1

Statistical Learning as a Regression Problem

1.1 Getting Started

As a first approximation, one can think of statistical learning as the "muscle car" version of Exploratory Data Analysis (EDA). Just as in EDA , the data are approached with relatively little prior information and examined in a highly inductive manner. Knowledge discovery can be a key goal. But thanks to the enormous developments in computing power and computer algorithms over the past two decades, it is possible to extract information that would have previouslybeen inaccessible. In addition, because statistical learning has evolved in a number of different disciplines, its goals and approaches are far more varied than conventional EDA.

In this book, the focus is on statistical learning procedures that can be understood within a regression framework. For a wide variety of applications, this will not pose a significant constraint and will greatly facilitate the exposition. The researchers in statistics, applied mathematics and computer science responsible for most statistical learning techniques often employ their own distinct jargon and have a penchant for attaching cute, but somewhat obscure, labels to their products: bagging, boosting, bundling, random forests, the lasso, and others. There is also widespread use of acronyms: CART, MARS, MART, LARS, and many more. A regression framework provides a convenient and instructive structure in which these procedures can be more easily understood.

After a brief discussion of how statisticians think about regression analysis, the chapter introduces a number of key concepts and raises broader issues that reappear in later chapters. It may be a little difficult for some readers to follow parts of the discussion, or its motivation, the first time around. However, later chapters will flow far better with some this preliminary material on the table, and readers are encouraged to return to this chapter as needed.

R.A. Berk, *Statistical Learning from a Regression Perspective*,
DOI: 10.1007/978-0-387-77501-2_1, © Springer Science+Business Media, LLC 2008

1.2 Setting the Regression Context

We begin with a brief consideration of what regression analysis is. A knee-jerk response in many academic disciplines and policy applications may be to equate regression analysis with causal modeling. This is too narrow and even misleading. Causal modeling is actually an interpretive framework that is imposed on the results of a regression analysis. An alternative knee-jerk response may be to equate regression analysis with the general linear model. At most, the general linear model can be seen as a special case of regression analysis.

Statisticians commonly define regression so that the goal is to understand "as far as possible with the available data how the conditional distribution of some response **y** varies across subpopulations determined by the possible values of the predictor or predictors" (Cook and Weisberg, 1999: 27). That is, interest centers on the distribution of the response variable Y conditioning on one or more predictors X.

This definition includes a wide variety of elementary procedures easily implemented in R. (See, for example, Maindonald and Braun, 2007: Chapter 2.) For example, consider Figures 1.1 and 1.2. The first shows the distribution of SAT scores for recent applicants to a major university, who self-identify as "Hispanic." The second shows the distribution of SAT scores for recent applicants to that same university, who self-identify as "Asian."

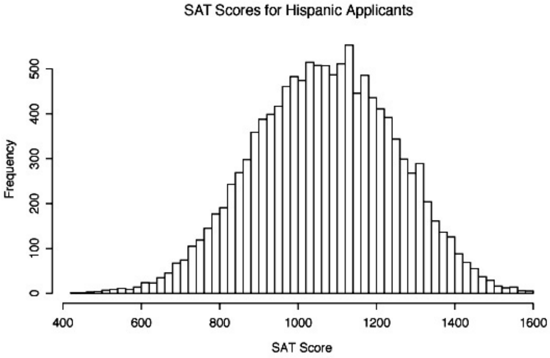

Fig. 1.1. Distribution of SAT scores for Hispanic applicants.

It is clear that the two distributions differ substantially. The Asian distribution is shifted to the right, leading to a distribution with a higher mean (1227 compared to 1072), a smaller standard deviation (170 compared to 180), and greater skewing. A comparative description of the two histograms alone constitutes a proper regression analysis. Using various summary statistics, some key features of the two displays are compared and contrasted (Berk, 2003: Chapter 1).

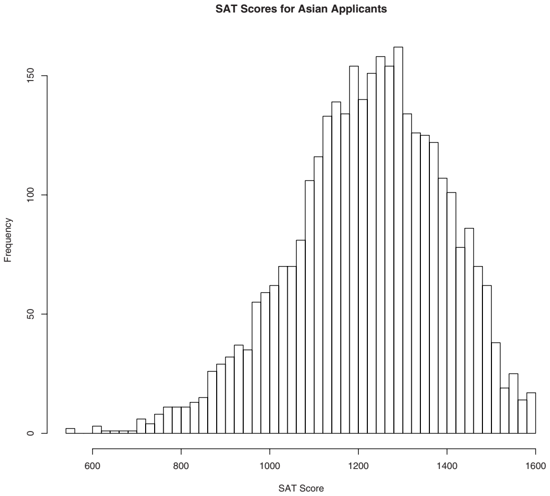

Fig. 1.2. Distribution of SAT scores for Asian applicants.

Should one be especially interested in a comparison of the means, one could proceed descriptively with a conventional least squares regression analysis as a special case. That is, for each observation i, one could let

$$\hat{y}_i = \beta_0 + \beta_1 x_i, \tag{1.1}$$

where the response variable y_i is each applicant's SAT score, x_i is an indicator variable coded "1" if the applicant is Asian and "0" if the applicant is Hispanic, β_0 is the mean SAT score for Hispanic applicants, β_1 is how much larger (or smaller) the mean SAT score for Asian applicants happens to be, and i is an index running from 1 to the number of Hispanic and Asian applicants, N. Here, $\beta_0 = 1072$ and $\beta_1 = (1227 - 1072) = 155$.

One can reverse the roles of the two variables and undertake another legitimate kind of regression analysis. Figure 1.3 shows a scatterplot with the SAT score on the horizontal axis and on the vertical axis an indicator variable coded "1" if the applicant self-identifies as Asian and "0" if the applicant self-identifies as Hispanic. The points in the plot have been jittered vertically to make the scatterplot easier to to read. Jittering adds a bit of noise to each observation, in this case for the Asian indicator variable.

Because the higher points in Figure 1.3 (around 1.0) are to the right of the lower points (around 0.0), the proportion of Asians increases moving from left to right. That is, the conditional proportion increases with SAT score. A least squares regression line overlaid on the scatterplot can quantify the association. It is of the same form as Equation 1.1 with the roles of Y and X exchanged. The slope, β_1, indicates that for each additional 100 SAT points, Asian representation, compared to Hispanics, increases about 8% on the av-

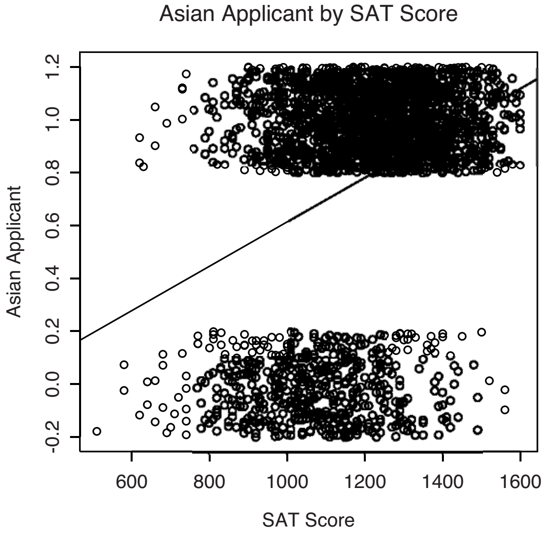

Fig. 1.3. Asian Applicants by SAT Score

erage. The intercept, β_0, is in this case a negative number, indicating that using a straight line may not be the best way to describe the relationship. An S-shaped function such as the logistic curve might do a better job. Moving to logistic regression would have still been a regression analysis.

There is no requirement that either variable be measured on an equal interval scale. Both variables can be categorical. Table 1.1 shows a cross-tabulation, using those same college applicants, for whether the applicant self-identifies as African-American and whether the applicant falls within a special admissions category of "athlete." The athlete designation usually places the applicant in a special pool that only includes other athletes.

It is readily apparent from the marginal distributions that athlete applicants and African-American applicants represent very small fractions of the total number of applicants (about .6% and 4.6%, respectively). One can also see from the within-row percentages that 2.1% of all African-American applicants are designated as athletes whereas around .5% of all other applicants are. This is a difference of 1.6%. Stated differently, African-American applicants are over four times more likely to be placed in the athlete applicant pool.

Table 1.1 is another example of a regression analysis. The proportion placed in the athlete applicant pool is computed conditional on whether the applicant self-identifies as African-American. And as before, one can arrive at the very same results with a least squares regression analysis. In Equation 1.1, the response y_i is an indicator variable coded "1" if the applica-

	Not an Athlete (%)	Athlete (%)	Row Percentage
Not Black	99.5	0.5	95.3
Black	97.9	2.1	4.6
Column Percentage	99.4	0.6	100=96,277

Table 1.1. Ethnicity by athlete designation.

tion is placed in the athlete pool and coded "0" otherwise. The explanatory variable x_i is an indicator variable coded "1" if the applicant is self-identifies as African-American and "0" if not. The value of β_0 is then .005, and $\beta_1 = .021 - .005 = .016$.

Figure 1.4 displays the most familiar kind of plot for regression analysis. The SAT score is plotted against an applicant's household income, with the regression line overlaid. Both variables are on an equal interval scale. The scatterplot was constructed for a random sample of 500 applicants to make the graph more legible.

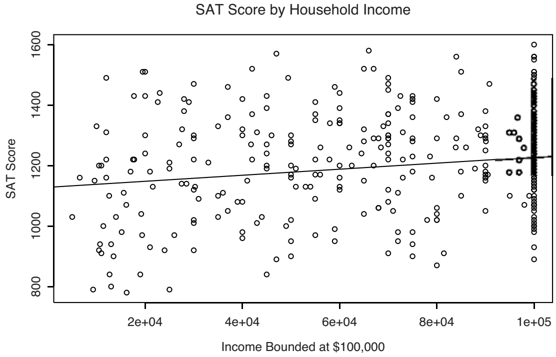

Fig. 1.4. SAT scores by family income.

As one moves from left to right, one can see how the conditional mean of the SAT score (y_i) changes with the household income (x_i). The value of β_0 is about 1200, which is the mean SAT score for households with no reported income. In this instance, it is not clear that the value of β_0 makes any substantive sense, but it is needed to locate the regression line. For each \$1000 of income, the mean SAT score increases about 1 point. But note that families with more than \$100,000 of income are treated as having no more than \$100,000 of income (because of the way the application forms are filled out). Consequently, the slope would be perhaps a little steeper than had the full range of income figures been available.

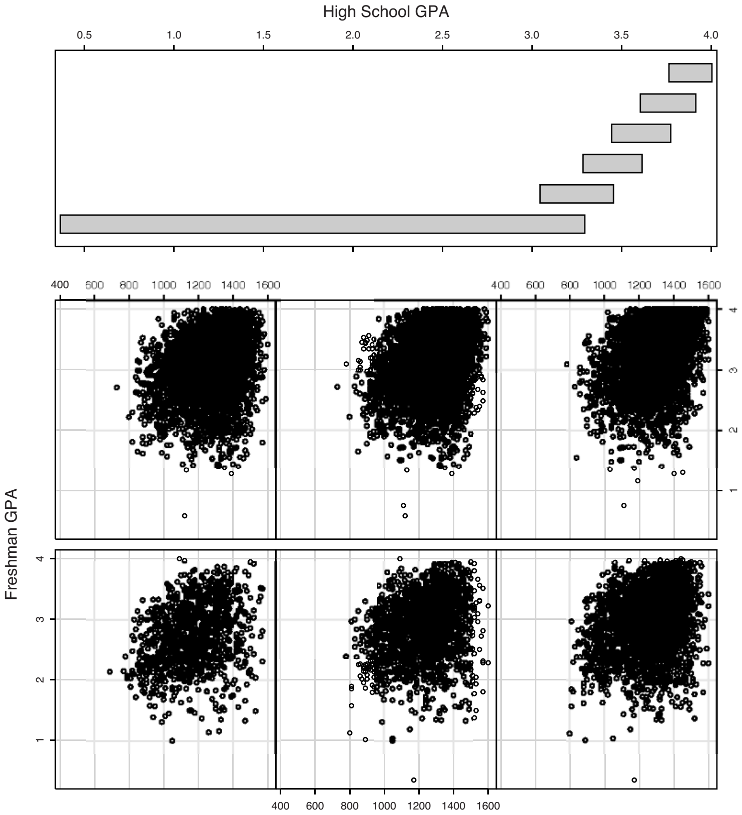

Fig. 1.5. Freshman GPA on SAT holding high school GPA constant.

Each of the four examples could easily have included more than one predictor. For example, Figure 1.5 shows a conditioning plot, sometimes called a "coplot" (Cleveland, 1993: 182–190). There are scatterplots of the Grade Point Average (GPA) of college freshmen, from that same university, against their SAT scores, holding constant their high school GPA. The scatterplots are read left to right starting with the bottom row. For the second row, one starts with the first plot on the left side. The conditioning subsets, defined by high school GPA, are shown in the top panel. The spans of the conditioning variable and amount of overlap between the subsets have been tuned to allow for a sufficient number of observations in each.

One can see that there is a positive association between SAT score and freshman GPA, within ranges of high school GPA. Holding the high school GPA approximately constant, the SAT score is related to performance in the first year of college. One can also see that the plots shift upward as one moves from the lower-left corner to the upper-right corner, indicating that the high school GPA is also positively related to performance in the first year of college, holding the SAT score constant. Note, for instance, that the vertical slice of points at an SAT score of 1200 rises toward the upper boundary of the plots. Finally, there is apparently a ceiling effect of the 4.0 upper limit for freshman GPA in the top three graphs, implying that the relationship between SAT score and freshman GPA would probably be stronger if students who performed especially well could receive grades higher than 4.0. More precise statements of these sorts could be made by adding a second predictor to a regression equation of the form shown in Equation 1.1.

There are several broad lessons in these initial illustrations. First, regression analysis seeks to characterize conditional distributions. The response variable and the predictors can be categorical or quantitative variables. That's the long and the short of it.

Second, within that definition, one is free to choose whatever procedures seem to be the most useful. Graphs, for instance, are not automatically better or worse than numerical summaries, and a wide variety of each can be helpful, at least in principle. The choice depends on the nature of the information to be extracted from the data and the audience for the results. For example, graphs can be more effective than numerical summaries when broad patterns in the data are more important than a few precise values. If numerical summaries are desirable, there are no necessary restrictions on the functional forms used or on how the numbers are computed. Classical linear regression is just a special case. Regression analysis is a "big-tent" procedure.

Third, although analyses such as these may reflect cause-and-effect relationships or motivate a search for causal explanations, there is nothing in a regression analysis that requires inferences about cause and effect. And there is certainly no requirement that the regression analysis be formulated as a "causal model" in which the causal mechanisms by which the data were produced are explicitly represented (Berk, 2003: Chapter 1). The fact that regression equations are often advertised as causal models does not mean the two are the same.

Fourth, there is also nothing in regression analysis that requires statistical inference: formal tests of null hypotheses or confidence intervals. These can sometimes be very useful but go beyond the definition of a regression analysis (Berk, 2003: Chapter 1). Theyare an add-on. Moreover, even if the data are generated in a manner that can justify statistical inference, there are real questions about how to interpret the p-values that result after an extensive exploratory analysis. In general, the p-values will be too small, sometimes dramatically so. A bit more is said about this shortly in the context of "data snooping." In later chapters, a more detailed discussion is undertaken under the rubric of overfitting.

Finally, a regression analysis can serve a variety of purposes. Most directly, a regression analysis can be used to describe the relationships between variables. For example, Figure 1.5 addresses the nature of the association between freshman GPA and SAT score, holding high school GPA constant. Insofar as a relationship can be found, it may also serve as the basis for useful forecasting. For instance, if SAT is an effective predictor of later performance in college, beyond what one might learn from a student's GPA in high school, an SAT score might be an important piece of information to use in an admission decision. And, although a regression analysis by itself is silent on cause and effect, it can under some circumstances be applied to characterize relationships taken to be causal because of additional information about how the data were generated. In perhaps the best situation, a regression analysis is applied to data from a randomized experiment in which one or more interventions are consciously manipulated, and the goal is to estimate a "contrast" with respect to the outcome for different treatment groups. For example, although a person's race cannot be manipulated, the race recorded in a home mortgage application can be. Then one can estimate the causal effect of a racial label on the interest rate offered, other things being equal. (Studies like this have been done.) It cannot be overemphasized that causal inference comes from knowledge about the experiment, not from the regression analysis. The regression analysis merely describes relationships that are already demonstrably causal.

1.3 The Transition to Statistical Learning

Statistical learning within a regression framework retains the focus on the conditional distribution of a response variable with respect to one or more predictors. Various features of that conditional distribution can be relevant, but the conditional mean will play a central role. How does the conditional mean of the response vary depending on the values of its predictors?

Where statistical learning can differ from conventional linear regression is in how that conditional relationship comes to be characterized. In conventional linear regression, functional forms linking the predictors to the response are determined before the fitting process begins. The same is true of the generalized linear model (e.g., logistic regression and Poisson regression) and conventional nonlinear regression. In that sense, all of these procedures can be called parametric.

In statistical learning, there is far less reliance on prior information when functional forms are determined to link predictors to the response. Although there will sometimes be constraints on the kinds of functions permitted, the functional forms are, by and large, arrived at inductively from the data. In that sense, statistical learning procedures can be called nonparametric.

Statistical learning is likely to shine when the functional forms are unknown and substantially nonlinear. Readers familiar with stepwise regression already have some appreciation for the look and feel of several statistical learning features. Readers familiar with smoothers can probably anticipate that smoothing a scatterplot can be a form of statistical learning. The use of the word "learning" is a metaphor for the exploratory manner in which relationships between variables are determined.

1.3.1 Some Goals of Statistical Learning

As with all statistical procedures, statistical learning necessarily raises a number of "meta-issues." We need to consider these over the next several pages so that the key features of the context and underpinnings of statistical learning are familiar. Parts may seem a little abstract but they lay the foundation for the more nuts-and-bolts material that follow. Some of the material may benefit from rereading after later chapters have been read.

As with any statistical procedure, the goals of statistical learning depend fundamentally on how the data were generated. It is often useful to think about the data on hand as the product of a specific data-generation process, also sometimes called a data-generation mechanism . The data-generation process is a product of natural forces and the activities of researchers.

An example of data generated primarily by natural forces might be a time series of air quality measures in a particular metropolitan area. An example of data generated primarily by researchers might be a clinical trial for a new cancer treatment. An example of data generated by a rich mix of the two might be a probability sample of registered voters in which voting preferences are reported. The degree to which researchers intervene in a natural process determines how much of the data-generation process will be characterized by research protocols such as random assignment or probability sampling.

A conceptual distinction is often made between two kinds of data-generation processes. One kind conceives of a stochastic process with the observations on hand a realization of that real process. For example, suppose that there are $i = 1, 2, \ldots, N$ observations in a dataset. Nature generates each observed value of y_i using, say, of $y_i = 5 + 3x_i + 1.5z_i + \varepsilon_i$, where x_i and z_i are predictors, and the value ε_i behaves as if drawn from a single distribution with a mean of 0.0, independently of any other ε_j, and independently of x_i and z_i. The systematic part of the data-generation process is $5 + 3x_i + 1.5z_i$. This is usually treated as fixed. The stochastic part is ε_i. Chance is built in solely through ε_i so that y_i is a random variable. Sometimes the distribution from which ε_i is drawn is said to be of a particular form such as the normal. In more formal terms: $\varepsilon_i \sim NIID(0, \sigma^2)$.

Another kind of data-generation process assumes that there exists a population of potential observations. Suppose that in this population there are, again, three variables. There is a response variable y_i and two predictors x_i and z_i. All three variables in the population are fixed; there is no stochastic component. If one were able to observe all of the variables for all elements in the population over and over, their values would not change. For each possible configuration of values for x_i and z_i, there is a mean value for the response. Nature computes these means using, for instance, $5 + 3x_i + 1.5z_i$, where for each mean, the index i is limited to those observations with the same values for x_i and z_i.

Commonly, there will be variation in the values of y_i around each conditional mean. The values of y_i around each conditional mean are unrelated to the values of the predictors, and are sometimes said to have a particular distribution, often the normal. A probability sample of size N is taken from the population. The three variables are now random variables. If a second probability sample were drawn, many (or even all) of the values for each of the variables would be different. In the sample, one can write, as before, $y_i = 5 + 3x_i + 1.5z_i + \varepsilon_i$, where $\varepsilon_i \sim NIID(0, \sigma^2)$. But now, ε_i results from the probability sampling, not from nature. It is common to treat the predictors as fixed, once they materialize in a given sample. In short, both kinds of data-generation processes can lead to the same formal expression of how y_i came to be.

Whether the data are a realization or a probability sample, an important goal can be to estimate from the data on hand how the two predictors are related to the response. For the stochastic process, that would imply trying to accurately represent the systematic component of y_i. For the probability sample, that would imply trying to accurately represent how the conditional means in the population are constructed from the predictors. Thus, both such enterprises are really the same. Indeed, effectively the same statistical tools can be used whether the data are treated as a realization of a stochastic process or a probability sample drawn from a population. Nevertheless, the two accounts can have different implications for the credibility of any subsequent analysis. Assumptions built into the data analysis need to be justified by a credible explanation of how the data were actually generated. In practice, therefore, that account will either be about a stochastic process or about what is going on in the population from which the data were sampled at random. It can also be important to verify what the sampling design was and whether it was implemented properly.

It is common to represent the data-generation process with a statistical model. A broad and popular class of data-generation processes can be written as $Y = f(X) + \varepsilon$, where Y is the response variable, X is a set of predictors, ε is a disturbance term, and $f(X)$ is some function mapping the predictors to the systematic part of Y. For a conventional linear model, $f(X)$ is a linear combination of the p predictors: $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$. An alternative might be to exponentiate the linear combination of predictors, as is common in Poisson regression. In statistical learning, $f(X)$ is far more open-ended. However, for a set of N observations $i = 1, 2, \ldots, N$, each ε_i is often assumed to have the same convenient properties as it does for classical linear regression: $\varepsilon_i \sim NIID(0, \sigma^2)$.

A credible parametric regression model depends on specifying a number of details justified with reference to subject matter knowledge and past research. For example, what is it about how the data were generated that permits one to assume that the predictors enter the model in a particular manner? Although statistical learning is often less assumption bound, it is never assumption free. There can be, for instance, the need to explain why the disturbances represented by ε_i have an expectation of zero, are independent of the predictors (or at least uncorrelated with them), and are independent of one another. These are very strong statements about how the data were generated. They require careful thought and justification, often beginning with a decision about whether to treat the data as a random realization, a random sample, or "just" a dataset.

Such questions are often difficult when $f(X)$ is known to take a particular parametric form. When $f(X)$ is to be largely determined by the data, the questions can be daunting. For example, how does one argue that ε_i is at least uncorrelated with the predictors when the transformations to be applied to these predictors are not yet known? And if any of the assumptions made about ε_i are substantially wrong, the results of the data analysis can be substantially wrong as well.

Much the same formulation can apply to categorical outcomes, so that one might write $G = f(X) + \varepsilon$. It is more common to write the expression for the conditional expectation of G and then to characterize the uncertainty separately. In the case of a binary response, for example, one could write $G = f(X) + \varepsilon$, with G coded as "1" or "0". Usually, however, one replaces G with the mean-value parameter (McCullagh and Nelder, 1989: 30), in this instance a probability, alters the $f(X)$ accordingly, and then indicates that the probability is a parameter in the binomial distribution by which the binary outcomes are generated. The binomial distribution is responsible for the uncertainty. But just as for quantitative outcomes, a strong subject matter case needs to be made that a particular formulation applies. Thus, what is the rationale for assuming that all observations with the same set of predictor values are subject to the exact same conditional probability? Why is there no heterogeneity in that conditional probability? Or, what reason is there to believe that, conditional on the predictor values, the binary events are independent of one another.

Taking the data-generation process into account is not usually by itself sufficient. It can also be important to consider what use will be made of the data analysis. One key dimension is what information from a data analysis will figure in the conclusions to be drawn. Sometimes interest centers primarily on the fitted values for the $f(X)$. There is no concern with representing how the predictors are related to the response. Building on an earlier example, the goal may be to make admissions decisions to a university based on information about the performance in their freshman year of previous successful applicants. Why some students do better than others does not matter because the admissions office is in no position to do anything about that.

Alternatively, the primary concern may be in learning how inputs are related to outputs. Organizations on campus that provide support services, such as tutoring for matriculating students, will want some guidance on where to intervene. Are students for whom English is not a primary language, for instance, at greater risk for poor academic performance? In short, whether attention is directed toward the fitted values, the relationships between inputs and outputs, or both, will affect how the data analysis is done and assessments of its worth.

There is another use that should be briefly mentioned, but does not fall within a regression perspective. One might want to compute summary statistics for the response, conditional upon certain values of the predictors. For example, one might want an estimate of number of students for whom English is a second language, and who will not graduate. No comparisons will be made to other students, and whether any graduation problems stem from language difficulties or from other factors is not of immediate interest. There is no concern with how the response changes depending on the values of predictors. Such applications are not considered here.

Another important factor shaping a data analysis can be what kind of story is likely to be told. Just as in parametric regression, there can be four kinds of stories.

1. *A Causal Story*—A data-generation process is assumed and given a causal interpretation (Freedman, 2004). For any study unit i subject to this process, the values of any X_{ij} can be set (manipulated) independently of the values of any other predictor. A researcher or nature determines these values. Then some natural process maps X_i onto $f(X_i)$ and attaches a value of ε_i. The value of ε_i behaves as if drawn at random from some distribution, independently of the values of X_i. Once again, the absence of a linear correlation can often be sufficient. Other units are subjected to the same process with each value of ε_i drawn independently of one another. In addition, the values of X_i set for any given unit do not alter the response of any other unit. All this can be framed as the stochastic process responsible for the data on hand, or for the population from which a random sample is drawn.

   In this setting, the job of statistical learning is to recover the $f(X)$ nature uses. The residuals, defined as the arithmetic difference between the observed response values and the estimated $f(X)$, are used to characterize the distribution of the ε_i. These goals are the same as for all causal modeling. What statistical learning offers is powerful tools to help learn inductively what $f(X)$ may be. One can think of this as function estimation. For example, one might be interested in the function that turns a person's human capital into earnings, or volatile hydrocarbons into ozone.

2. *A Conditional Distribution Story*—The basic formulation is the same; there is an assumed model with many of its features to be informed by the data. However, no causal interpretation is given. The $f(X)$ is descriptive only. One is satisfied saying something such as $Y \sim N(f(X), \sigma^2)$, where $f(X)$ now represents the conditional mean or perhaps the condi-

tional expected value. To take a simple illustration, for bivariate linear regression one might write $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, with $f(X) = \beta_0 + \beta_1 x_i$ as the expected value of y_i.

Interest centers on the conditional distribution $Y|X$, but no claims are made that manipulating the value of X alters Y. So, there is no need to assume that one or more of the predictors can be individually manipulated, or that setting a predictor at a particular value for the ith case does not affect the response of the jth case. The key assumptions center on distribution of the ε_i. For example, one must persuasively argue that ε_i is at least uncorrelated with, and ideally independent of, the predictors. As in the causal story, this is a demanding task when the functions to be applied to X are not yet known.

Just as for the causal modeling story, statistical learning is used to characterize the $f(X)$, taken to be real and "out there." This too may be seen as function estimation. Thus, one might be interested in the function that associates a score on the mathematics Scholastic Aptitude Test with a score on the verbal Scholastic Aptitude Test, or the function that associates the number of predators with the number of prey.

3. *A Data Summary Story*—Although there can be a data-generation process in principle, it plays no direct role in the data analysis. Thus, there is no assumed model by which the data were generated. There are only the data. Statistical learning is then used as a data reduction tool. The aim is to provide through some data-derived $f(X)$ an accurate and accessible summary of how, in the data on hand, y_i is systematically related to a set of predictors $x_{ij}, i = 1, 2, \ldots, N; j = 1, 2, \ldots, p$. In the same spirit, there can be residuals just as in the two earlier stories, but these have no necessary correspondence to some ε_i; residuals are computed from the data whereas the ε_i is a feature of a hypothetical model.

   One may only be interested in the conditional distribution of the data being analyzed or one may also wish to generalize beyond the data to other similar settings. In some cases, interest ultimately may be in the $f(X)$, but its recovery is at least premature. For example, one might still be interested in the function that turns a person's human capital into earnings but have only proxy measures for human capital (e.g., years of education).

4. *A Forecasting Story*—Using the data on hand, one constructs a function with which to make forecasts. If there is some real $f(X)$ "out there" that can be exploited for forecasting, all the better. But there is no interest in causal effects, no interest in the conditional distribution $Y|X$, and no interest in data reduction unless they can improve forecasting skill. Statistical learning is a tool for developing a useful forecasting apparatus.

> For example, one might be interested in forecasting water quality at local beaches from the previous week's precipitation.

The difference between the causal story and the data reduction story is sometimes a matter of degree. Consider a simple example. Suppose there is a single predictor, and the $f(X)$ is actually a parabola: $f(X) = \beta_0 + \beta_1 X + \beta_2 X^2$. For the causal story, the goal is to determine from the data that $f(X)$ takes this exact form, including accurate estimates of the values for its three parameters. For the data summary story, a data analyst might be satisfied learning that the $f(X)$ is smooth and convex, and the approximate value of X at which $f(X)$ is minimized.

Why would the researcher settle for less than the full causal story if $f(X)$ is causal? Perhaps the statistical learning procedure applied does not work well for these kinds of functions. Or perhaps the predictor is measured poorly, or is at best a proxy for the predictor needed. Indeed, the predictor used may be the wrong one altogether. Or perhaps the response is measured with so much noise that the $f(X)$ is effectively obscured.

A consideration of the four possible stories raises a topic to which we return many times: the appropriate loss function to be employed. Almost inevitably, the empirical correspondence between Y or G and the $\hat{f}(X)$ will be imperfect. The fitting enterprise, therefore, depends on how the disparities between the observed response variable values and the fitted response variable values are treated. Usual practice is to minimize a loss function with these disparities as inputs. In conventional least squares regression, for example, the loss function is quadratic.

In the pages ahead, a variety of loss functions are considered. Some include a penalty for fitted values that are unnecessarily complex. Some allow for asymmetric losses so that for classification exercises, false positives can be weighted differently from false negatives. For example, the costs of mistakenly concluding that a patient has cancer are likely to be very different from the costs of mistakenly concluding that a patient is cancer free. We show that the loss function one chooses can dramatically affect the fitted values, with important implications for the story to be told and how that story will be used.

1.3.2 Statistical Inference

In principle, statistical tests and confidence intervals can be important for each of the four stories. If there is a data-generation process to be characterized, it is common to proceed as if the data have been produced in a manner that introduces some randomness. As already mentioned, sometimes that randomness is a product of probability sampling undertaken by the researcher. Sometimes that randomness is taken to be an inherent part of how nature generated the data (sometimes called "model-based sampling"). In either case, if the data were generated again, they are almost certain to be different, at least a bit, because of “chance.” So, the role of chance needs to be addressed as part of the data analysis.

For example, the expression $Y = f(X) + \varepsilon$, ε is often seen as the source of the randomness, whether it be a product of actions undertaken by the researcher or by nature. Should parametric regression be applied to characterize the $f(X)$, statistical tests and confidence intervals can naturally follow. The same can hold for statistical learning.

Statistical inference has no role when description of the data on hand is the only goal or when the links to a population or stochastic process cannot be credibly articulated in a manner the p-values require. For example, testing hypotheses about how average earnings vary with seniority for a quota sample of clerical workers intercepted in a local shopping mall will likely produce uninterpretable p-values, even if it were possible to figure out the population to which inferences were being made. Absent random sampling, what would be required is a credible account of a natural data-generation process meeting the requisite assumptions. As already noted, such accounts are often very difficult to construct.

Statistical inference can be important for forecasting. But, the forecasts must be into a probability sample from the same population as the data on hand, or into a realization of the same stochastic process. Otherwise, the computed probabilities are likely to have no useful meaning. If, for example, the forecasts are made into a random sample from a different population, or into a convenience sample, the forecasts can differ from what actually transpires because the response is not related to the predictors in the same fashion it was in the data from which the forecasts were constructed. More than random error is involved. A 95% confidence interval, for example, will not cover the population value 95% of the time because each interval is offset by some amount of bias.

Some readers may wonder why there has been no discussion of random assignment as a way in which a chance process can affect data. Random assignment to a treatment group or to a control group within a randomized clinical trial, for instance, can be formulated within a probability sampling framework. But the inferential issues are somewhat different from random sampling in observational studies. Under random assignment, the uncertainty involves chance variation in estimated treatment effects because of the way in which a fixed group of study units is assigned to treatments and control conditions. There is usually no larger population to which inferences are formally being drawn. There is also commonly the need to construct a theoretical model of causal effects. In short, although many randomized experiments in principle can be analyzed using statistical learning procedures (e.g., within a dose-response framework), they are rarely needed and even more rarely used. Randomized experiments are not considered further in this book.

To summarize, statistical inference can in principle play a useful role across a wide range of statistical learning applications. In practice, however, we show that statistical inference is not particularly salient in statistical learning. Even if the data are known to have been generated in a manner that might justify statistical inference, and even if all of the necessary prerequisites for statistical inference are present (e.g., all of the needed predictors), the way statistical learning is undertaken can make statistical inference inappropriate.

1.3.3 Some Initial Cautions

Expositions of statistical learning commonly assume that the goal is to tell a causal story or a conditional distribution story. There is some truth external to the data that statistical learning will help to reveal. One wants to know the $f(X)$ and if the $f(X)$ can be represented as parametric, the values of its parameters as well.

It is important to be clear from the start that no credible statistician would ever claim that even when all of the necessary predictors are present and perfectly measured, there are one or more statistical learning procedures that will exactly capture the $f(X)$. The data with which one works will necessarily be an imperfect reflection of the $f(X)$ because of the impact of ε; the values from the $f(X)$ and ε are thoroughly commingled as Y is generated. Because ε is unobservable, it cannot be removed from Y in order to obtain the $f(X)$. This is a fundamental problem inherent in all estimation.

Matters are further complicated by the need to learn from the data both the underlying functional forms and the values of key parameters. We show later that the need to learn about the functional forms can place very heavy demands on a dataset. Large samples are often necessary with the observations more densely packed where the nonlinear functions are changing more rapidly. We also show that the performance of all statistical learning procedures is significantly determined by tuning parameters for which only very broad guidelines are likely to exist. Craft lore rather than proved theorems can dominate practice.

Just as in parametric regression, researchers have to be satisfied with one of two possible fallback positions. The first entails desirable finite sample properties such as unbiasedness and efficiency. For reasons that become clear later, it is difficult for statistical learning procedures to satisfy these requirements. The second fallback position entails desirable asymptotic properties such as consistency. Of late there have been some successes for a number of statistical learning procedures (Breiman, 2004; Jiang, 2004; Lugosi and Vayatis, 2004; Efron et al., 2004; Zhang and Yu, 2005; Bickel et al., 2006; Bühlmann, 2006; Traskin, 2008), but as considered in subsequent chapters, these formal results often do not answer the questions that applied researchers would make a top priority. And there remains, as always, the matter of how best to make use of asymptotic results for the sample on hand.

In practice, moreover, real world studies rarely cooperate with what the theoretical statistical work requires. Perhaps most obviously, if the goal is to recover the $f(X)$, X must be known; each and every predictor must be identified. Then, all of the predictors must be in the dataset to be analyzed and measured without error (even random measurement error). It is difficult to find examples in which this is even approximately true, and most applications are not even close.

If the goal is to recover the $f(X)$, often the best that one can do is to proceed with the understanding that one's results will be biased and inconsistent, sometimes substantially. Because both the direction and the magnitude of these difficulties are usually unknown—if they were known, they would not be difficulties—it can be difficult to determine what to make of the fitted values.

By default, therefore, real applications are usually about data reduction, and occasionally about forecasting. One important implication is that for these stories, a substantial portion of the theoretical justification for many statistical learning procedures provides indirect guidance at best. Consequently, rationales for particular statistical learning applications tend to rely on in-sample features of the fit, forecasting skill, and subject matter knowledge. This is a point to which we return many times in the pages ahead.

Finally, there is nothing in any of the four stories that requires statistical learning. One can apply parametric regression to the very same ends. Statistical learning earns its keep when, for causal and estimation stories, the $f(X)$ is not well understood but is likely to be substantially nonlinear and hence, complex. In the same spirit, statistical learning earns its keep for data reduction and forecasting when the systematic information in the data is best captured with a complex fit.

"Complexity" can be conceptualized in different ways, many of which are not easily represented within a statistical framework (Zellner et al., 2001). For example, is "simple" the opposite of "complex" or is "parsimonious" a better choice? And if parsimonious, how might one translate that into statistical concepts? As a practical matter, complexity is commonly represented by the degrees of freedom "used up" in the fitting process. A statistical learning procedure produces a more complex fit when more degrees of freedom are used up. In parametric regression, for instance, a larger number of regression coefficients implies that a larger number of degrees of freedom will be spent, and that a more complex rendering of the $f(X)$ will result. We show that this is a special case of how complexity is often measured in statistical learning. Indeed, relying solely on the degrees of freedom used up can be unsatisfying. Which is more complex: $\hat{y}_i = \hat{\beta}x_i$ or $\hat{y}_i = \hat{\beta}x_i^2$? Or, are they equally complex? We show that with some of the most recent and advanced statistical learning applications, complexity is even more difficult to conceptualize and measure.

1.3.4 A Cartoon Illustration

One can get a more grounded sense of the issues by comparing a hypothetical fit using linear regression to a hypothetical fit that might result from statistical learning. Suppose one is concerned about the number of misconduct incidents committed by prison inmates. The response variable is the number of such incidents reported for each inmate during a one year period. The predictor is the nominal sentence length of the prison term each prisoner received.

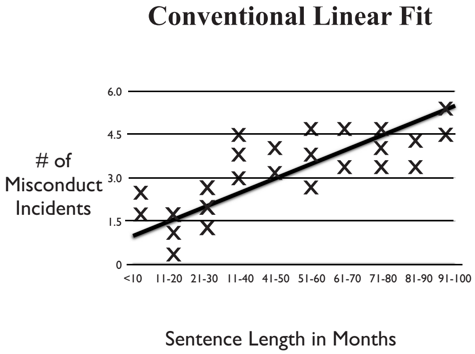

Fig. 1.6. Imposing a least squares line on the data.

Figure 1.6 is a hypothetical scatterplot with a least squares regression line overlaid. The regression line shows that in general, the relationship is positive. The number of misconduct incidents increases with sentence length. To the eye, the fit is quite good, and a positive relationship is hardly surprising. Longer prison sentences are generally associated with more serious crimes and the criminal histories of "habitual offenders." Both are thought to characterize inmates who would not "program" well. So, a researcher might well be satisfied with the results.

However, Figure 1.7 shows that if the data are allowed to play a larger role in determining the functional form, a somewhat different story emerges. The number of misconduct incidents decreases with sentences up to 20 months, increases rapidly with sentences from 20 to 40 months, and is almost flat thereafter. There is probably no simple explanation for this pattern.

The sentence lengths for which the relationship is nearly flat may represent older inmates subject to sentence length enhancements because of earlier convictions. It is well known that older inmates are much less likely to get into trouble in prison. The sentence lengths associated with the rapid increase in misconduct incidents may reflect the behavior of younger inmates, often gang members, convicted of serious crimes, but not yet subject to sentencing enhancements (i.e., "gang-bangers"). The sentence lengths associated with declines in misconduct could represent inmates with short sentences who wish to stay out of any trouble that would jeopardize their release. The relationship is negative, perhaps because the risks of becoming involved in a misconduct incident increase with time behind bars. So, those inmates who are thinking ahead to their parole dates may be especially careful if their period of risk is longer.

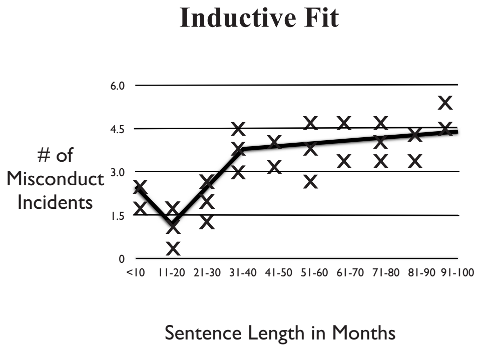

Fig. 1.7. Letting the data determine the functional form.

It is also possible that the inductive fit in Figure 1.7 is dominated by happenstance. Note that the nonlinear fit is substantially driven by three inmates with sentences between 11 and 20 months. If for those three observations, the number of misconduct incidents were a bit greater, or if those three observations were missing altogether, the original linear fit would be far more satisfactory.

A comparison between Figure 1.6 and Figure 1.7 raises an important issue to which we return many times. When the linear fit was imposed in Figure 1.6, the support for that line came from the full range of the available data. With the inductive fit, the support was highly local. For these data, the linear fit may not capture as well the relationship between sentence length and misconduct. But the linear fit is likely to be more stable under random variation in the data. The inductive fit may capture better the relationship between sentence length and misconduct. But the inductive fit may be relatively unstable under random variation in the data. In other words, the inductive fit may be too much a product of overfitting. Random variation is being interpreted as systematic variation. These and related points are more formally addressed in the pages ahead.

To complete the story, Figure 1.8 shows an inductive fit for a binary outcome, coded "1" for misconduct and "0" for no misconduct. Each "X" in

Figure 1.8 represents many data points that cannot be seen because of overprinting. The same issues arise. One can impose a functional form, such as the logistic or, as in Figure 1.8, allow the functional form to respond substantially to the data.

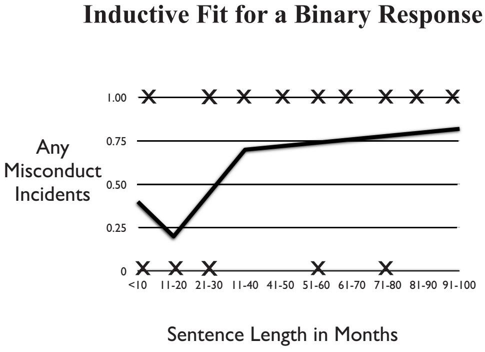

Fig. 1.8. Letting the data determine the functional form for a binary response.

1.3.5 A Taste of Things to Come

Figures 1.6 through 1.8 begin to raise a number of difficult questions that are addressed throughout the course of the book. Sometimes there some are good answers, but often the best answers are highly provisional. And often the answers come in the form of statistical procedures that at first seem somewhat curious. Consider the following sequences of operations.

1. Fit the data with some conventional procedure.
2. Compute the residuals as the difference between the fitted values and the actual values.
3. Compute a measure of fit from the residuals, such as the error sum of squares (also known as the residual sum of squares).
4. Apply the fitting procedure again, but weight the observations so that the cases for which the absolute value of the residuals is larger receive more weight, and the cases for which the absolute value of the residuals is smaller receive less weight.
5. Repeat the first four steps 1000 times.
6. Compute the final set of fitted values as a weighted average of the fitted values over the 1000 passes through the data, with the weights a function

of the error sum of squares computed in Step 3. Fitted values with a better overall fit are given more weight in the averaging. That is, each $\hat{y}_i$ is a weighted average of 1000 fitted values for each observation i, weighted so that the fitted values that perform better for a given pass are given more weight than those that perform worse.

7. Output the averaged fitted values, the predictor values, and some measure of fit between the fitted values of the response and the actual values of the response.

In this procedure, the averaged fitted values and the predictor values characterize the response function. The response function is a way to represent how the predictors are associated with the response. The overall measure of fit conveys how well the response function corresponds to the data. But a good fit has no necessary implications for whether the response function is "correct." It is not even clear what "correct" means for this procedure. No data-generation process has been proposed, let alone a causal model. There is also no apparent role for statistical inference. We begin and end with the data on hand.

The procedure just outlined has a lot in common with "boosting" (Schapire, 1999), a statistical learning procedure that is considered later. An interesting feature is that with each iteration, observations that are more difficult to fit are given more weight. We show, however, that boosting has more in common with conventional statistical procedures than might first appear. There is a loss function being minimized, just as in conventional parametric regression. What can be novel is the particular loss function being used and the manner in which a very flexible fitting function is constructed.

In many situations, boosting fits the data well, and often substantially better than procedures that make only one pass through the data. This can mean that it will forecast more accurately too. By these criteria, boosting is likely to outperform conventional linear regression even if the linear regression is implemented in a stepwise manner. The superior performance can become more apparent as the sought-after response function becomes more nonlinear.

Figure 1.9 provides an example of some boosting output. The response variable is the number of homeless individuals in a census tract in log units. The single predictor is the percentage of the land in a census tract used for residential purposes. Plotted are the fitted values.

One might well have expected a negative relationship overall, but had a linear regression model been imposed, the conclusions would have been quite misleading. There is a precipitous drop in the number of homeless as the percentage of residential land use in a tract varies from about 5% to about 15%. On either side, the relationship is essentially flat. The pattern looks like a neighborhood tipping effect, which might not have been anticipated. It is very unlikely that the precise location of the transition would have been anticipated. This is just the kind of relationship with which statistical learning procedures can excel, and conventional regression can stumble.

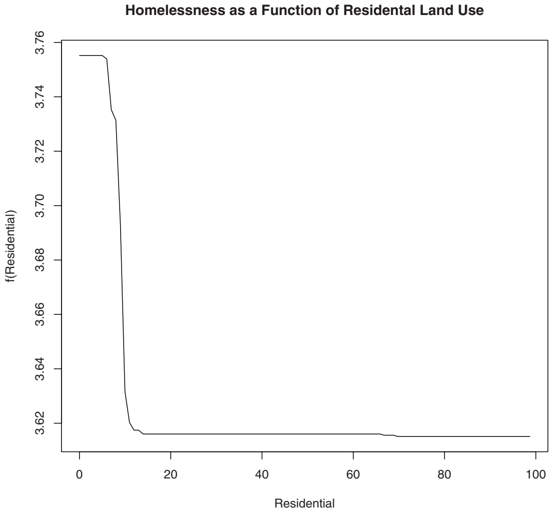

Fig. 1.9. Log of the number of homeless as a function of the percentage of land devoted to residential use.

Figure 1.9 was constructed from 100 passes through the data. A number of the other statistical learning procedures to be considered also make many passes through the data, but with very interesting twists and turns. For example, one can let each pass through the data be based on a random sample with replacement of the data on hand. One averages as before, but without any weighting. This is a first approximation of a procedure called “bagging” (Breiman, 1996).

The quick introduction to boosting and bagging no doubt looks very different from parametric regression. One puts one’s faith in a computer algorithm and pushes the run key. No data-generation process need be specified and so there seems to be no need for a statistical model. For these reasons, Breiman (2001b) has called such methods “algorithmic.” However, the break with conventional regression need not be that dramatic. Statistical learning, as the basis for the causal story or the conditional distribution story, relies on a specified data-generation process and a statistical model. It is the data summary story and the forecasting story that can be seen as “algorithmic” in the sense Breiman meant.

1.4 Some Initial Concepts and Definitions

Given the regression analysis framework, a wide variety of statistical learning procedures and approaches are examined. But, before going much farther down that road, a few definitions and concepts are necessary. They play a key role in the chapters ahead and, at this point, benefit from a brief introduction. We return to this content many times, so nothing like mastery is required now. And that's a good thing because some readers will find the content challenging, at least at a first reading.

1.4.1 Overall Goals

The procedures we examine have been described in many different ways (Sutton and Barto, 1999; Christianini and Shawe-Taylor, 2000; Witten and Frank, 2000; Hand et al., 2001; Hastie et al., 2001; Breiman 2001b; Dasu and Johnson, 2003), and associated with them are a variety of names: statistical learning, machine learning, reinforcement learning, algorithmic modeling, and others. "Statistical learning" as used in the pages that follow, is based on the following notions.

There may or may not be some data-generation process "out there" whose features we wish to learn about from the data. Such a construct is outside of the data and can help set the goals of a data analysis: what kind of conclusions are to be drawn from the results of the data analysis? A proposed data-generation process can also help provide a rationale for one statistical learning procedure rather than another. But much of the hands-on job of applying statistical learning to data proceeds in the same manner whether or not a data-generation mechanism has been proposed.

The earlier definition of regression analysis applies. Thus, for a quantitative response variable the goal is to examine $Y|X$ for a response Y and a set of predictors X. If the response variable is categorical, the goal is to examine $G|X$ for a response G and a set of predictors X. X may be categorical, quantitative, or a mix of the two. Consistent with common regression practice, the observed values of X are usually treated as fixed.

Many different features of $Y|X$ can be examined, but the conditional mean, $\bar{Y}|X$, is usually a key concern. This is the feature of $Y|X$ that has to date received the most attention. $\bar{Y}|X$ is sometimes interpreted as an expected value. For $G|X$, the conditional proportion is usually of interest for each of its K categories. $\bar{G}|X$ is sometimes interpreted as a conditional probability. In either case, $G|X$ is typically linked to response categories, often called "classes." The goal is to assign cases to classes. Then, the task can be called "classification," and the procedure employed can be called a "classifier."

1.4.2 Loss Functions and Related Concepts

In real applications, any efforts to fit the values of the response variable with one or more functions of the predictors will almost always be less than perfect. Indeed, if the fit is perfect, it is likely that some serious mistake has been made. Nevertheless, it often makes good sense to try to fit the response values as well as possible. This implies that a fitting criterion needs to be defined in order to characterize how good the fit is.

Fitting criteria are commonly called loss functions, cost functions, or objective functions. We use those terms interchangeably. The most common loss function for quantitative response variables is squared error loss, also called quadratic loss: $[Y - \hat{f}(X)]^2$. Recall that the mean is the central tendency measure that minimizes the sum of the squared deviations around itself. The fitted values that minimize squared error loss are the conditional means for each configuration of x-values. That is, compute $\bar{y}|X = x$ for each $X = x$.

If one wants to treat the data as a random sample for a well-defined population or as a random realization from a well-defined stochastic process, it follows that these conditional means are unbiased estimates and have other desirable formal properties. By a "well-defined" population, one means that it is possible to determine which units are in the population and which are not. For example, "all living adults in the United States" is not well defined until "adult" and "in the United States" are defined. A stochastic process is "well-defined" when it is thoroughly and precisely described. This will usually require one or more mathematical expressions. Thus, "a set of coin flips" by itself is not a sufficient definition. One would need to specify a particular binomial process (if that is the intent).

If concern is with bias, one can think of the conditional means as the "gold standard" under squared error loss. In practice, however, the gold standard can have some undesirable side effects. In particular, there may be for any particular $X = x$ no values of the response variable, or so few that the computed mean has a very large amount of sampling error.

There are a number of potential fixes. One is to compute the response variable conditional means not for $X = x$, but rather for X close by x. For example, if the predictors are age and education, rather than computing the conditional mean of income for, say, 28-year olds with four years of work experience, one might compute the conditional mean of income for all individuals between 25 and 27 who have between three and five years of work experience. For unbiasedness to be maintained, the true mean for the first set of individuals must be the same as the true mean for the second (and larger) set of individuals. "True" denotes the conditional mean in the population or the mean associated with the stochastic process responsible for the data. The goal is still to estimate the true mean income for individuals who are 28 and have four years of work experience, but information from nearby ages and years of work experience is being used.

In practice, one will rarely know if such assumptions are correct, but there may be evidence that they are close enough; the conditional means for the two groups are not likely to differ by enough to matter. For example, from past studies one many know that incomes usually change slowly as age and seniority change. Consequently, the bias that would be produced is negligible. Nearest neighbor methods, discussed later, build on this approach.

Computing different conditional means for different ranges of predictor values is the same as assuming the $f(X)$ is a step function. Alternatively, one might assume that the $f(X)$ is linear or some other smooth function of X. In effect, the data are pooled once again, but in a different manner so that more information is brought to bear on each conditional mean. Each conditional mean now can be computed using least squares regression, for example, with information from all of the observations in the dataset. For unbiasedness to be maintained, the true conditional means must fall on the assumed smooth function of the predictors. In this sense, the model for the conditional means is "right."

Squared error loss imposes a particular way of weighting the disparities between the response and the fitted values that needs to be considered carefully in the context of the data's properties and how the results of the data analysis will be used. Squaring makes large disparities especially influential in the fitting process, which can make the fitted values vulnerable to outliers (i.e., values that lie some distance from the mass of the data). In addition, the weights are symmetric; fitted values above the response are treated the same as fitted values below the response. Yet, symmetric weights are often inappropriate.

The tradition of resistant/robust estimation grew in part as a reaction to the problems caused by outliers under squared error loss. One option that we consider in a later chapter is linear loss: $|Y - \hat{f}(X)|$. Recall that the median is the central tendency measure that minimizes the sum of the absolute values of the deviations around itself. Thus, linear loss leads to computing the conditional median rather than the conditional mean. Then, the statistical issues that follow are much like those associated with the conditional mean.

The consequences of symmetric weighting are a very important issue to which we return later. We show that taking asymmetric costs into account can significantly change the fitted values. Decisions made from these fitted values can significantly change as well. Consider again, for example, the number of homeless in a census tract as the response variable, and predictors that are features of census tracts. Overestimating the number of homeless individuals in a census tract can have very different implications from underestimating the number of homeless individuals in a census tract. Yet, a symmetric loss function would assume that in the metric of costs their consequences are exactly the same.

Consider now a categorical response variable and a set of predictors. Just as for a quantitative response variable, one can proceed with a symmetric loss function. Suppose there are K distinct and mutually exclusive classes. Any misclassification—the fitted class is the wrong class—is given the same weight of 1.0. For example, the error of asserting that a high school student is a dropout when that student is not is given the same weight as asserting that a high school student is not a dropout when that student is. In both cases, the errors are given a value of 1.0. Correct classifications are given a value of 0.0. To minimize the sum of these errors over classes, it is clear that one can simply classify by the most common class. Because of the 0/1 coding of losses, using the sum of the squared errors gives the same result. In other terms that are used a lot later, the fitted class is determined by a "vote" in which the class with the plurality wins. When there are two classes, classification is by majority vote. Such procedures can be placed in a Bayesian decision theory framework (e.g., Bishop, 2006: 38–46) and are often called “Bayes classifiers.” The proportion misclassified by this approach is often called the “Bayes error rate.”

For example, given $X = x$ and a response of dropout or no dropout, if 65% of the students are not dropouts, all of the students for which $X = x$ are assumed to have not dropped out. Then, 35% of the students for which $X = x$ are misclassified. Suppose now that there are three response categories: drop out, no dropout, and moved to another school. Also suppose that 45% of the students fall in the no dropout category and that this is the largest percentage of the three for $X = x$. Then, all students for whom $X = x$ are assumed to have not dropped out and 55% are misclassified. In both illustrations, not dropping out is the fitted class. A similar rationale can be applied to each subset of observations defined by each unique configuration of x-values. Then it is possible to sum misclassifications over these unique values to obtain an overall proportion of cases misclassified.

Just as for quantitative response variables, one may use this reasoning to obtain useful estimators for parameters of a population or of a stochastic process. For example, if the data are a random sample from a well-defined population, each conditional proportion computed from the sample is an unbiased estimate of its population conditional proportion. That is, the proportion computed for each class is an estimate of the population proportion within that class. The term “Bayes risk” is sometimes applied to the expected value of the classification error when, in just such circumstances, the response variable is a random variable. A useful and accessible discussion from a more purely Bayesian perspective can be found in Ripley (1996: Section 2.1). A complementary discussion from a computer science perspective can be found in Bishop (2006: Section 1.5).

Finally, treating all classification errors as generating the same costs is often inappropriate, especially when real decisions will be made based on the classifications. As before, symmetric loss functions can be misleading. For example, the costs to the student and the school of failing to identify a student as a potential dropout may be very different from the costs to the student and the school of incorrectly identifying a student as a potential dropout. If these costs can be usefully approximated, it only makes sense to take them into account before actions are taken. We show later that building in the costs of classification errors can dramatically alter the classifications themselves.

1.4.3 Linear Estimators

Within the context of a regression analysis, consider a dataset with N observations. There is a single predictor X and a single value of X, x_0. Generalizations to more than one predictor are provided in a later chapter. The fitted value for $\hat{y}_0$ at x_0 can be written as

$$\hat{y}_0 = \sum_{j=1}^{N} \mathbf{S}_{0j} y_j. \tag{1.2}$$

S is an N by N matrix of fixed weights and is sometimes called a "smoother matrix." The subscript 0 denotes the row corresponding to the case whose fitted value of y is to be constructed. The subscript j denotes the column in which the weight is found. In other words, the fitted value $\hat{y}_0$ at x_0 is a linear combination of all N values of y_i, with the weights determined by $\mathbf{S}_{0j}$. In many applications, the weights decline with the distance from x_0. Sometimes the declines are abrupt, as in a step function. In practice, therefore, a substantial number of the values in $\mathbf{S}_{0j}$ can be zero.

If formal estimation of a conditional mean of the population is the goal, one has a linear estimator $\bar{y}|x$. It is a linear estimator because with **S** fixed, each value of y_i is multiplied by a constant before the y_i are added together; $\hat{y}_0$ is a linear combination of the y_i. Linear estimators play a central role in all of the chapters ahead. Linearity can make it easier to determine the formal properties of an estimator, and linear estimators and are often easier to understand. But one must be clear that even if an estimator is linear, the relationship between $\hat{y}$ and x can still be highly nonlinear, as we soon show.

$\mathbf{S}_{0j}$ has much in common with the hat matrix from conventional linear regression analysis. Recall that

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}. \tag{1.3}$$

The hat matrix **H** transforms the y_i in a linear fashion into $\hat{y}_i$. $\mathbf{S}_{0j}$ performs the same function, but can be constructed using more general procedures.

Consider the following cartoon illustration in matrix format. There are five observations constituting a time series. The goal is to compute a moving average of three observations going from the first observation to the last. In this case, the middle value is given twice the weight of values on either side. Endpoints are often a complication in such circumstances and here, the first and last observations are simply taken as is.

$$\begin{pmatrix} 1.0 & 0 & 0 & 0 & 0 \\ .25 & .50 & .25 & 0 & 0 \\ 0 & .25 & .50 & .25 & 0 \\ 0 & 0 & .25 & .50 & .25 \\ 0 & 0 & 0 & 0 & 1.0 \end{pmatrix} \begin{pmatrix} 3.0 \\ 5.0 \\ 6.0 \\ 9.0 \\ 10.0 \end{pmatrix} = \begin{pmatrix} 3.00 \\ 4.75 \\ 6.50 \\ 8.50 \\ 10.00 \end{pmatrix}. \tag{1.4}$$

The leftmost matrix is **S**. It is post multiplied by the vector **y** to yield the fitted values $\hat{\mathbf{y}}$. But from where do the values in $\mathbf{S}_{0j}$ come? If there are predictors, it only makes sense to try to use them. Consequently, $\mathbf{S}_{0j}$ is usually constructed from X.

1.4.4 Degrees of Freedom

Recall that, loosely speaking, the degrees of freedom associated with an estimate is the number of observations that are free to vary, given how the estimate is computed. Consider a variable with N observations. In the case of the mean, if one knows the values of $N-1$ of those observations, and one knows the value of the mean, the value of the remaining observation can be easily obtained. Given the mean, $N-1$ observations are free to vary. The remaining observation is not. So, there are $N-1$ degrees of freedom associated with the estimator of the mean.

This sort of reasoning carries over to many common statistics including those associated with parametric regression analysis. The number of degrees of freedom "used up" when the fitted values are computed is the number of regression parameters whose values need to be obtained (i.e., the intercept plus the regression coefficients). The degrees of freedom remaining, often called the "residual degrees of freedom," is the number of observations minus the number of these parameters. One of the interesting properties of the hat matrix is that the sum of its main diagonal elements (i.e., the trace) equals the number of regression parameters estimated. This is of little practical use with parametric regression because one can arrive at the same number by simply counting all of the regression coefficients and the intercept. However, the similarities between the **H** and **S** (Hastie et al., 2001: 129–130) mean that the trace of **S** can be interpreted as the degrees of freedom used up. Its value is sometimes called the "effective degrees of freedom" and can roughly be interpreted as the "equivalent number of parameters" (Ruppert et al., 2003: Section 3.13). That is, the trace of **S** can be thought of as capturing how much less the data are free to vary given the calculations represented in **S**. The residual degrees of freedom can then be computed by subtraction (see also Green and Silverman, 1994: Section 3.3.4).

There are other definitions of the degrees of freedom associated with a smoother matrix. In particular, Ruppert and his colleagues (2003: Section 3.14) favor

$$df_{\mathbf{S}} = 2\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{S}\mathbf{S}^T). \tag{1.5}$$

In practice, the two definitions of the smoother degrees of freedom will not often vary by a great deal, but whether the two definitions lead to different conclusions depends in part on how they are used. If used to compute an estimate of the residual variance, their difference can sometimes matter. If used to characterize the complexity of the fitting function, their differences are usually less important because one smoother is compared to another applying the same yardstick. The latter application is far more salient in subsequent discussions.

Beyond its relative simplicity, there seem to be interpretive reasons for favoring the first definition (Hastie et al., 2001: 130–133). Consequently, we use the trace of **S** as the smoother degrees of freedom. We show that the larger the value of the effective degrees of freedom, the more flexible is the fitting function and the more complex the fit. We also show that the effective degrees of freedom does not have to be an integer.

1.4.5 Model Evaluation

Just as in any fitting exercise, there needs to be a way to evaluate the quality of the fit. This evaluation is best done combining quantitative information obtained during the data analysis with subject matter expertise and policy concerns. A model that fails in subject matter or policy terms is of little use no matter how well it scores on statistical criteria. This might mean, for example, choosing a smoother fit than might be favored by a some statistical measure if the substantive implications are more easily understood and more consistent with existing subject matter knowledge.

But what kind of quantitative measure should be used? For the conventional linear regression, it is common to work with the mean squared error, $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, or some standardized version such the R^2. The R^2 is usually interpreted as the proportion of the variance of the response that is accounted for by the predictors. So, larger values of R^2 can be considered better than smaller values of R^2, and it follows that fitted values with a larger R^2 can be considered better than fitted values with a smaller.

The mean squared error and the R^2 are "resubstitution" fit statistics because the data used to evaluate the model is exactly the same as the data used to build the model. Such measures can convey unjustified optimism about the quality of the fit. In an effort to minimize the error sum of squares, the fitted values will respond as best they can to the data on hand, some features of which may be idiosyncratic. Then, the results will not generalize well to new samples from the same population. The unjustified optimism is exacerbated when the number of regression coefficients being estimated is large relative to the sample size because the fitting function gains flexibility relative to the amount of data. "Overfitting" is sometimes the term used to describe how unjustified optimism can be produced.

In response, a measure of fit can be used that attempts to adjust for the overfitting. A simple alternative to R^2 is R^2 adjusted for degrees of freedom:

$$\mathrm{AdjR}^2 = 1 - \left[(1-\mathrm{R}^2)\left(\frac{N-1}{N-p-1}\right)\right], \tag{1.6}$$

where N is the of number observations, and p is the number of parameters whose values are determined by the data. For a given number of observations, increasing the number of unknown parameters reduces the measure of fit. By "unknown," one means parameters whose values are to be determined by the data. Just as with its unadjusted cousin, bigger is better.

Taking the degrees of freedom into account leads to a conceptual improvement over the unadjusted R^2 because an effort is made to discount fit quality resulting solely from the complexity of the fitting function. However, the adjusted R^2 lacks much formal justification and is not easily generalized beyond a least squares context. The common use of terms such as "pseudo R^2" in such settings is telling.

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are alternatives to the adjusted R^2. They are based on far more rigorous statistical theory, which can be an important consideration. The AIC represents the relative amount of information lost (using the Kullbeck–Leibler information) when a given model, whose parameters are estimated from the data, is compared to the unknown but true processes that generated the data (Akaike 1973). Thus, the baseline is found in a conceptual entity beyond the data themselves. The smaller the AIC the better the approximation to the truth. The BIC is based on the Bayesian posterior probability of a given model (Schwartz, 1978; Raftery, 1995: Sections 4.1–4.2) compared to the "null model" with no predictors. There is again the construct of a true model serving as a target. A larger posterior probability implies that the model is more credible and that one has a better approximation of the truth. The BIC is smaller when the posterior probability is larger.

The AIC and the BIC can be properly applied to a much larger set of fitting procedures than the various kinds of R^{2}s. But as with the adjusted R^2, the AIC and BIC can be seen as altering the model's measure of fit by imposing a penalty for complexity. The penalty for the AIC and BIC increases with the number of unknown parameters in the model and decreases with the sample size. In other words, the penalty is larger when the number of parameters increases relative to the number of observations.

The AIC can be written in a number of ways, but one common expression is

$$\text{AIC} = \log\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right] + \frac{2p}{N}. \tag{1.7}$$

Likewise, the BIC can be written in a number of ways. A common expression is

$$\text{BIC} = \log\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right] + \frac{[\log(N)]p}{N}. \tag{1.8}$$

The BIC imposes a heavier penalty for the number of regression parameters. But neither is easy to interpret without some baseline or other means of comparison. We return to this issue shortly.

The AIC and BIC can be generalized so that in principle they are useful fit statistics for statistical learning procedures. In particular, it is common to replace p with $\text{tr}(\mathbf{S})$. However, both measures can prove to be inadequate when it is not apparent how to quantify the effective degrees of freedom.

For example, statistical learning procedures are often applied several times to the data with one or more tuning parameters varied. The AIC may be computed for each. But each AIC is ignorant about the information obtained from prior fitting attempts and how many degrees of freedom were expended in the process. Matters are even more complicated if some of the variables are transformed or recoded after examining descriptive statistics before the fitting begins. Often, the effective degrees of freedom used in the AIC and BIC will be too few. Some unjustified optimism remains.

As an alternative, it can be useful to think about the various measures of fit as efforts to characterize how well the fitting function forecasts. Then, one popular definition of fit is the expected prediction error (also called "expected forecasting error," or "expected generalization error")

$$\text{PRE} = E[(Y - f(X))^2]. \tag{1.9}$$

PRE is the mean squared error in the population from which the data were sampled or over limitless independent realizations of the stochastic process that generated the data. An alternative definition is the expectation of the sum of the absolute values of $(Y - f(X))$. How does one go about estimating such quantities?

An excellent option is to work with two (or more) random samples from the same population. One sample is treated as the "training sample" and the other sample is treated as the "test sample." A fitting function built from the training sample is applied to the test sample, and a measure of prediction error computed. That is, data from the test sample are used with the fitting function from the training sample to produce fitted values. These are paired with the observed values in the test sample when, for example, the mean squared error is calculated.

A lot also can be learned by unpacking the overall measure of test sample prediction error. It can be instructive to learn which observations are being underestimated and which are being overestimated, and by how much. For example, if the response variable is the number of homeless individuals in a census tract, it would be important to know if the fitted values tend to substantially underestimate homeless counts in census tracts where a large number of homeless individuals are likely to be found. Or the problems could be spatial; predictions for census tracts in one part of a metropolitan area may in general be less accurate. Such information can provide a more sensitive evaluation of the fitted values and sometimes suggest ways the fitting function might be improved.

Often there is no test sample. Under these circumstances, there are interesting alternatives that nevertheless draw directly on the idea of a training sample and a test sample. "Drop-one" cross-validation is a popular example. Drop-one cross-validation is also called "leave-one-out" cross-validation, "jackknife" cross-validaton, and "N-fold" cross-validation.

Imagine a statistical learning procedure that is applied N times to the data. Each observation in turn is dropped from the dataset and its fitted value computed. That is, each fitting is based on $N-1$ observations, from which the fitted value of the dropped observation can be computed. The mean squared error computed from the dropped values and their corresponding fitted values is a cross-validation measure of fit and an estimate of prediction error.

More formally,

$$CV = \sum_{i=1}^{N} [y_i - \hat{f}_i^{-i}(\mathbf{X}_i)]^2, \tag{1.10}$$

where the superscript $-i$ signifies that observation i has been dropped. One nice feature of cross-validation is that the effective degrees of freedom does not enter explicitly into the calculations. Another nice feature is that it has a PRE interpretation.

The Generalized Cross-Validation statistic (GCV) can be a handy cross-validation approximation. It can be applied to the existing data as a whole and is easily computed as

$$GCV = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{y_i - \hat{f}_i(\mathbf{X}_i)}{1 - \mathrm{tr}(\mathbf{S})/N} \right]^2. \tag{1.11}$$

The GCV requires the user to decide which definition of the effective degrees of freedom is to be used. As shown, the trace of the smoother matrix is the usual choice.

Another approach in the same spirit exploits a bootstrap procedure (Efron, 1983; Efron and Tibshirani, 1993: Section 17.7). In its simplest form, one takes B samples of size N, with replacement, from the data. Each random sample serves as a training sample so that for each, a fitting function is constructed. The original (unsampled) data serve as a test sample. One can then compute a mean square error B times using the fitting function from each of the training samples and the data from the test sample. Averaging over the B samples provides an estimate of prediction error.

The simple bootstrap estimate of prediction error is not entirely satisfactory. Each bootstrap sample is drawn from the original sample; the original sample is not really a pure test sample. The overlap leads to estimates of prediction error that are too small.

A better approach is to borrow some ideas from cross-validation. For any given bootstrap of size N, about a third of the observations will by chance not be selected. These observations can serve as a test sample when estimates of prediction error are computed. With a sufficient number of bootstrap samples, any given observation will likely fall in a test sample several times. The average mean squared error for each observation in its test samples provides an observation-specific estimate of prediction error. Averaging these over the N observations leads to an overall estimate of prediction error based on the "drop-one" bootstrap.

However, because each bootstrap sample will on the average contain only about two-thirds of the unique observations of the original sample, the data used to construct the fitting function will be more sparse than the full dataset. If $f(X)$ is complex, some of its features will be missed. Bias results. Consequently, the estimate of prediction error will be inflated. A correction can be introduced that leads to the following expression for the estimated prediction error.

$$\widehat{\text{PRE}} = .368(\text{MSE}_R) + .632(\text{MSE}_B), \tag{1.12}$$

where MSE_R is the resubstitution mean squared error from the original sample, and MSE_B is the mean squared error from the drop-one bootstrap. Further improvements are possible (Hastie et al., 2001: 219–220).

Finally it is also possible, with small modifications, to treat the AIC and BIC as estimates of prediction error. Thus,

$$\text{AIC} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \frac{2p\sigma^2}{N}. \tag{1.13}$$

In this form, the AIC is known as the C_p statistic and can provide an unbiased estimate of the prediction error (Efron and Tibshirani, 1993: 242). The BIC is now computed as

$$\text{BIC} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \frac{[\log(N)]\sigma^2 p}{N}, \tag{1.14}$$

and can provide a consistent estimate of the prediction error (Efron and Tibshirani, 1993: 242).

In both equations, p may be replaced by $\text{tr}(\mathbf{S})$, and an estimate of σ^2 is required. Estimates of σ^2 are usually constructed from the observed values of the response and the $\hat{f}(X)$. One proceeds as if the $\hat{f}(X)$ has negligible bias and the disturbances behave as the model $Y = F(X) + \varepsilon$ requires. Moreover, the model is specified independently of the data being analyzed; there can be no data snooping (Luo et al., 2006: 165–167). If these requirements are substantially violated, $\hat{\sigma}^2$ can be substantially biased so that the estimates of prediction error will be substantially biased as well.

These implications of data snooping can generalize. All of the methods to properly represent prediction error can provide overly optimistic results if the impact of data snooping is not taken into account. For example, if drop-one methods are employed with data already altered using information from earlier models, there is no way for the procedures used to estimate prediction error to know about the earlier snooping. Falsely small estimates can result. Or as we show later, the summary statistics just described can be used to tune a model. This can be a useful way to undertake the inductive development of a model. But, it is very easy to slip into "overtuning." For example, neither the AIC or BIC can know about models previously considered and rejected. Likewise, cross-validation depends on random splits of the data into training and test samples whose independence, such as it is, can be rapidly compromised when the data are used repeatedly in the same modeling enterprise. Falsely optimistic measures of model performance can result.

In summary, there are many choices available for obtaining more honest measures of fit. The resampling methods yield estimates having an intuitive appeal, but can be computationally taxing, especially in many statistical learning applications. The adjustments to the resubstitution mean squared error are far easier to compute, but require a credible value for the effective degrees of freedom and for some, a credible estimate of σ^2. There also can be important differences in performance for given samples. Finally, there remain a number of unresolved issues even for the linear estimators emphasized here; nonlinear estimators are more demanding still (Efron, 2004). Therefore, there seems to be no definitive guidance on which measures to use and in practice, those that can be computed easily seem to dominate. The best course, when feasible, is to have a training sample and a true test sample. A key asset is that one can use as test data the dataset as it was before any data snooping was undertaken. A more honest assessment of overfitting can result, especially if the fit measures are not overused.

1.4.6 Model Selection

It is a relatively small step from model evaluation to model selection. A model that performs better is chosen over a model that performs worse. In parametric regression, model selection usually means deciding by some quantitative yardstick which predictors, and transformation thereof, belong in the model. One way to think about this process is that among all of the available regressors and/or their transformations that could be included, some have their regression coefficients set to zero. Regressors with such regression coefficients are, by definition, excluded from the model.

With parametric regression, the model selection process can be undertaken in at least three ways. First, hypothesis tests may be used when the candidate models are nested within a single, all-encompassing model (Cook and Weisberg, 1999: 266–272). The usual null hypothesis is defined by one or more constraints on the regression coefficients. The most common constraints require that a subset of the regression coefficients in the all-encompassing model be equal to zero. The smaller model is assumed to be the correct model unless the null hypothesis is rejected. Likelihood ratio and f-tests are popular.

Second, one may apply various regression diagnostics from which model flaws can sometimes be identified. After that, remedies can be sought. For example, plots of a model's residuals against its fitted values can reveal nonlinearities in the data that have been overlooked. Likewise, added variable plots (Cook and Weisberg, 1999: Section 10.5) can point to possible omitted variables and/or needed transformations.

Third, one may compare various candidate models by some goodness-of-fit measure, and choose the one that fits the data best. As just discussed, this is best done discounting fit quality for the number of unknown regression parameters. The AIC and BIC are common choices.

The AIC is one of a class of "asymptotically efficient" model selection tools (Hurvich and Tsai, 1989) that tries to find the model that loses the least information relative to how the data were actually generated. No claims are necessarily made, however, that the model selected is "correct" or even that a correct model is in the set of models examined. The hope is that the model selected is a useful approximation of the "truth."

The BIC is one of a class of consistent model selection tools that in principle will find the correct predictors, if they are included within the set of models examined, by determining the proper dimension of the regressor matrix. It is this dimension that is consistently estimated. Suppose there is a set of predictors in the data set placed in any arbitrary order. Some of the predictors are unrelated to the response. Starting at the top of the regressor ordering, the BIC estimates how many of these predictors should be included. If there are, say, 15 ordered predictors, but none of the predictors after the first 9 are related to the response, the selected models should contain the first 9 regressors. Some of these may not be related to the response. But in principle, all of the predictors that belong in the true model are selected and the false positives make little difference because they are not associated with the response.

If the data analyst believes that the true model is within the set to be examined, some argue that the BIC can be a better model selection tool. If the data analyst does not believe that the true model is within the set to be examined, some argue that the AIC can be a better model selection tool. However, there are other versions of the AIC which are arguably superior to the BIC when the correct model is among those that could be constructed from the dataset (Simonoff and Tsai, 1999). At this point, there seems to be no clear consensus on which selection criterion is best. For the kinds of procedures considered in this book, the idea that there is a correct model to be found in the data seems somewhat anachronistic. Other approaches to model selection are considered in later chapters.

For all three approaches to model selection, model parsimony is also important. Simpler models are preferred, other things equal, and sometimes when they are not equal. Simpler models can be more stable to modest changes in the model, otherwise inconsequential differences in how the variables are measured, and random perturbations of the data, such as occur under random sampling. Simpler models may also be more easily interpreted. This can mean, for example, that a model selected using the AIC may not be simple enough despite adjustments for the effective number of parameters.

Sometimes model selection procedures are usefully automated. The canonical illustration is conventional stepwise regression. In the forward selection case, among all the candidate variables, the one with the largest correlation with the response is included in the model first. The slope and intercept are estimated. Among all of the remaining candidate regressors, the one with the largest partial correlation (conditioning on the regressor already in the model) is included. The values of the two slopes and intercept are then computed. This process is continued until no new candidate regressor improves the model fit sufficiently. Other criteria can be used to decide which variables to include, such as hypothesis tests or a measure of fit as with the ones just discussed. A key point is that after each predictor is introduced into the model, the values of all of the model's parameters are recomputed. This is different from stagewise regression, discussed in later chapters, in which the values of regression coefficients computed in one step are not recomputed when a new predictor is added to the model.

Additional examples of automated model selection include backward selection stepwise regression in which less important predictors are eliminated one by one from the all-encompassing model, and all-subsets regression in which all possible submodels are compared. Among the risks of automation are that subject matter information is neglected so that the resulting models are not informative.

Both stepwise regression and stagewise regression are examples of "greedy algorithms," which figure significantly in later chapters. They are greedy because at each step or stage the single best predictor is selected but then not reconsidered later. As a result, overall optimization can be sacrificed to the local optimization undertaken at each step or stage; the long term can be jeopardized by short-term thinking. Yet, greedy algorithms are often practical and effective. The alternative of searching over all possible models can become computationally taxing, or even intractable, if there are a large number of predictors.

Whether automated or not, all model selection procedures can risk serious overfitting. For example, t-tests applied to later models do not take into account the t-tests applied to earlier models. One result can be spuriously small p-values. Then, predictors may be retained when they should be removed; a null hypothesis is falsely rejected. An unnecessarily complicated model can result, which may not generalize well. The same concerns apply to the various measures of fit just discussed, especially when they are used repeatedly. At a deeper level, all statistical inference can be seriously jeopardized when the same data are used to inductively build a model and then to estimate the model's final set of parameters. More is said about this in the next chapter.

Despite the many unresolved problems with model selection for parametric regression, we soon show that, broadly speaking, the same three model selection strategies are commonly applied in statistical learning. There are also many opportunities for automation. At the same time, however, model selection needs to be placed in a broader context.

With statistical learning, models may be characterized not just by constraints placed directly on individual regression coefficients one by one, but by constraints placed on the objective function being minimized. For example, penalties for model complexity, in much the same spirit as those used in the AIC and BIC, can be imposed when a least squares fitting criterion is applied. One goal can be to reduce the risks of overfitting in the model itself, not just alter the measure of fit. This is a theme that surfaces a number of times in later chapters. The resulting model can have regression coefficients that are generally smaller in absolute value (but not necessarily zero) than they would have been had the complexity penalty not been imposed.

If an important goal of a statistical learning procedure is to arrive at a model in which subject matter conclusions depend on the particular regressors included, the model-building process should be able to force unnecessary regression coefficients to be zero. Then, model selection becomes regressor selection. If instead, interest centers more on the fitted values, there is no requirement that in the model building any regression coefficients should be zero. For example, a useful balance between bias and variance may require altering each of the regression coefficients a bit without trying to force any of them to zero. There is no intent to weed out any regressors. In short, what makes one model better than another depends in part on which model outputs are more important.

1.4.7 Basis Functions

Basis functions play a key role in all of the statistical learning procedures discussed. Basis functions are transformations of predictors that can allow for a more flexible fitting function by increasing the dimensionality of the regressor matrix. A set of p predictors becomes a set of predictors greater than p. This can allow the fitted values to be more responsive to the data.

Consider first the case when there is but a single predictor. X contains two columns, one column with the values of that single predictor and one column solely of 1s for the intercept. The $N \times 2$ matrix is sometimes called the "basis" of a bivariate regression model. This basis can be expanded if one allows transformations of X. A very powerful and flexible set of transformations can be written as

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X). \tag{1.15}$$

There are M transformations of X, which can include the untransformed predictor and a column of 1s (allowing for a y-intercept). β_m is the weight given to the mth transformation, and $h_m(X)$ is the mth transformation of X. Consequently, $f(X)$ is a linear combination of transformed values of X. The right-hand side is sometimes called a "linear basis expansion" of X.

One common transformation employs polynomial terms such as 1, x, x^2, x^3. Then, Equation 1.15 takes the form

$$f(X) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3. \tag{1.16}$$

When least squares is applied, a conventional hat matrix follows from which fitted values may be constructed.

Another popular option is to construct a set of indicator variables. For example, one might have predictor z, transformed in the following manner.

$$f(Z) = \beta_0 + \beta_1(I[z > 5]) + \beta_2(I[z > 8|z > 5]) + \beta_3(I[z < 2]). \tag{1.17}$$

As before, fitting by least squares leads to a conventional hat matrix from which the fitted values may be constructed.

Equation 1.15 can be generalized so that $p > 1$ predictors may be included:

$$f(X) = \sum_{j=1}^{p} \sum_{m=1}^{M_j} \beta_{jm} h_{jm}(X). \tag{1.18}$$

Each predictor has its own set of transformations. Then, all of the transformations for all predictors, each with its own weight β_{jm}, are combined in a linear fashion. For example, one could combine Equations 1.16 and 1.17 with both X and Z as predictors.

Why use the additive formulation when there is more than one predictor? With each additional predictor, the number of observations needed can increase enormously; the volume to be filled with data goes up as a function of the power of the number of predictor dimensions. This is what lies behind the "curse of dimensionality." One important result can be data that are too sparse for the intended analysis. In addition, there can be very taxing computational demands. So, it is often necessary to restrict the class of functions of X examined. One hopes that the response variable will be fitted sufficiently well by a model that is less flexible than what one ideally might like. We show one manner in which this plays out when smoothers are discussed in the next chapter.

Equation 1.18 has the additional benefit of retaining some of the same look and feel as conventional linear regression. This can lead to simplifications in the underlying mathematics, more effective computer algorithms, and more transparent interpretations. We show soon that Equation 1.18 leads to surprisingly flexible and effective fitting procedures in part because many complex functions can be well approximated by low-order polynomials and other relatively simple transformations.

But if Equation 1.18 is essentially multiple regression, where is the statistical learning? The answer will at this point probably seem a bit perplexing. The parametric structure of each basis function by itself can lead to a nonparametric fitting function when all of the pieces are used at once. And how these pieces are used can be substantially determined in an inductive manner by the data. It is a bit like constructing a highly nonlinear function from a large number of very small line segments connected end to end.

An Illustration

Consider, for example, how Equation 1.16 might be used within the following algorithm for a single response variable and a single predictor.

1. Select a target value of y, y_0, with its associated x_0.
2. Find the 20% of the observations whose values on x place them nearest to x_0.

3. Estimate for that subset of observations a cubic regression equation, weighting each of these observations so their weights decline linearly with distance from x_0.
4. From the results, construct the predicted value of y_0, $\hat{y}_0$.
5. Repeat Steps 1–4 for each unique value of x.

The result is a set of fitted values that can be called a "smoothed" version of y. The procedure goes under various names such as "locally weighted regression" (Cleveland, 1979). As we show later, there are often better ways to do locally weighted regression, but locally weighted regression can be written within a basis function framework.

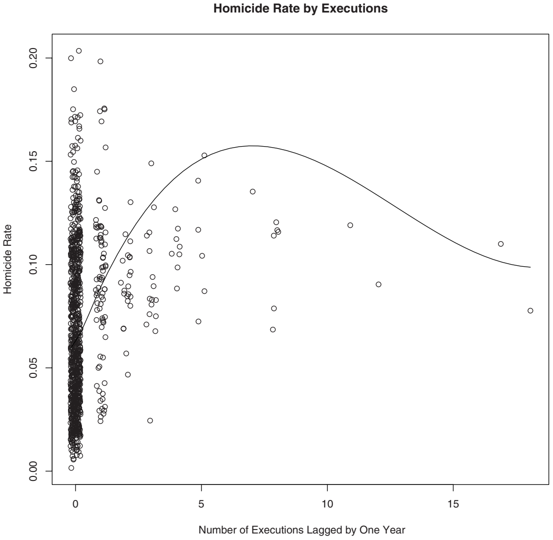

Fig. 1.10. The homicide rate per 1,000 as a function of the number of executions

Figure 1.10 provides an illustration. The data include the set of all 50 states each year from 1978 to 1998, for a total of 1000 observations. Each year, the homicide rate and the number of executions for capital crimes are recorded. Data such as these have been central in a recent debate about the deterrent value of the death penalty (Berk, 2005a).

In Figure 1.10, executions lagged by one year is on the horizontal axis, and the homicide rate per 1000 people is on the vertical axis. There are 1000 observations: 50 states times 20 years. Consequently, an observation is a state-year. To make the scatterplot more comprehensible, the number of executions has been jittered. But in most years, most states execute no one. Over 80% of the observations have zero executions. A very few states in a very few years execute more than five individuals. Years in which more than five individuals in a state are executed represent about 1% of the data (i.e., 11 observations out of 1000).

The fitted values for a locally weighted regression are overlaid, constructed from the original (unjittered) data. One can see that for five executions or less, the relationship between the number of executions and the homicide rate one year later is positive. More executions are followed one year later by more homicides. Thus, there is a positive relationship for 99% of the data. When a given state in a given year executes six or more individuals, the relationship turns negative. With more executions, there are fewer homicides one year later. But one can see that there are almost no data supporting this relationship and in fact, a proper confidence interval around that portion of the curve would show that the true curve could easily be flat and even positive. In short, for 99% of the data the relationship is positive and for the atypical 1%, one really cannot tell. (For more details, see Berk, 2005a.)

Figure 1.10 represents a descriptive exercise. No data-generation mechanism was proposed, let alone a causal model. The goal was to provide an instructive visual summary of how the homicide rate is related to the number of executions one year earlier. A key point is that no functional form was imposed on the response function despite the application of parametric cubic regression in many local regions of the data. Had a single, parametric, linear relationship been imposed on the data, a misleading negative slope would have materialized. The few observations on the far right side of the plot are highly influential.

As an alternative fitting approach, consider a step function. For Equation 1.17, the choices of where to segment z can also be determined empirically. We consider in some detail later how this is done, but the basic idea is to choose each break point so that the fit is improved the greatest amount possible. For example, an initial break point at $z = 5$ might be found by trying all possible break points and choosing the one that reduced the error sum of squares the most. Then, within the two subsets of observations ($z > 5$ and $z \leq 5$), the next two break points would be independently chosen with the same goal in mind: to reduce the error sum of squares the largest amount possible. Classification and Regression Trees (Breiman et al., 1984) is based on this general idea and figure significantly in material presented in later chapters. Classification and regression trees can also be written as a set of basis functions that produce a nonparametric fit.

What would be the result if the indicator variable approach were applied to the data in Figure 1.10? The result is a single break point at $z = 1$. In effect, a split is made between no executions and one or more executions. The conditional mean when there are no executions is .066. The conditional mean when there is one execution or more is .094. When there is one or more executions, the homicide rate per 1000 people is about 50% greater the following year. An overlay of these fitted values on Figure 1.10 would reveal a step function. There would be a horizontal line between 0 and 1 at a value of .066, and another horizontal line between 1 and the 18 (the largest value for the number of executions) at a value of .094. A vertical line at 1 would connect the two.

One important implication of the step function is that the fit is not improved a meaningful amount by segmenting the data any further. It is not useful to distinguish between, say, six executions and ten executions. This underscores the earlier point that data beyond five executions are far too sparse to be of much use.

1.5 Some Common Themes

Although the variety of procedures that are discussed later can differ greatly in look and feel, and although they often come from rather disparate intellectual traditions, there are several themes running through the material. These are usefully flagged before getting into lots of details.

- *Goals*—The goals of the procedures discussed are usually some mix of description, classification, and forecasting. Sometimes, the intent is to characterize a data-generation mechanism. But the techniques used in statistical learning are not usually considered causal models, and causal inference is rarely a goal. Likewise, statistical inference is not usually considered a key activity, sometimes because the data do not justify it, sometimes because a data generation mechanism has not been clearly articulated, and sometimes because credible procedures for statistical inference have not be developed.
- *Forecasting Skill*—In much of the statistical learning literature on which we rely, the true test of a procedure is not how well it fits the data on hand, but how well it forecasts. Forecasting skill is the gold standard. We show that although forecasting skill is widely accepted as the key performance criterion, how forecasting skill is defined and operationalized can vary substantially.
- *The Bias–Variance Tradeoff*—The bias-variance tradeoff is very visible for all of the procedures discussed. The basic point is that more flexible fitting functions will usually fit the data better but will typically generate less stable results. A better fit can imply less bias but more variance. Insofar as a true mean function exists, the fitted values have less systematic error. However, the fitted values for a new random sample from the same population will differ more from the original set of fitted values. Conversely, less flexible fitting functions will usually have more systematic error, but typically generate more stable results. A worse fit can imply more bias but less variance. Insofar as a true mean function exists, the fitted values have more systematic error. However, the fitted values for a new random sample from the same population will differ less from the original set of fitted values. A key goal in statistical learning can be to strike a useful balance between the bias and the variance or better still, find a way around

it. We also show that different statistical learning procedures can address the bias-variance tradeoff in different ways.

- *Loss Functions*—All of the statistical learning procedures discussed try to fit the data taking the disparities between the fitted values and the actual values into account. These disparities are "losses" that need to be weighted, aggregated, and minimized in some sensible fashion as the fitted values are computed. There are very important differences in how the disparities are handled from one statistical learning procedure to another.
- *Overfitting*—A major difficulty in statistical learning is overfitting. Very flexible fitting procedures will tend to respond to idiosyncratic features of the data, producing results that do not generalize well to new data. The results tend to be dataset-specific. The results can elicit very creative subject matter interpretations that, unfortunately, are stories about the noise not the signal.
- *Tuning*—For all of the statistical learning procedures examined, there are choices to be made about "tuning parameters." These are not population parameters of subject matter or statistical interest. They are parameters, much like dials on a machine, that determine how a procedure functions. There is often little statistical theory to guide the selection of tuning parameter values. Consequently, data analysts will proceed by craft lore or trial and error. There is nothing inherently wrong with such practices, but there will commonly be a strong temptation to try a large number of different tuning parameter settings. This can lead to "overtuning" and hence, overfitting. Sometimes the tuning is done using one or more measures of model performance, such as goodness-of-fit. Here, too, overuse can lead to overfitting even when the fit statistic has been designed to counter certain causes of overfitting.
- *Interpretability*—Within a regression framework, results that are difficult to interpret in subject matter terms, no matter how good the fit, are often of little use. This will sometimes lead to another kind of tradeoff. Fitting functions that perform very well by various technical criteria may stumble when the time comes to understand what the results mean. Important features of the data may be lost. It will sometimes be useful, therefore, to relax the technical performance criteria a bit in order to get results that make sense.
- *Differences That Make No Difference*—In almost every issue of journals that publish work on statistical learning and related procedures, there will be articles offering some new wrinkle on existing techniques, or even new procedures, often with strong claims about superior performance compared to some number of other approaches. Such claims are often data-specific but even if broadly true, rarely translate into important implications for practice. Often the claims of improved performance are small by any standard. Some claims of improved performance are unimportant for the subject matter problem being tackled. But even when the improvements seem

to be legitimately substantial, they often address secondary concerns. In short, the newest is not necessarily the best.

- *Rapid Development*—The concepts, understandings, and tools that fall under the rubric of statistical learning are evolving rapidly. It is very difficult to keep up with the field, and today's breakthrough can be tomorrow's bust. Moreover, the pace of technical development has to date vastly outstripped the pace at which hands-on experience with real data has accumulated. As a result, it is very difficult to provide grounded advice to data analysts working on real scientific and policy problems. There is so far relatively little data analysis lore for many of the newer tools. In addition, most popular software packages are several years behind the curve. Researchers who want to use the most recent advances either have to work in the software environment where the tools are being developed (e.g., R or Matlab), or in special-purpose proprietary packages such as the one available from Salford Systems (http://www.salford-systems.com/).
- *Data Quality Really Matters*—Just as in any form of regression analysis, good data are a necessary prerequisite. If there are no useful predictors, if the data are sparse, if key variables are highly skewed or unbalanced, or if the key variables are poorly measured, it is very unlikely that the choice of one among several statistical learning procedures will be very important. The problems are bigger than that. It is rare indeed when even the most sophisticated and powerful statistical learning procedures can overcome the liabilities of bad data.

1.6 Summary and Conclusions

The statistical learning emphasized in this book is treated as a form of regression analysis, broadly defined, with no necessary commitment a priori to any particular functional relationship between predictors and the response. The relationships between the predictors and the response are substantially determined from the data. The stance taken is within the spirit of procedures such as stepwise regression, but beyond allowing the data to determine which predictors are useful, the data are allowed to help determine what predictor functions are most appropriate. In practice, this means subcontracting a large part of the data analysis to one or more computer algorithms.

What role can subject matter "theory" have? Subject matter theory can be very important in

1. Framing the empirical questions to be addressed
2. Defining a data-generation mechanism
3. Designing and implementing the data collection
4. Determining which variables in the dataset are to be inputs and which are to be outputs
5. Settling on the values of tuning parameters

6. Deciding which results make sense

But none of these activities is necessarily formal or deductive, and they leave lots of room for interpretation. If the truth be told, subject matter theory plays much the same role in statistical learning as it does in most conventional analyses. But in statistical learning, there is often far less posturing.

At least as important as subject matter theory is information on how the statistical learning results are to be used. Central in these discussions is the concept of a loss function, which determines the costs of inaccurate fitted values. There is really no way to avoid a consideration of loss functions for any statistical learning approach (and for most other statistical procedures as well). A loss function typically is assumed, even if it is not explicitly acknowledged by the data analyst. And the loss function employed can have an enormous impact on the results. Subject matter expertise is necessarily brought to bear when loss functions are selected, and the practical applied decisions to be made also can play a key role.

Finally, the nature of exploratory data analysis needs to be briefly revisited to put both "data snooping" and statistical learning in their proper places. Data snooping is under-the-table exploratory data analysis. The data are studied at great length, various transformations are undertaken, a large number of statistical procedures are applied and then, only the best results are reported. This is a common ruse in most sciences that creates technical problems (e.g., invalidating statistical tests) and misleading results. An exploratory data analysis is presented as if it were a confirmatory data analysis.

A form of data snooping can also be undertaken with more apparent legitimacy when systematic model selection procedures are employed. If data used for model selection are also used to construct and evaluate the chosen model, data snooping is again at work. The procedures used may be forthrightly described, but as a formal matter, statistical inference can be undermined. The usual regression statistical inference is undertaken conditional upon a model known before the data are examined. There is a small but instructive literature showing that the unconditional distribution of the post model-selection estimator cannot be arrived at with sufficiently useful accuracy, even asymptotically (Freedman et al., 1988; Danilov and Magnus, 2004; Leeb and Pötscher, 2006). So, the usual conditional tests do not address the question that data analysts typically want to answer, and the unconditional tests can be very problematic. We return to this later. For now, the point is that model selection procedures share a lot with conventional exploratory data analysis.

Statistical learning is usually exploratory as well. But at least in principle, the exploration is undertaken by a set of very explicit rules represented in the algorithms employed. No one is hiding the ball. Equally important, we show later that great pains are taken to avoid the seductions of overfitting. There is often a training dataset to which the procedures are applied and a test dataset to determine whether the results are too good to be true. When there are no test data, there can be clever resampling techniques or helpful adjustments to measures of fit that can sometimes achieve much the same result. Rather than reporting only the best results, a conscious effort is made to report only the honest results.

Despite these good intentions, statistical learning practice can be subverted by data snooping. In particular, a procedure that by itself does not overfit can be applied to the data many times. After each pass through the data, the results are examined, and the statistical learning algorithm is tuned in the hope of producing better results. This sequence of fitting attempts can lead to overfitting despite protections against overfitting built into the algorithm when it is applied just once. The difficulties are compounded when the tuning process goes unreported. Sensitivity to overfitting is an important strength of statistical learning. But that sensitivity does not confer immunity.

Exercises

The purpose of these exercises is to provide a bit of practice doing regression analyses by examining conditional distributions without the aid of conventional linear regression. You will see that regression analysis does not require a parametric model.

Problem Set 1

Load the R dataset "airquality" using data(airquality). Learn about the data set using help(airquality). Attach the dataset "airquality" using attach(airquality). If you do not have access to R, or choose to work with other software, exercises in the same spirit can be easily undertaken. Likewise, exercises in the same spirit can be easily undertaken with other data sets.

1. Using pairs(), construct of a scatterplot matrix for all of the variables except for "Month" and "Day." Describe the relationships between each pair of variables.

2.

3. boxplot Using boxplot(), construct side-by-side boxplots for ozone concentrations against month and ozone concentrations against day. Does the ozone distribution vary by month of the year and day of the month? In what ways?

4. What would one have to assume to use month or day of the month in scatterplots?

5. Construct a three-dimensional scatterplot with ozone concentrations as the response and temperature and wind speed as predictors. (Maybe use cloud() from the lattice package.) What patterns can you make out? It

is difficult to see much. Other kinds of plots for three variables are often more useful.

6. Construct a conditioning plot using coplot() with ozone concentrations as the response, temperature as a predictor, and wind speed as a conditioning variable. How does the conditioning plot attempt to hold wind speed constant?

7. Consider all the conditioning scatterplots. What common patterns do you see? What does this tell you about how ozone concentrations are related to temperature with wind speed held constant?

8. How do the patterns differ across the conditioning scatter plots? What does that tell you about how wind is related to ozone concentration holding temperature constant? What does that tell you about how the relationship between ozone concentrations and temperature can differ for different wind speeds?

9. Construct an indicator variable for missing data. Using table() or xtab(), cross-tabulate the indicator against month. What do you learn about the pattern of missing data? How might your earlier analyses using the conditioning plot be affected?

10. Write out the parametric regression model that seems to be most consistent with what you have learned from the conditioning plot. Try to justify all of the assumptions you are imposing.

11. Implement your regression model in R using lm() and examine the results. How do your conclusions about the correlates of ozone concentrations learned from the regression model compare to the conclusions about the correlates of ozone concentrations learned from the conditioning plot?

Problem Set 2

The purpose of this exercise is to give you some understanding about how the complexity of a fitting function affects the results of a regression analysis and whether popular measures of fit compensate sensibly.

1. Construct the data as follows. For your predictor: x = rep(1:20, times = 10). This will give you 200 observations with values 1 through 20. For your response: y = rnorm(200). This will give you 200 random draws from the standard normal distribution.

2. Plot the response against the predictor and describe what you see.

3. Apply a bivariate regression using lm() and then glm(). Describe what overall conclusions you draw from the two sets of output. (The fit should be the same but the output from the two procedures are a bit different.)

4. Repeat the two bivariate regressions with the predictor as a factor. Use the same R code as before but use as.factor(x) instead of x.

5. How do the two sets of output differ from the previous sets? Focus on the overall measures of fit. Do the adjustments for the degrees of freedom used up seem to be effective in this case?

Problem Set 3

This purpose of this exercise is to explore how degrees of freedom used up across models can affect results and whether popular measures of fit compensate sensibly.

Construct a dataset as follows. Using R, construct 50 variables as independent random draws from a standardized normal distribution. Each predictor should have 100 observations. One easy way to do this is drawing 50 times 100 values using rnorm() and then formatting the result into a 50 by 100 matrix with matrix(). It will turn out to be helpful if that matrix is made into a data frame using data.frame(). Then, attach the data frame using the attach() command.

Now run a linear regression using that data frame. Apply lm() and assign the output to some name so that you can retrieve it later. To keep things simple, the only argument should be the data frame name. The first column will automatically be chosen as the response variable. Do not look at the output (yet).

Now apply a stepwise regression to the data. There are several stepwise regression procedures in R, but stepAIC() in the MASS library is one good one. Just feed the saved output object from lm() into stepAIC(). Again, save the output by assigning it to some name. For this exercise, accept the default settings. Do not look at the output (yet).

1. What do you expect the output from the first regression analysis (not the stepwise regresson) to look like: the regression coefficients, the t-tests, the R^2, the adjusted R^2, and the F-test?

2. Now look at the output from the first regression. How does the actual output compare with your expectations?

3. What do you conclude about how well linear regression produces results consistent with how you know the data were generated?

4. Now examine the output from the stepwise regression: the regression coefficients, the *t*-tests, the R^2, the adjusted R^2, and the *F*-test. How do these compare to the output from your initial regression analysis?

5. From that comparison, what are the possible implications for model selection and overfitting?

6. How would these implications change if the ratio of the number of observations to the number of predictors were much larger (e.g., 10 predictors with 100 observations) or much smaller (e.g., 90 predictors with 100 observations)? Try it. What happens? How do the comparisons between the conventional regression results and the stepwise regression results change depending on the ratio of the number of observations to the number of predictors?

7. What are some lessons for model selection and overfitting?