Chapter 8: Replacing Missing and Empty Values

Overview

The presence of missing or empty values can make problems for the modeler. There are several ways of replacing these values, but the best are those that not only are well understood by the modeler as to their capabilities, limits, and dangers, but are also in the modeler's control. Even replacing the values at all has its dangers unless it is carefully done so as to cause the least damage to the data. It is every bit as important to avoid adding bias and distortion to the data as it is to make the information that is present available to the mining tool.

The data itself, considered as individual variables, is fairly well prepared for mining at this stage. This chapter discusses a way to fill the missing values, causing the least harm to the structure of the *data set* by placing the missing value in the context of the other values that are present. To find the necessary context for replacement, therefore, it is necessary to look at the data set as a whole.

8.1 Retaining Information about Missing Values

Missing and empty values were first mentioned in Chapter 2, and the difference between missing and empty was discussed there. Whether missing or empty, many, if not most, modeling tools have difficulty digesting such values. Some tools deal with missing and empty values by ignoring them; others, by using some metric to determine "suitable" replacements. As with normalization (discussed in the last chapter), if default automated replacement techniques are used, it is difficult for the modeler to know what the limitations or problems are, and what biases may be introduced. Does the modeler know the replacement method being used? If so, is it really suitable? Can it introduce distortion (bias) into the data? What are its limitations? Finding answers to these questions, and many similar ones, can be avoided if the modeler is able to substitute the missing values with replacements that are at least neutral, that is, introduce no bias—and using a method understood, and controlled by, the modeler.

Missing values should be replaced for several reasons. First, some modeling techniques cannot deal with missing values and cast out a whole instance value if one of the variable values is missing. Second, modeling tools that use default replacement methods may introduce distortion if the method is inappropriate. Third, the modeler should know, and be in control of, the characteristics of any replacement method. Fourth, most default replacement methods discard the information contained in the missing-value patterns.

8.1.1 Missing-Value Patterns

A point to note is that replacing missing values, without elsewhere capturing the information that they were missing, actually removes information from the data set. How is this? Replacing a missing value obscures the fact that it was missing. This information can be very significant. It has happened that the pattern of missing values turned out to be the most important piece of information during modeling. Capturing this information has already been mentioned in Chapter 4. In Figure 4.7, the single-variable CHAID analysis clearly shows a significant relationship between the missing-value pattern variable, _Q_MVP, and the variable SOURCE in the SHOE data (included on the CD-ROM). Retaining the information about the pattern in which missing values occur can be crucial to building a useful model.

In one particular instance, data had been assembled into a data warehouse. The architects had carefully prepared the data for warehousing, including replacing the missing values. The data so prepared produced a model of remarkably poor quality. The quality was only improved when the original source data was used and suitably prepared for modeling rather than warehousing. In this data set, the most predictive variable was in fact the missing-value pattern—information that had been completely removed from the data set during warehousing. The application required a predictive model. With warehoused data, the correlation between the prediction and the outcome was about 0.3. With prepared data, the correlation improved to better than 0.8.

Obviously a change in correlation from 0.3 to 0.8 is an improvement, but what does this mean for the accuracy of the model? The predictive model was required to produce a score. If the prediction was within 10% of the true value, it was counted as “correct.” The prediction with a correlation of 0.3 was “correct” about 4% of the time. It was better than random guessing, but not by much. Predicting with a correlation of about 0.8 produced “correct” estimates about 22% of the time. This amounts to an improvement of about 550% in the predictive accuracy of the model. Or, again, with a 0.3 correlation, the mean error of the prediction was about 0.7038 with a standard deviation of 0.3377. With a 0.8 correlation, the mean error of the prediction was about 0.1855 with a standard deviation of 0.0890. (All variables were normalized over a range of 0–1.)

Whatever metric is used to determine the quality of the model, using the information embedded in the missing-value patterns made a large and significant difference.

8.1.2 Capturing Patterns

The *missing-value pattern* (MVP) is exactly that—the pattern in which the variables are missing their values. For any instance of a variable, the variable can either have a value (from 0–1) or not have any value. If it has no numerical value, the value is “missing.” “Missing” is a value, although not a numerical value, that needs to be replaced with a numerical value. (It could be empty, but since both missing and empty values have to be

treated similarly and replaced, here they will be discussed as if they are the same.) For each variable in the data set, a flag, say, “P” for present and “E” for empty, can be used to indicate the presence or absence of a variable’s value in the instance value. Using such flags creates a series of patterns, each of which has as many flags as there are dimensions in the data. Thus a three-dimensional data set could have a maximum of eight possible MVPs as shown in Table 8.1

TABLE 8.1 Possible MVPs for a three-dimensional data set.

Pattern number	Pattern
1	PPP
2	PPE
3	PEP
4	PEE
5	EPP
6	EPE
7	EEP
8	EEE

The number of possible MVPs increases very rapidly with the number of dimensions. With only 100 dimensions, the maximum possible number of missing-value patterns is far more than any possible number of instance values in a data set. (There are over one nonillion, that is, 1 x 1030, possible different patterns.) The limiting number quickly becomes the maximum number of instances of data. In practice, there are invariably far fewer MVPs than there are instances of data, so only a minute fraction of the possible number of MVPs actually occur. While it is quite possible for all of the MVPs to be unique in most data sets, every practical case produces at least some repetitive patterns of missing values. This is particularly so for behavioral data sets. (Chapter 4 discussed the difference between physical and behavioral data sets.)

MVPs aren't invariably useful in a model. However, surprisingly often the MVPs do contribute useful and predictive information. The MVPs are alpha labels, so when they are extracted, the MVPs are numerated exactly as any other alpha value. Very frequently the MVPs are best expressed in more than one dimension. (Chapter 6 discusses numeration of alpha values.) Where this is the case, it is also not unusual to find that one of the multiple dimensions for MVPs is especially predictive for some particular output.

8.2 Replacing Missing Values

Once the information about the patterns of missing values is captured, the missing values themselves can be replaced with appropriate numeric values. But what are these values to be? There are several methods that can be used to estimate an appropriate value to plug in for one that is missing. Some methods promise to yield more information than others, but are computationally complex. Others are powerful under defined sets of circumstances, but may introduce bias under other circumstances.

Computational complexity is an issue. In practice, one of the most time-consuming parts of automated data preparation is replacing missing values. Missing-value estimating methods that produce mathematically optimal values can be highly complex, and vary with the type of data they are to be applied to. Highly complex methods are too time-consuming for large data sets. Even with the speed of modern computer systems, the amount of processing required simply takes too long to be reasonable, especially for time-sensitive business applications. Also, preparation techniques need to be as broadly applicable as possible. Replacement methods with advantages in specific situations, but that are unworkable or introduce bias in others, are simply not suitable for general application.

In addition to first doing as little harm as possible under all circumstances, and being computationally tractable, whatever method is used has to be applicable not only to the identified MVPs, but to any new missing values that arise in execution data sets. Chapter 7 discussed estimating the probability that the population maximum was found in any particular sample. That discussion made it clear that out-of-range values were to be expected during execution. The same is true for MVPs—and for the missing values that they represent. There is always some probability that individual variable values that were not ever missing in the sample will be found missing in the execution data set, or that MVPs occurring in the population are not actually present in a sample. The PIE-I needs to be able to deal with these values too—even though they were never encountered as missing during the building of the PIE.

8.2.1 Unbiased Estimators

An *estimator* is a device used to make a justifiable guess about the value of some particular value, that is, to produce an *estimate*. An *unbiased* estimator is a method of guessing that doesn't change important characteristics of the values present when the

estimates are included with the existing values. (Statistically, an unbiased estimator produces an estimate whose “expected” value is the value that would be estimated from the population.)

For instance, consider the numbers 1, 2, 3, *x*, 5, where “*x*” represents a missing value. What number should be plugged in as an unbiased estimate of the missing value? Ideally, a value is needed that will at least do no harm to the existing data. And here is a critical point—what does “least harm” mean exactly? If the mean is to be unbiased, the missing value needs to be 2.75. If the standard deviation is to be unbiased, the missing value needs to be about 4.659. The missing-value estimate depends as much on which characteristic is to be unbiased as it does on the actual values. Before deciding what “least harm” means in practice, it is important to discover which relationships need to be preserved, both within and between variables. Finding which are the important relationships to preserve will indicate how to find the estimate that best preserves these relationships—that is, that is least biased for these particular relationships.

8.2.2 Variability Relationships

Chapter 7 discussed variability and redistributing a variable’s distribution. The transform produced a rectangular distribution insofar as the actual values allowed for it. Each variable was considered individually, and was redistributed individually without reference to the distributions of other variables. In addition to this within-variable relationship, there is also a critical between-variable relationship that exists for all of the variables. The between-variable relationship expresses the way that one variable changes its value when another variable changes in value. It is this multiple-way, between-variable relationship that will be explored by any modeling tool. Since the chosen modeling tool is going to explore these relationships between variables, it is critical to preserve them, so far as possible, when replacing missing values.

Variability forms a key concept in deciding what values to use for the replacement. Standard deviation is one measure of that variability (introduced in Chapter 5). To exemplify the underlying principles of preserving variability, consider a single variable whose values are transformed into a rectangular distribution. Figure 8.1 shows the values of such a variable. The 11 values of the original series are shown in the column headed “Original sample.” Suppose that the value in series position 11 is missing and is to be replaced. Since in this example the actual series value is present, it is easy to see how well any chosen estimator preserves the relationships.

Position	Original sample	Position 11 missing	Preserve mean as estimate	Preserve variance as estimate
1	0.0886	0.0886	0.0886	0.0886
2	0.0684	0.0684	0.0684	0.0684
3	0.3515	0.3515	0.3515	0.3515
4	0.9874	0.9874	0.9874	0.9874
5	0.4713	0.4713	0.4713	0.4713
6	0.6115	0.6115	0.6115	0.6115
7	0.2573	0.2573	0.2573	0.2573
8	0.2914	0.2914	0.2914	0.2914
9	0.1662	0.1662	0.1662	0.1662
10	0.4400	0.4400	0.4400	0.4400
11	0.6939	?	**0.3731**	**0.6622**
Mean	0.4023	0.3731	0.3731	0.3994
Standard deviation	0.2785	0.2753	0.2612	0.2753
Size of error in the estimate			0.3208	0.0317

Figure 8.1 Estimating the value of position 11, given only the values in positions 1 through 10.

Position 11 has an actual value of 0.6939. The mean for the original 11-member series is 0.4023. If the value for position 11 is to be estimated, only positions 1 through 10 are used since the actual value of position 11 is assumed unknown. The third column (headed "Position 11 missing") shows the 10 values used to make the estimate, with the mean and standard deviation for these first 10 positions shown beneath. The mean for these first 10 positions is 0.3731. Using this as the estimator and plugging it into position as an estimate of the missing value (position 11) changes the standard deviation from about 0.2753 to 0.2612. The column mean is unchanged (0.3731). Using the mean of instances 1 through 10 has least disturbed the *mean* of the series. The actual value of the "missing" value in position 11 is 0.6939, and the mean has estimated it as 0.3731—a discrepancy of 0.3208.

Suppose, however, that instead of using the mean value, a value is found that least disturbs the standard deviation. Is this a more accurate estimate, or less accurate? Position 11 in column five uses an estimate that least disturbs the standard deviation. Comparing the values in position 11, column four (value not disturbing mean as replacement) and column five (value not disturbing standard deviation as replacement), which works best at estimating the original value in column two?

The column four estimator (preserving mean) misses the mark by

$$0.6939 - 0.3731 = 0.3208.$$

The column five estimator (preserving standard deviation) only misses the mark by

$$0.6622 - 0.6939 = 0.0317.$$

Also, preserving the standard deviation (column five) moved the new mean closer to the original mean value for all 11 original values in column one. This can be seen by

comparing the mean and standard deviation for each of the columns. The conclusion is that preserving standard deviation does a much better job of estimating the “true” mean and provides a less biased estimate of the “missing” value.

Was this a convenient and coincidental fluke? Figure 8.2 shows the situation if position 1 is assumed empty. The previous example is duplicated with values for preserving mean and standard deviation shown in separate columns. As before, generating the replacement value by preserving standard deviation produces a less biased estimate.

Position	Original sample	Position 1 missing	Preserve mean as estimate	Preserve variance as estimate
1	0.0886	?	**0.4336**	**0.1479**
2	0.0684	0.0684	0.0684	0.0684
3	0.3515	0.3515	0.3515	0.3515
4	0.9874	0.9874	0.9874	0.9874
5	0.4713	0.4713	0.4713	0.4713
6	0.6115	0.6115	0.6115	0.6115
7	0.2573	0.2573	0.2573	0.2573
8	0.2914	0.2914	0.2914	0.2914
9	0.1662	0.1662	0.1662	0.1662
10	0.4400	0.4400	0.4400	0.4400
11	0.6939	0.6939	0.6939	0.6939
Mean	0.4025	0.4336	0.4336	0.4076
Standard deviation	0.2791	0.2723	0.2584	0.2723
Size of error in the estimate			0.3450	0.0593

Figure 8.2 Estimating the value of position 1, given only the values in positions 2 through 11.

Why is it that preserving variability as measured by standard deviation produces a better estimate, not only of the missing value, but of the original sample mean too? The reason is that the standard deviation reflects far more information about a variable than the mean alone does. The mean is simply a measure of central tendency (mentioned in Chapter 7). The standard deviation reflects not just central tendency, but also information about the variability within the variable’s distribution. If the distribution is known (here made as rectangular as possible by the methods described in Chapter 7), that knowledge contributes to determining a suitable replacement value. It is this use of additional information that produces the better estimate.

Preserving variability works well for single missing values in a variable. If multiple values are missing, however, the estimator still produces a single estimate for all missing values, just as using the mean does. If both positions 1 and 11 were missing in the above example, any estimator preserving standard deviation would produce a single estimate to replace both missing values. So, there is only one single replacement value to plug into all the missing values. Does this cause any problem?

8.2.3 Relationships between Variables

Since it is important to retain the relationship between the variables as well as possible, the question becomes, Does assigning a single value to a variables' missing values maintain the between-variable relationship?

As a system of variables, there exists some relationship linking the variables' values to each other. It may be stronger or weaker depending on which variables are compared and which portions of the range are compared. Nonetheless, since there is a relationship, whatever it may be, if the value of one variable is found to be at a specific value, the values of all the other variables will be expected to be at particular values in their ranges. The linkage expresses the amount and direction of change in value of each variable. This amounts to no more, for instance, than saying, "Size of home owned increases with income level." True or not, this statement expresses a relationship between house size and income level. As one changes, so do expected values of the other.

If missing values of, say, income, are replaced with any fixed values, it does not allow for, or reflect, the value of house size associated with that missing value. No notice is taken of what might be an "appropriate" value to use for the missing value of income, given the matching value of house size. Assigning a constant value to one variable for all of its missing values will certainly distort the relationship between the variables. What is more, as already observed, if the missing values are not missing at random, then using replacements that all have the same value will not only be inappropriate, but will actually add bias, or distortion. This will show up as the nonrandom pattern of fixed values plugged into the missing values.

Replacing missing values in one variable, then, needs to take account of whatever values are actually present for the other variables in a specific instance value. Given multiple MVPs, several variables may be simultaneously missing values, but never all of the variables at once. Whatever variable values are present can be used to estimate what the appropriate level of the missing variable values should be.

There are several ways of estimating the appropriate missing values, and it is here that the demonstration software takes a shortcut to make the computations tractable. At this point in the data preparation, the modeler does not know the precise nature of the relationship between the variables. Discovering and explicating the nature of that relationship is, in fact, one of the main tasks of data mining in the first place—specifically the part called modeling that comes after preparation. It is likely that the exact relationship is not actually linear. In a linear relationship, if the value of one variable changes by a particular amount, then the value of another variable changes by another particular amount, and in a specific direction. To return to house size and income for a moment, it means that however much the house size increases for a $1,000 rise in income at $25,000 a year, it increases a similar amount for any similar $1,000 rise at any other income level, say, $50,000. This may or may not be the case. For the demonstration software, however, the relationship is assumed to be linear. In practice, this assumption introduces very little if any bias for most data sets. Nonetheless, the modeler needs to be

aware of this.

A key point is that, although the replacement values are indeed predictions, it is not the accuracy of these predictions that is of most importance when replacing missing values. The key concern is that the predictions produce a workable estimate that least distorts the values that are actually present. It is this “least distortion” that is the important point. The purpose of replacing missing values is not to use the values themselves, but to make available to the modeling tools the information contained in the other variables’ values that are present. If the missing values are not replaced, the whole instance value may be ignored. If not ignored, the missing value may be replaced with some default unknown to the modeler that introduces bias and distortion. Avoiding bias is important. Although not perfect, multiple linear estimation is enormously preferable to, and produces much less bias than, any replacement with a constant value (like the mean value), or a procedure that ignores problematic instances.

However, it must be noted that a more accurate replacement can be achieved by using one of the more computationally intensive methods of determining multiple nonlinear relationships between variables. Although the demonstration software uses a modified multiple linear regression technique, which is examined next, a brief review of some possible alternatives in discussed in the Supplemental Material section at the end of this chapter. (The full data preparation and survey suite, on which the demonstration code is based, uses a nonlinear estimator optimized for least information change. This method is based on the principles discussed here. Where intervariable relationships are nonlinear or even discontinuous, such methods minimize the potential distortion to information content in a data set.)

8.2.4 Preserving Between-Variable Relationships

Regression methods are inherently mathematical and normally are themselves very sensitive to missing values. The Supplemental Material section at the end of this chapter discusses exactly how regression methods can be modified to determine appropriate missing values.

As a conceptual overview, recall that what is needed is some way of determining how the value of one variable changes as the value of another (or several others) changes. This simply means measuring what values of one variable are when another variable has particular values.

Table 8.2 shows, for instance, that if the value of *y* was missing in a particular instance, but the value of *x* was 3, then an appropriate value to plug in for *y* would be about 14. But what is needed for automated replacement is not a table to look at to find a replacement value, but a formula that relates one set of values to the other. What would this formula look like?

TABLE 8.2 Values of two variables.

Variable *x* value	Variable *y* value
1	10
2	12
3	14
4	16
5	18
6	20

Start by trying to find a way to predict *y* from *x*. Notice that the value of *y* increases by 2 for every increase in *x* of 1, and decreases by 2 for every decrease in *x* of 1. So notice that if *x* = 0, *y* must = 8. So the relationship must be that *y* equals 8 plus 2 times the amount of *x*. Or

$$y = 8 + 2x$$

A little mathematical manipulation of this formula gives an expression for determining values of *x* from values of *y*:

$$x = (y - 8)/2$$

These formulas represent the essence of what needs to be done to find missing values. Of course, something other than an intuitive mechanism is needed for discovering the expressions needed, and more of those details are explored in the Supplemental Material section at the end of this chapter. However, the whole method is based on preserving between-variable variability, which is no more than saying preserving the between-variable relationships, which is what the example just did.

This is a very simple example of what regression analysis achieves. How it is done in practice, why preserving variability is critical, and how the methods work are all covered in more conceptual detail also in the Supplemental Material section at the end of this

chapter.

8.3 Summary

Replacing missing values in a data set is very important. However, it is at least as important to capture the patterns of values that are missing, and the information that they contain, as it is to replace the values themselves. The values that are missing must be replaced with extreme sensitivity for not disturbing the patterns in the data that already exist. Using inappropriate values easily disturbs those patterns and introduces spurious patterns, called bias or noise. Such artificially introduced patterns damage the information carried in the between-variable variability in a data set. This hides or distorts the existing patterns, thus hiding or distorting the information embedded in the data set.

Capturing the variability that is present in a data set in the form of the ratios between various values can be used to infer, or impute, appropriate missing values that do least damage to the information content of the data set. All methods of replacing missing values are a compromise. Please see the Supplemental Material section at the end of this chapter for a more detailed explanation.

From here forward, the data (variables) are prepared for modeling. What remains is to look at the issues that are still outstanding in preparing the data set as a whole.

Supplemental Material

Using Regression to Find Least Information-Damaging Missing Values

In any data set, with or without missing values, there is enfolded information. The information is carried in the relationships between the values both within a single variable (its distribution), and in the relationships with the patterns in other variables. If some of these values are missing, for reasons already discussed, they need to be replaced. But replacing values is a tricky business. It is critically important that the replacement values do not introduce a pattern into the data that is not actually present in the measured values. Such a pattern may later be “discovered” by the miner, and may actually appear to be meaningful, but it carries no real information since it is simply an artifact of the replacement process.

However, just as important as not introducing an artificial pattern into a data set by replacement values is the maintenance of the existing pattern. Some patterns, maybe critically important ones, are fragile and delicate, and great care must be taken to maintain them undistorted in the data set. How is this delicate balancing act to be accomplished? There are a number of techniques available. The one explained here is used primarily for ease of understanding. There are more complex approaches available that perform slightly better under some circumstances, and those are discussed in the second section

of this supplemental material.

Regressions

Multiple linear regression is a generalized extension of ordinary linear regression, which is more accessible for explanatory purposes than multiple linear regression since it uses only two variables. The idea is simply to determine the value of one variable given the value of the other. It assumes that the variables' values change, the one with the other, in some linear way. The technique involved simply fits a manifold, in the form of a straight line, through the two-dimensional state space formed by the two variables. Figure 8.3 shows this for three of the variables in the CARS data set.

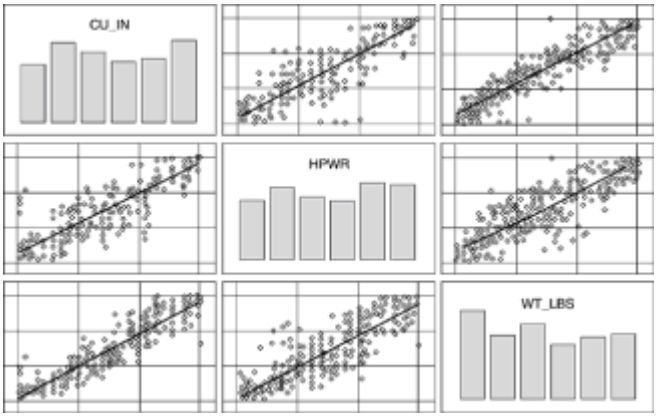

Figure 8.3 Linear regression manifolds (lines) for 2D spaces. The bar charts show the distributions of the variables CU_IN (cubic inches), HPWR (horsepower), and WT_LBS (weight).

Figure 8.3 shows the state space for each pair of variables. The bar graphs show the distribution for each of the three variables. Since they have been normalized and redistributed, the distributions are approximately rectangular. The lines drawn through each graph show the linear regression line that best fits each. For two-dimensional state spaces, the linear regression manifold is a line. In three dimensions it is a plane, and just such a plane is shown for all three variables simultaneously in Figure 8.4. In more than three dimensions, the manifold is impossible to show but still forms a rigid hyperdimensional surface that fits the points in state space.

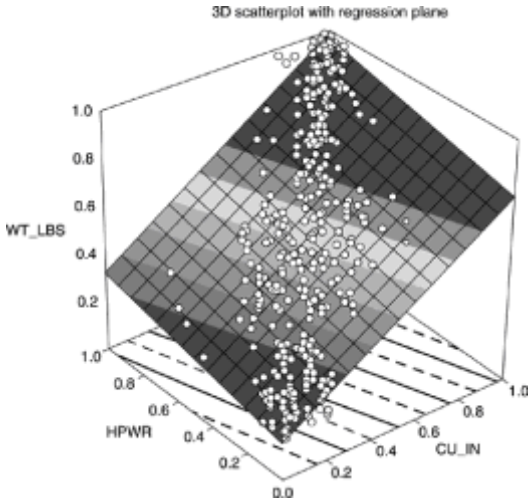

Figure 8.4 Multiple linear regression manifold in a 3D state space. The manifold is a flat plane. The flatness is characteristic of linear regressions in any number of dimensions. The variables used are the same as shown in Figure 8.3.

The linear regression technique involves discovering the *joint* variability of the two variables and using this to determine which values of the predicted variable match values of the predictor variable. *Joint variability*, the measure of the way one variable varies as another varies, allows between-variable inferences to be made of the same sort that were made within-variable in the example above. The shortcoming in the example shown in Figures 8.1 and 8.2 is that when multiple values are missing in a single variable, only a single replacement for all the missing values can be estimated. Preserving between-variable variability allows a suitable value to be found for a missing value in one variable if the value of its partner in the other variable is known. In a sense, linear regression involves finding the joint variability between two variables and preserving it for any needed replacement values. Multiple linear regression does the same thing, but weighs the contributions of several variables' joint distributions to estimate any missing value. With more variables contributing to the joint variability (more evidence if you will), it is usual to find a better estimate of any missing value than can be found by using only one other variable.

Linear Regression

Linear regression is relatively straightforward. It involves no more than discovering a specific expression for the straight line that best fits the data points in state space. The expression describing a straight line is

$$y = a + bx$$

This expression gives the appropriate y value for any value of x. The expression a is a constant that indicates where the straight line crosses the y-axis in state space, and b is

an expression that indicates how much the line increases (or decreases if it is a negative number) its position for an increase in *x*. In other words, to find a value for *y*, start at *a* and go up (or down) by amount *b* for every unit of *x*. Figure 8.5 illustrates the expression for a straight line. Linear regression involves finding the appropriate *a* and *b* values to use in the expression for a straight line that make it best fit through the points in state space. The linear regression comes in two parts, one to find *b* and the other to find *a* when *b* is known. These two expressions look like this:

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a = \overline{y} - b\overline{x}$$

where

$\overline{y}$ is the mean value of y
$\overline{x}$ is the mean value of x

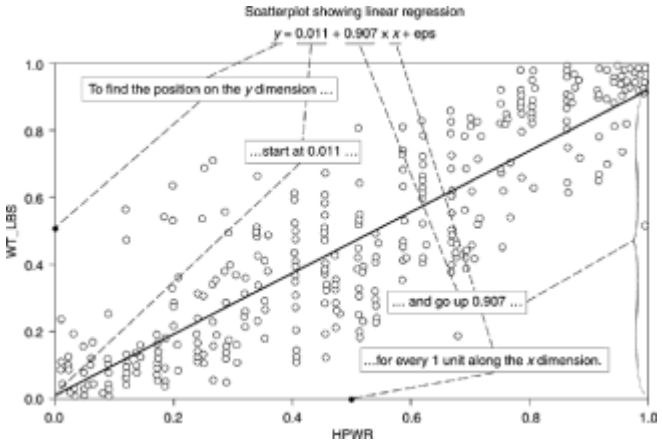

Figure 8.5 Showing the straight-line equation found that best fits the point distribution for the two variables WT_LBS and HPWR from the CARS data set.

Looking at these two expressions shows that the pieces of information required are

n	the number of instance values
Σ*x*	the sum of all the *x* values
Σ*y*	the sum of all the *y* values
Σx^2	the sum of all the values of *x* squared
Σ*xy*	the sum of all *x* values multiplied by all *y* values

Calculating these quantities is simple only if all of the values are actually present. Table 8.3 illustrates the problem.

TABLE 8.3 The effect of missing values (?.??) on the summary values of *x* and *y*.

n	*x*	*y*	*x*2	*y*2	*xy*
1	0.55	0.53	0.30	0.28	0.29
2	0.75	0.37	0.56	0.14	0.28
3	0.32	0.83	0.10	0.69	0.27
4	0.21	0.86	0.04	0.74	0.18
5	0.43	0.54	0.18	0.29	0.23
Sum	**2.26**	**3.13**	**1.20**	**2.14**	**1.25**
1	0.55	0.53	0.30	0.28	0.29
2	*?.??*	0.37	*?.??*	0.14	*?.??*
3	0.32	0.83	0.10	0.69	0.27
4	0.21	*?.??*	0.04	*?.??*	*?.??*
5	0.43	0.54	0.18	0.29	0.23
Sum	**?.??**	**?.??**	**?.??**	**?.??**	**?.??**

The problem is what to do if values are missing when the complete totals for all the values are needed. Regressions simply do not work with any of the totals missing. Yet if any single number is missing, it is impossible to determine the necessary totals. Even a single missing *x* value destroys the ability to know the sums for *x*, x^2, and *xy*! What to do?

Since getting the aggregated values correct is critical, the modeler requires some method to determine the appropriate values, even with missing values. This sounds a bit like pulling one's self up by one's bootstraps! Estimate the missing values to estimate the missing values! However, things are not quite so difficult.

In a representative sample, for any particular joint distribution, the *ratios* between the various values x*x* and xx^2, and x*y* and xy^2 remain constant. So too do the ratios between x*x* and x*xy* and x*y* and x*xy*. When these ratios are found, they are the equivalent of setting the value of *n* to 1. One way to see why this is so is because in any representative sample the ratios are constant, regardless of the number of instance values—and that includes *n* = 1. More mathematically, the effect of the number of instances cancels out. The end result is that when using ratios, *n* can be set to unity. In the linear regression formulae, values are multiplied by *n*, and multiplying a value by 1 leaves the original value unchanged. When multiplying by *n* = 1, the *n* can be left out of the expression. In the calculations that follow, that piece is dropped since it has no effect on the result.

The key to building the regression equations lies in discovering the needed ratios for those values that are jointly present. Given the present and missing values that are shown in the lower part of Table 8.3, what are the ratios?

Table 8.4 shows the ratios determined from the three instance values where *x* and *y* are both present. Using the expressions for linear regression and these ratios, what is the estimated value for the missing *y* value from Table 8.3?

TABLE 8.4 Ratios of the values that are present in the lower part of Table 8.3.

	xx2	***xy2***	***xxy***
Ratio x*x* to:	0.45		0.61
Ratio x*y* to:		0.66	0.42

In addition to the ratios, the sums of the *x* and *y* values that are present need to be found. But since the ratios scale to using an *n* of 1, so too must the sums of *x* and *y*—which is identical to using their mean values. The mean values of variable *x* and of variable *y* are taken for the values of each that are jointly present as shown in Table 8.5.

TABLE 8.5 Mean values of *x* and *y* for estimating missing values.

n	***x***	***y***