estimate is 73.0. (If the two middle observations were different, they could be averaged to produce the single estimate of the median.) The middle columns of Figure 2.10 present the prediction $\hat{Y}_i$, the error $e_i = Y_i - \hat{Y}_i$, and the absolute error based on the median as the estimate b_0. In this case, total error = 200.3. Any other estimate for β_0 would produce a larger value for SAE. Although it is not obvious from Figure 2.10, the 25th and 26th largest absolute errors are 3.20 and 3.30, so the median absolute deviation or MAD = 3.25.

If SSE is adopted as the criterion, then the best estimate for β_0 is the mean. The average of the 50 observations gives 72.806 as the estimate b_0. The last set of columns

FIGURE 2.9 Percentage of households that had internet access in 2013 by US state (sorted by percentage)

i	*State*	*Percentage*	*Rank*	*i*	*State*	*Percentage*	*Rank*
24	MS	57.4	1	49	WI	73.0	26
4	AR	60.9	2	12	ID	73.2	27
1	AL	63.5	3	3	AZ	73.9	28
31	NM	64.4	4	13	IL	74.0	29
18	LA	64.8	5	9	FL	74.3	30
48	WV	64.9	6	8	DE	74.5	31
40	SC	66.6	7	32	NY	75.3	32
36	OK	66.7	8	45	VT	75.3	33
42	TN	67.0	9	50	WY	75.5	34
17	KY	68.5	10	28	NV	75.6	35
14	IN	69.7	11	46	VA	75.8	36
25	MO	69.8	12	23	MN	76.5	37
22	MI	70.7	13	39	RI	76.5	38
33	NC	70.8	14	7	CT	77.5	39
41	SD	71.1	15	37	OR	77.5	40
35	OH	71.2	16	5	CA	77.9	41
43	TX	71.8	17	11	HI	78.6	42
26	MT	72.1	18	20	MD	78.9	43
10	GA	72.2	19	47	WA	78.9	44
15	IA	72.2	20	2	AK	79.0	45
38	PA	72.4	21	30	NJ	79.1	46
34	ND	72.5	22	6	CO	79.4	47
19	ME	72.9	23	21	MA	79.6	48
27	NE	72.9	24	44	UT	79.6	49
16	KS	73.0	25	29	NH	80.9	50

in Figure 2.10 gives the values of $\hat{Y}$ (or b in the case of the simple model), $e = Y - \hat{Y}$, and e_i^2. Note that the sum of the errors equals zero exactly and, necessarily, that the sum of the data observations equals the sum of the predictions. This is characteristic of predictions based on minimizing the SSE. The actual SSE equals 1355.028. Again, any other estimate for β_0 would produce a larger SSE. The variance or, more generally, the MSE equals:

$$s^2 = \frac{\text{SSE}}{n-1} = \frac{1355.028}{49} = 27.654$$

and the standard deviation or root-mean-squared error equals:

$$s = \sqrt{\text{MSE}} = \sqrt{27.654} = 5.259$$

Finally, CV is given by:

$$\frac{s}{\bar{Y}} = \frac{5.259}{72.806} = .072$$

Note that in this particular example the three estimates of β_0 (using the three definitions of error) were very similar: 73 (the first mode), 73.0, and 72.806. This is often the case for "well-behaved" data, but there is no guarantee that data will be well

FIGURE 2.10 Predictions and errors using the median and mean to estimate b_0 in the simple model

			Median			*Mean*		
i	*US state*	*Percentage*	$\hat{Y}_i$	e_i	$\lvert e_i \rvert$	$\hat{Y}_i$	e_i	e_i^2
1	AK	79.0	73	6.0	6.0	72.806	6.194	38.366
2	AL	63.5	73	–9.5	9.5	72.806	–9.306	86.602
3	AR	60.9	73	–12.1	12.1	72.806	–11.906	141.753
4	AZ	73.9	73	0.9	0.9	72.806	1.094	1.197
5	CA	77.9	73	4.9	4.9	72.806	5.094	25.949
6	CO	79.4	73	6.4	6.4	72.806	6.594	43.481
7	CT	77.5	73	4.5	4.5	72.806	4.694	22.034
8	DE	74.5	73	1.5	1.5	72.806	1.694	2.870
9	FL	74.3	73	1.3	1.3	72.806	1.494	2.232
10	GA	72.2	73	–0.8	0.8	72.806	–0.606	0.367
11	HI	78.6	73	5.6	5.6	72.806	5.794	33.570
12	IA	72.2	73	–0.8	0.8	72.806	–0.606	0.367
13	ID	73.2	73	0.2	0.2	72.806	0.394	0.155
14	IL	74.0	73	1.0	1.0	72.806	1.194	1.426
15	IN	69.7	73	–3.3	3.3	72.806	–3.106	9.647
16	KS	73.0	73	0.0	0.0	72.806	0.194	0.038
17	KY	68.5	73	–4.5	4.5	72.806	–4.306	18.542
18	LA	64.8	73	–8.2	8.2	72.806	–8.006	64.096
19	MA	79.6	73	6.6	6.6	72.806	6.794	46.158
20	MD	78.9	73	5.9	5.9	72.806	6.094	37.137
21	ME	72.9	73	–0.1	0.1	72.806	0.094	0.009
22	MI	70.7	73	–2.3	2.3	72.806	–2.106	4.435
23	MN	76.5	73	3.5	3.5	72.806	3.694	13.646
24	MO	69.8	73	–3.2	3.2	72.806	–3.006	9.036
25	MS	57.4	73	–15.6	15.6	72.806	–15.406	237.345
26	MT	72.1	73	–0.9	0.9	72.806	–0.706	0.498
27	NC	70.8	73	–2.2	2.2	72.806	–2.006	4.024
28	ND	72.5	73	–0.5	0.5	72.806	–0.306	0.094
29	NE	72.9	73	–0.1	0.1	72.806	0.094	0.009
30	NH	80.9	73	7.9	7.9	72.806	8.094	65.513
31	NJ	79.1	73	6.1	6.1	72.806	6.294	39.614
32	NM	64.4	73	–8.6	8.6	72.806	–8.406	70.661
33	NV	75.6	73	2.6	2.6	72.806	2.794	7.806
34	NY	75.3	73	2.3	2.3	72.806	2.494	6.220
35	OH	71.2	73	–1.8	1.8	72.806	–1.606	2.579
36	OK	66.7	73	–6.3	6.3	72.806	–6.106	37.283
37	OR	77.5	73	4.5	4.5	72.806	4.694	22.034
38	PA	72.4	73	–0.6	0.6	72.806	–0.406	0.165
39	RI	76.5	73	3.5	3.5	72.806	3.694	13.646
40	SC	66.6	73	–6.4	6.4	72.806	–6.206	38.514
41	SD	71.1	73	–1.9	1.9	72.806	–1.706	2.910
42	TN	67.0	73	–6.0	6.0	72.806	–5.806	33.710
43	TX	71.8	73	–1.2	1.2	72.806	–1.006	1.012
44	UT	79.6	73	6.6	6.6	72.806	6.794	46.158
45	VA	75.8	73	2.8	2.8	72.806	2.994	8.964
46	VT	75.3	73	2.3	2.3	72.806	2.494	6.220
47	WA	78.9	73	5.9	5.9	72.806	6.094	37.137
48	WI	73.0	73	0.0	0.0	72.806	0.194	0.038
49	WV	64.9	73	–8.1	8.1	72.806	–7.906	62.505
50	WY	75.5	73	2.5	2.5	72.806	2.694	7.258
Sum		3640.3	3650	–9.7	200.3	3640.3	.000	1355.03

behaved and that the three estimates will be similar. Later we will see that a major discrepancy between the three estimates, especially between the median and the mean, should alert us to special problems in the analysis of such data. Note also in this example that the median absolute deviation and the standard deviation produce different estimates for the typical error—3.25 and 5.26, respectively. This is not surprising given the different definitions of error used.

SUMMARY

In terms of the basic data analysis equation:

DATA = MODEL + ERROR

the simple model with one parameter is expressed as:

$$Y_i = \beta_0 + \varepsilon_i$$

Fitting the model to the data consists of finding the estimator b_0 of β_0 that makes the errors e_i as small as possible in the equation:

$$Y_i = b_0 + e_i$$

To fit the simple model to data, we must first define how the individual error terms e_i are to be aggregated into a summary index of error. Once we have chosen an aggregate index of error, we can find, by trial and error if necessary, the best estimate b_0 of β_0 that minimizes error. Important definitions of aggregate error are: (a) the count of errors, (b) the sum of absolute errors, and (c) the sum of squared errors. For each of these definitions, there is a different, well-defined best estimator for β_0 that can be found by calculation rather than by trial and error. These estimators are, respectively: (a) the mode—the most frequent value of Y_i, (b) the median—the middle value of all the Y_i, and (c) the mean—the arithmetic average of all the Y_i. Also, for each of these three definitions of aggregate error there is an expression for representing the "typical" value for e_i. These expressions are, respectively: (a) the modal error, (b) the median absolute deviation, and (c) the standard deviation. Collectively, these best estimators and these expressions for the typical error are known as *descriptive statistics* because they provide a first-cut description of a batch of data. We prefer to view them simply as estimators for the simple model using different definitions of error.

In Chapter 3, we will choose one of the three definitions of error to be our standard definition. We will make this choice on the basis of reasoned principles. However, the estimates and aggregate indices for the sum of squared errors are more easily obtained from the standard computer statistical systems than are those for other definitions of aggregate error. The reader should therefore anticipate that we will choose the sum of squared errors to be our standard definition of aggregate error.