1 Introduction to Data Analysis

This book is about data analysis. In the social and behavioral sciences we often collect batches of data that we hope will answer questions, test hypotheses, or disprove theories. To do so we must analyze our data. In this chapter, we present an overview of what data analysis means. This overview is intentionally abstract with few details so that the "big picture" will emerge. Data analysis is remarkably simple when viewed from this perspective, and understanding the big picture will make it much easier to comprehend the details that come later.

OVERVIEW OF DATA ANALYSIS

The process of data analysis is represented by the following simple equation:

DATA = MODEL + ERROR

DATA represents the basic scores or observations, usually but not always numerical, that we want to analyze. MODEL is a more compact description or representation of the data. Our data are usually bulky and of a form that is hard to communicate to others. The compact description provided by the model is much easier to communicate, say, in a journal article, and is much easier to think about when trying to understand phenomena, to build theories, and to make predictions. To be a representation of the data, all the models we consider will make a specific prediction for each observation or element in DATA. Models range from the simple (making the same prediction for every observation in DATA) to the complex (making differential predictions conditional on other known attributes of each observation). To be less abstract, let us consider an example. Suppose our data were, for each state in the United States, the percentage of households that had internet access in the year 2013; these data are listed in Figure 1.1. A simple model would predict the same percentage for each state. A more complex model might adjust the prediction for each state according to the age, educational level, and income of the state's population, as well as whether the population is primarily urban or rural. The amount by which we adjust the prediction for a particular attribute (e.g., educational level) is an unknown *parameter* that must be estimated from the data.

The last part of our basic equation is ERROR, which is simply the amount by which the model fails to represent the data accurately. It is an index of the degree to which the model mispredicts the data observations. We often refer to error as the *residual*—the part that is left over after we have used the model to predict or describe the data. In other words:

ERROR = DATA – MODEL

FIGURE 1.1 Percentage of households that had internet access in the year 2013 by US state

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

The goal of data analysis is then clear: We want to build the model to be a good representation of the data by making the error as small as possible. In the unlikely extreme case when ERROR = 0, DATA would be perfectly represented by MODEL.

How do we reduce the error and improve our models? One way is to improve the quality of the data so that the original observations contain less error. This involves better research designs, better data collection procedures, more reliable instruments, etc. We do not say much about such issues in this book, but instead leave those problems to texts and courses in experimental design and research methods. Those problems tend to be much more discipline specific than the general problems of data analysis and so are best left to the separate disciplines. Excellent sources that cover such issues are Campbell and Stanley (1963), Cook and Campbell (1979), Judd and Kenny (1981a), Maruyama and Ryan (2014), Reis and Judd (2014), Rosenthal and Rosnow (2008), and Shadish, Cook, and Campbell (2002). Although we often note some implications of data analysis procedures for the wise design of research, we in general assume that the data analyst is confronted with the problem of building the best model for data that have already been collected.

The method available to the data analyst for reducing error and improving models is straightforward and, in the abstract, the same across disciplines. Error can almost always be reduced (never increased) by making the model's predictions conditional on additional information about each observation. This is equivalent to adding parameters to the model and using data to build the best estimates of those parameters. The meaning of "best

estimate" is clear: we want to set the parameters of the model to whatever values will make the error the smallest. The estimation of parameters is sometimes referred to as "fitting" the model to the data. Our ideal data analyst has a limited variety of basic models. It is unlikely that any of these models will provide a good fit "off the rack"; instead, the basic model will need to be fitted or tailored to the particular size and bulges of a given data customer. In this chapter, we are purposely vague about how the error is actually measured and about how parameters are actually estimated to make the error as small as possible because that would get us into details to which we devote whole chapters later. But for now the process in the abstract ought to be clear: add parameters to the model and estimate those parameters so that the model will provide a good fit to the data by making the error as small as possible.

To be a bit less abstract, let us again consider the example of internet access by state. An extremely simple model would be to predict a priori (that is, without first examining the data) that in each state the percentage of households that has internet access is 75. This qualifies as a model according to our definition, because it makes a prediction for each of the 50 states. But in this model there are no parameters to be estimated from the data to provide a good fit by making the error as small as possible. No matter what the data, our model predicts 75. We will introduce some notation so that we have a standard way of talking about the particulars of DATA, MODEL, and ERROR. Let Y_i represent the *i*th observation in the data; in this example Y_i is simply the percentage of households that have internet access for the *i*th state. Then our basic equation:

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

for this extremely simple model becomes:

$$Y_i = 75 + \text{ERROR}$$

We can undoubtedly improve our model and reduce the error by using a model that is still simple but has one parameter: predict that the percentage is the same in all states, but leave the predicted value as an unspecified parameter to be estimated from the data. For example, the average of all 50 percentages might provide a suitable estimate. We will let β_0 represent the unknown value that is to be estimated so that our slightly more complex, but still simple, model becomes:

$$Y_i = \beta_0 + \text{ERROR}$$

It is important to realize that we can never know β_0 for certain; we can only estimate it.

We can make our model yet more complex and reduce the error further by adding more parameters to make *conditional predictions*. For example, innovations reputedly are adopted on the east and west coasts before the middle of the country. We could implement that in a model that starts with a basic percentage of internet use (β_0) for all states, which is adjusted upward by a certain amount (β_1) if the state is in the Eastern or Pacific time zones and reduced by that same amount if the state is in the Central or Mountain time zones. More formally, our basic equation now has a more complex representation, namely:

$$Y_i = \beta_0 + \beta_1 + \text{ERROR if the state is in the Eastern or Pacific time zones}$$

$$Y_i = \beta_0 - \beta_1 + \text{ERROR if the state is in the Central or Mountain time zones}$$

In other words, our model and its prediction would be conditional on the time zone in which the state is located.

Another slightly more complex model would make predictions conditional on a continuous, rather than a categorical, predictor. For example, we might make predictions conditional on the proportion of college graduates in a state, presuming that college graduates are more likely to be internet users. We again start with a basic percentage of internet users (β_0) for all states, which is adjusted upward by a certain amount (β_1) for each percentage point a state's proportion of college graduates is above the national average and reduced by the same amount for each percentage point a state's proportion of college graduates is below the national average. More formally, letting X_i represent the amount a state's proportion of college graduates is above or below the national average:

$$Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$$

In words, the percentage of college graduates is the condition in this model on which we base our differential or conditional prediction of a state's internet use.

We can continue making our model yet more complex by adding parameters to make similar adjustments for income, urban versus rural population, etc. By so doing we will be adding still more implicit hypotheses to the model.

It might appear that the best strategy for the data analyst would be to add as many parameters as possible, but this is not the case. The number of observations in DATA imposes an inherent limit on the number of parameters that may be added to MODEL. At the extreme, we could have separate parameters in our model for each observation and then estimate the value of each such parameter to be identical to the value of its corresponding DATA observation. For example, our prediction might contain statements such as, *if the state is Kentucky, then estimate its parameter to be 68.5*, which is the percentage of households in Kentucky that have internet access. That procedure would clearly reduce the error to zero and provide a perfect fit. But such a model would be uninteresting because it would simply be a duplicate of data and would provide no new insights, no bases for testing our theories, and no ability to make predictions in slightly different circumstances. A paramount goal of science is to provide simple, parsimonious explanations for phenomena. A model with a separate parameter for each observation is certainly not parsimonious. Our ideal model, then, is a compact description of the data and has many fewer parameters than the number of observations in data.

We now have an obvious conflict. The goal of reducing the error and providing the best description of DATA leads us to add parameters to the model. On the other hand, the goal of parsimony and the desire for a compact, simple model lead us to remove parameters from the model. The job of the data analyst is to find the proper balance between these two conflicting objectives. Thus, the ultimate goal is to find the smallest, simplest model that provides an adequate description of the data so that the error is not too large ("too large" will be defined later). In still other words, the data analyst must answer the question of whether it is worthwhile to add yet more parameters to a model.

Returning to the example of internet access, we will want to ask whether the extra complexity of making predictions conditional on time zone, educational level, income, urban versus rural population, etc. is worth the trouble. By so doing, we will simultaneously be asking whether the hypotheses implicit in the more complex models are true. For example, if we decide that conditioning our prediction of internet access on

the percentage of college graduates is not worthwhile, then we will have effectively rejected the hypothesis that college education is related to higher internet access. This is the essence of testing hypotheses.

Although we are still being vague about how to measure the error, we can be more precise about what we mean by: "Are more parameters worthwhile?" We will call the model without the additional parameters the *compact model* and will refer to it as Model C. The alternative, *augmented model*, Model A, includes all the parameters, if any, of Model C plus some additional parameters. The additional parameters of Model A may reduce the error or leave it unchanged; there is no way the additional parameters can increase the error. So it must be that:

$$\text{ERROR(A)} \leq \text{ERROR(C)}$$

where ERROR(A) and ERROR(C) are the amounts of error when using Models A and C, respectively. The question of whether it is worthwhile to add the extra complexity of Model A now reduces to the question of whether the difference between ERROR(C) and ERROR(A) is big enough to worry about. It is difficult to decide based on the absolute magnitude of the errors. We will therefore usually make relative comparisons. One way to do that is to calculate the *proportional reduction in error* (PRE), which represents the proportion of Model C's error that is reduced or eliminated when we replace it with the more complex Model A. Formally:

$$\text{PRE} = \frac{\text{ERROR(C)} - \text{ERROR(A)}}{\text{ERROR(C)}}$$

The numerator is simply the difference between the two errors (the amount of error reduced) and the denominator is the amount of error for the compact model with which we started. An equivalent expression is:

$$\text{PRE} = 1 - \frac{\text{ERROR(A)}}{\text{ERROR(C)}}$$

If the additional parameters do no good, then ERROR(A) will equal ERROR(C), so PRE = 0. If Model A provides a perfect fit, then ERROR(A) = 0 and (assuming Model C does not also provide a perfect fit) PRE = 1. Clearly, values of PRE will be between 0 and 1. The larger the value of PRE, the more it will be worth the cost of increased complexity to add the extra parameters to the model. The smaller the value of PRE, the more we will want to stick with the simpler, more parsimonious compact model.

For example, ignoring for the moment how we calculate the error, assume that total ERROR = 50 for the simple model that says that internet access is the same in all states and that ERROR = 30 for the model with the additional parameter for the percentage of college graduates. Then, ERROR(C) = 50, ERROR(A) = 30, and:

$$\text{PRE} = 1 - \frac{30}{50} = .40$$

That is, increasing the complexity of the model by considering educational level would reduce the error by 40%.

Let us review where we are. We have transformed the original problem of the conflicting goals for the model (parsimony and accurate representation of data) into a

consideration of the size of PRE for comparing Model C and Model A. Unfortunately, we have still not solved the problem of the conflicting goals for the model, because now we must decide whether a given PRE (e.g., the 40% reduction in the previous example) is big enough to warrant the additional parameter(s). The transformation of the original problem has moved us closer to a solution, however, for now we have a PRE index that will be used no matter how we finally decide to measure the error. More importantly, PRE has a simple, intuitive meaning that provides a useful description of the amount of improvement provided by Model A over Model C.

Deciding whether a PRE of, say, 40% is really worthwhile involves inferential statistics. An understanding of inferential statistics must await the details of measuring the error, sampling distributions, and other topics that are developed in Chapters 2, 3, and 4. We can, however, now specify two considerations that will be important in inferential statistics. First, we would be much more impressed with a PRE of, say, 40% if it were obtained with the addition of only one parameter instead of with four or five parameters. Hence, our inferential statistics will need to consider the number of extra parameters added to Model C to create Model A. *PRE per parameter added* will be a useful index. Second, we noted that n, the number of observations in DATA, serves as an upper limit to the number of parameters that could be added to the model. We will be more impressed with a given PRE as the difference between the number of parameters that were added and the number of parameters that could have been added becomes greater. Hence, our inferential statistics will consider how many parameters *could have been added* to Model C to create Model A but were not. In other words, we will be more impressed with a PRE of 40% if the number of observations greatly exceeds the number of parameters used in Model A than if the number of observations is only slightly larger than the number of parameters.

The use of PRE to compare compact and augmented models is the key to asking questions of our data. For each question we want to ask of DATA, we will find appropriate Models C and A and compare them by using PRE. For example, if we want to know whether educational level is useful for predicting the percentage of households that have internet access, we would compare a Model C that does not include a parameter for educational level to a Model A that includes all the parameters of Model C plus an additional parameter for educational level. If Model C is a simple model (i.e., a single-parameter model that makes a constant prediction for all observations), then we are asking whether educational level *by itself* is a useful predictor of internet access. If there are other parameters in Model C, then we are asking whether educational level is a useful predictor of internet access *over and above* the other parameters. (We discuss this at length in Chapter 6.) As another example, if we want to ask whether several factors, such as time zone, educational level, urban versus rural population, and income, are simultaneously useful in predicting internet access, we would use PRE to compare a Model C that did not have parameters for any of those factors to a Model A that did have those parameters in addition to those in Model C.

In the usual language for statistical inference, Model C corresponds to the *null hypothesis* and Model A corresponds to the alternative hypothesis. More precisely, the null hypothesis is that all the parameters included in Model A but not in Model C are zero (hence, the name "null") or equivalently that there is no difference in error between Models A and C. If we reject Model C in favor of Model A, then we reject the null hypothesis in favor of the alternative hypothesis that is implied by the difference between

Models C and A. That is, we conclude that it is unreasonable to presume that all the extra parameter values in Model A are zero. We discuss this fully in Chapter 4.

NOTATION

To facilitate seeing interrelationships between procedures, we use consistent notation throughout. Y_i represents the ith observation from DATA. The first observation is numbered 1 and the last is n, for a total of n observations in DATA. $\hat{Y}_i$ represents the Model's prediction of the ith observation. Y will always represent the variable that we are trying to predict with our model. Other variables giving information about each observation on which we might base conditional predictions will be represented by X. In other words, X will always be used to represent the predictor(s) of Y. So, X_{ij} represents the value of the jth predictor variable for the ith observation. For example, in the internet access example the 50 percentages would be represented as, $Y_1, Y_2, \ldots Y_{50}$, with $n = 50$. X_{i1} could be 1 if the state is in the Eastern or Pacific time zones, and –1 if the state is in the Central or Mountain time zones. In this example, we could use X_{i2} to represent the proportion of college graduates, in which case the value of X_{i2} would be the proportion of college graduates for the ith state.

For model parameters we will use $\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_{p-1}$, for a total of p parameters. Even if we were to know the values of these parameters exactly, we would not expect the model to predict the data exactly. Instead, we expect that some random error will cause the model to predict less than perfectly. We let ε_i represent the unknown amount by which we expect the model to mispredict Y_i. Thus, for the simple model, the basic equation:

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

can be expressed in terms of the true parameter β_0 and the error ε_i as:

$$Y_i = \beta_0 + \varepsilon_i$$

We can never know the true β parameters (or ε) exactly. Instead, we will have estimates of β that we will calculate from the data. These estimates will be labeled, $b_0, b_1, \ldots, b_j, \ldots, b_{p-1}$, respectively. We use $\hat{Y}_i$ to represent the prediction for the ith observation based on the calculated b values. We then let e_i represent the amount by which the predicted value or $\hat{Y}_i$ mispredicts Y_i; that is:

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

The Greek letters β and ε represent the true but unknowable parameters and the Roman letters b and e represent estimates of those parameters calculated from DATA. For the simple model, the model part of the basic data analysis equation is:

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

and we can express that basic equation in terms of our parameter estimates as either:

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

or

$$Y_i = b_0 + e_i$$

Either of the two previous equations are estimates for the basic equation expressed in terms of the unknown parameters as:

$$Y_i = \beta_0 + \varepsilon_i$$

The quantity $\beta_j X_{ij}$ tells us how much we should adjust the basic prediction for the ith observation based on the jth predictor variable. For example, if X_{ij} equals the proportion of college graduates in a state, then β_j specifies how much to adjust, upward or downward, depending on the sign of β_j, the internet access prediction for particular states. A more complicated model involving more parameters can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_j X_{ij} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

$\hat{Y}_i$, the MODEL portion of the data analysis equation, is then represented in terms of the parameter estimates as:

$$\text{MODEL: } \hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_j X_{ij} + \ldots + b_{p-1} X_{i,p-1}$$

The equation:

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

can again be expressed in two ways:

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

or

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_j X_{ij} + \ldots + b_{p-1} X_{i,p-1} + e_i$$

Note that when β values are used on the right side of the equation the appropriate symbol for the error is always ε_i and when b values (i.e., estimates of βs) are used on the right side of the equation the appropriate symbol for the error is always e_i. The reason is that in the first instance the error ε_i is unknown, while in the second instance an estimated value e_i can actually be calculated once $\hat{Y}_i$ is calculated.

We will have to develop a few special symbols here and there, but in general the above notation is all that is required for all the models we consider in this book.

SUMMARY

The basic equation for data analysis is:

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

The data analyst using this equation must resolve two conflicting goals: (a) to add parameters to MODEL so that it is an increasingly better representation of DATA with correspondingly smaller ERROR, and (b) to remove parameters from MODEL so that it will be a simple, parsimonious representation of DATA. Resolving this conflict is equivalent to asking whether the additional parameters are worth it. We use PRE, the index of the proportional reduction in error, to answer this question by comparing appropriately chosen Models C and A. In the traditional language of statistical inference, this is equivalent to comparing a null hypothesis and an alternative hypothesis. The next

several chapters provide the information necessary to judge when PRE is large enough to warrant rejecting Model C in favor of Model A.

We use consistent notation throughout the book to specify our statistical models. Y represents the variable that we are trying to predict ("y-hat," that is, $\hat{Y}$, represents a predicted value of Y) and X represents a predictor variable. Lower-case Greek letters represent true, but unknown, characteristics of the population and Roman letters represent the estimates of those characteristics. Thus, β represents a true, but unknown, population parameter and b represents an estimate of that parameter, which is calculated from DATA. Similarly, ε_i represents the true, but unknown, error of prediction and e_i represents an estimate of that error, which is calculated from DATA. The same statistical model might therefore be expressed using unknown population parameters, for example, $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$, or using symbols that represent estimates that are calculated from DATA, for example, $Y_i = b_0 + b_1 X_{i1} + e_i$. Finally, we may also express the model in terms of predicted values, for example, $\hat{Y}_i = b_0 + b_1 X_{i1}$.