Chapter 1
Statistical Learning as a Regression Problem

Before getting into the material, it may be important to reprise and expand a bit on three points made in the first and second prefaces — most people do not read prefaces. First, any credible statistical analysis combines sound data collection, intelligent data management, an appropriate application of statistical procedures, and an accessible interpretation of results. This is sometimes what is meant by "analytics." More is involved than applied statistics. Most statistical textbooks focus on the statistical procedures alone, which can lead some readers to assume that if the technical background for a particular set of statistical tools is well understood, a sensible data analysis automatically follows. But as some would say, "That dog won't hunt."

Second, the coverage is highly selective. There are many excellent encyclopedic, textbook treatments of machine/statistical learning. Topics that some of them cover in several pages, are covered here in an entire chapter. Data collection, data management, formal statistics, and interpretation are woven into the discussion where feasible. But there is a price. The range of statistical procedures covered is limited. Space constraints alone dictate hard choices. The procedures emphasized are those that can be framed as a form of regression analysis, have already proved to be popular, and have been throughly battle tested. Some readers may disagree with the choices made. For those readers, there are ample references in which other materials are well addressed.

Third, the ocean liner is slowly starting to turn. Over the past decade, the 50 years of largely unrebutted criticisms of conventional regression models and extensions have started to take. One reason is that statisticians have been providing useful alternatives. Another reason is the growing impact of computer science on how data are analyzed. Models are less salient in computer science than in statistics, and

The original version of this chapter was revised: See the "Chapter Note" section at the end of this chapter for details. The erratum to this chapter is available at https://doi.org/10.1007/978-3-319-44048-4_10.

R.A. Berk, *Statistical Learning from a Regression Perspective*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-44048-4_1

far less salient than in popular forms of data analysis. Yet another reason is the growing and successful use of randomized controlled trials, which is implicitly an admission that far too much was expected from causal modeling. Finally, many of the most active and visible econometricians have been turning to various forms of quasi-experimental designs and methods of analysis in part because conventional modeling often has been unsatisfactory. The pages ahead will draw heavily on these important trends.

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 can be 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 previously been 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, and others. There is also widespread use of acronyms: CART, LOESS, MARS, MART, LARS, LASSO, and many more. A regression framework provides a convenient and instructive structure in which these procedures can be more easily understood.

After a discussion of how statisticians think about regression analysis, this 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 of this preliminary material on the table, and readers are encouraged to return to the chapter as needed.

1.2 Setting the Regression Context

We begin by defining regression analysis. A common conception in many academic disciplines and policy applications equates regression analysis with some special case of the generalized Linear model: normal (linear) regression, binomial regression, Poisson regression, or other less common forms. Sometimes, there is more than one such equation, as in hierarchical models when the regression coefficients in one equation can be expressed as responses within other equations, or when a set of

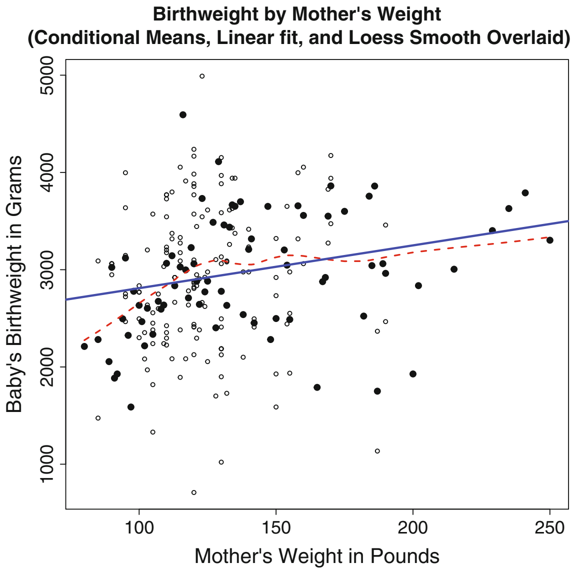


Fig. 1.1 Birthweight by mother's weight (*Open circles* are the data, *filled circles* are the conditional means, the *solid line* is a linear regression fit, the *dashed line* is a fit by a smoother. $N = 189$.)

equations is linked though their response variables. For any of these formulations, inferences are often made beyond the data to some larger finite population or a data generation process. Commonly these inferences are combined with statistical tests and confidence intervals. It is also popular to overlay causal interpretations meant to convey how the response distribution would change if one or more of the predictors were independently manipulated.

But statisticians and computer scientists typically start farther back. Regression is "just" about conditional distributions. 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. Regression analysis fundamentally is the about conditional distributions: $Y|X$.

For example, Fig. 1.1 is a conventional scatter plot for an infant's birth weight in grams and the mother's weight in pounds.[1] Birthweight can be an important indicator of a newborn's viability, and there is reason to believe that birthweight depends in part on the health of the mother. A mother's weight can be an indicator of her health.

In Fig. 1.1, the open circles are the observations. The filled circles are the conditional means and the likely summary statistics of interest. An inspection of the pattern of observations is by itself a legitimate regression analysis. Does the conditional distribution of birthweight vary depending on the mother's weight? If the conditional mean is chosen as the key summary statistic, one can consider whether the conditional means for infant birthweight vary with the mother's weight. This too

[1] The data, *birthwt*, are from the MASS package in R.

is a legitimate regression analysis. In both cases, however, it is difficult to conclude much from inspection alone. The solid blue line is a linear least squares fit of the data. On the average, birthweight increases with the mother's weight, but the slope is modest (about 44 g for every 10 pounds), especially given the spread of the birthweight values. For many, this is a familiar kind of regression analysis. The dashed red line shows the fitted values for a smoother (i.e., lowess) that will be discussed in the next chapter. One can see that the linear relationship breaks down when the mother weighs less than about 100 pounds. There is then a much stronger relationship with the result that average birthweight can be under 2000 g (i.e., around 4 pounds). This regression analysis suggests that on the average, the relationship between birthweight and mother's weights is nonlinear.

None of the regression analyses just undertaken depend on a "generative" model; no claims are made about how the data were generated. There are also no causal claims about how mean birthweight would change if a mother's weight is altered (e.g., through better nutrition). And, there is no statistical inference whatsoever. The regression analyses apply solely to the data on hand and are not generalized to some large set of observations. A regression analysis may be enhanced by such extensions, although they do not go to the core of how regression analysis is defined. In practice, a richer story would likely be obtained were additional predictors introduced, perhaps as "controls," but that too is not a formal requirement of regression analysis. Finally, visualizations of various kinds can be instructive and by themselves can constitute a regression analysis.

The same reasoning applies should the response be categorical. Figure 1.2 is a spine plot that dichotomizes birth weight into two categories: low and not low. For each decile of mothers' weights, the conditional proportions are plotted. For example, if a mother's weight is between 150 and 170 pounds, a little under 20 % of the

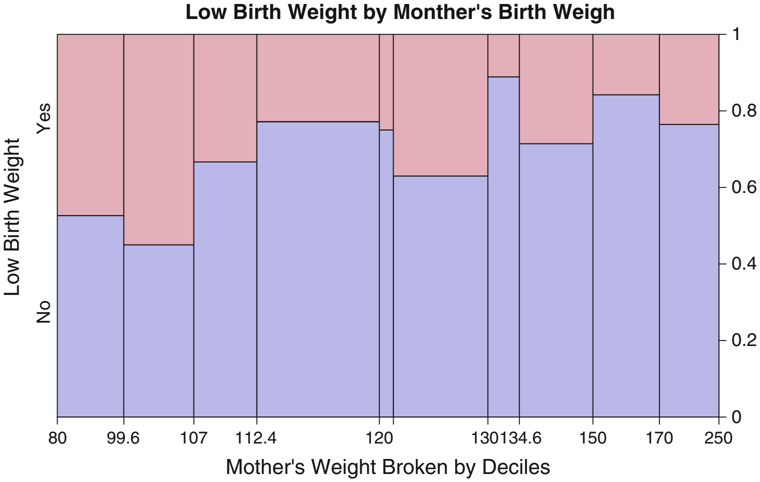


Fig. 1.2 Low birth weight by mother's weight with birth weight dichotomized (Mother's weight is binned by deciles. $N = 189$.)

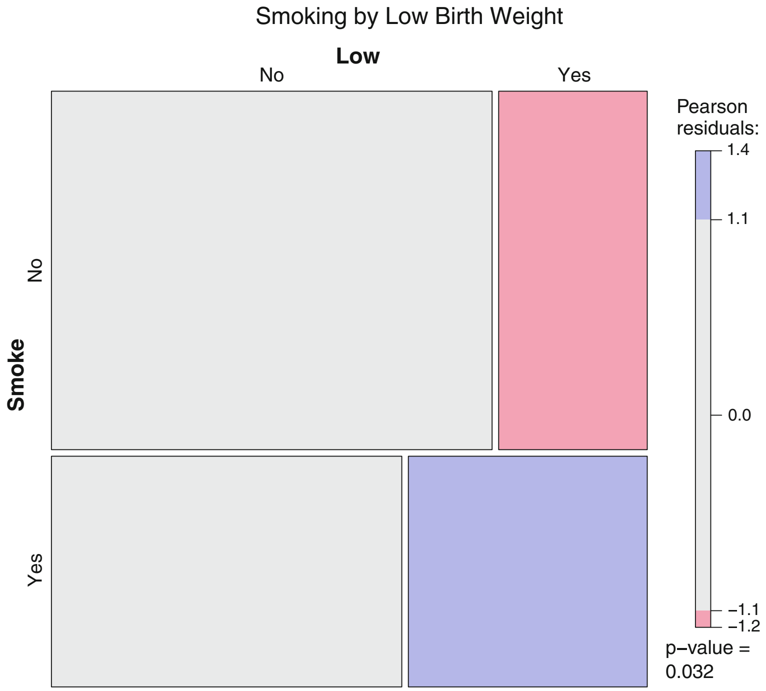


Fig. 1.3 Whether the mother smokes by low birth weight with Pearson residuals assuming independence (*Red* indicates fewer cases than expected under independence. *Blue* indicates more cases than expected under independence. $N = 189$.)

newborns have low birth weights. But if a mother's weight is less than 107 pounds, around 40 % of the newborns have low birth weights.

The reasoning applies as well if both the response and the predictor are categorical. Figure 1.3 shows a mosaic plot for whether or not a newborn is underweight and whether or not the newborn's mother smoked. The area of each rectangle is proportional to the number of cases in the respective cell of the corresponding 2 × 2 table. One can see that the majority of mothers do not smoke and a majority of the newborns are not underweight. The red cell contains fewer observations than would be expected under independence, and the blue cell contains more observations than would be expected under independence. The metric is the Pearson residual for that cell (i.e., the contribution to the χ^2 statistic). Mothers who smoke are more likely to have low birth weight babies. If one is prepared to articulate a credible generative model consistent with a conventional test of independence, independence is rejected at the .03 level. But even without such a test, the mosaic represents a legitimate regression analysis.[2]

[2] The spine plot and the mosaic plot were produced using the R package *vcd*, which stands for "visualizing categorical data." Its authors are D. Meyer et al. (2007).

There are several lessons highlighted by these brief illustrations.

- As discussed in more depth shortly, the regression analyses just conducted made no direct use of models. Each is best seen as a *procedure*. One might well have preferred greater use of numerical summaries and algebraic formulations, but regression analyses were undertaken nevertheless. In the pages ahead, it will be important to dispense with the view that a regression analysis automatically requires arithmetic summaries or algebraic models. Once again, regression is just about conditional distributions.
- Visualizations of various kinds can be a key feature of a regression analysis. Indeed, they can be the defining feature.
- A regression analysis does not have to make conditional means the key distributional feature of interest, although conditional means or proportions dominate current practice. With the increasing availability of powerful visualization procedures, for example, entire conditional distributions can be examined.
- Whether it is the predictors of interest or the covariates to "hold constant," the choice of conditioning variables is a subject-matter or policy decision. There is nothing in data by itself indicating what role, if any, the available variables should play.[3]
- There is nothing in regression analysis that requires statistical inference: inferences beyond the data on hand, formal tests of null hypotheses, or confidence intervals. And when statistical inference is employed, its validity will depend fundamentally on how the data were generated. Much more will said about this in the pages ahead.
- If there is to be cause-and-effect overlay, that too is a subject-matter or policy call unless one has conducted an experiment. When the data result from an experiment, the causal variables are determined by the research design.
- A regression analysis can serve a variety of purposes.

1. For a "level I" regression analysis, the goal is solely description of the data on hand. Level I regression is effectively assumption-free and should always be on the table. Too often, description is undervalued as a data analysis tool perhaps because it does not employ much of the apparatus of conventional statistics. How can a data analysis without statistical inference be good? The view taken here is that p-values and all other products of statistical inference can certainly be useful, but are worse than useless when a credible rationale cannot be provided (Berk and Freedman 2003). Assume-and-proceed statistics is not likely to advance science or policy. Yet, important progress frequently can be made from statistically informed description alone.
2. For a "level II" regression analysis, statistical inference is the defining activity. Estimation is undertaken using the results from a level I regression, often in

[3] Although there are certainly no universal naming conventions, "predictors" can be seen as variables that are of subject-matter interest, and "covariates" can be seen as variables that improve the performance of the statistical procedure being applied. Then, covariates are not of subject-matter interest. Whatever the naming conventions, the distinction between variables that matter substantively and variables that matter procedurally is important. An example of the latter is a covariate included in an analysis of randomized experiments to improve statistical precision.

concert with statistical tests and confidence intervals. Statistical inference forms the core of conventional statistics, but proper use with real data can be very challenging; real data may not correspond well to what the inferential tools require. For the statistical procedures emphasized here, statistical inference will often be overmatched. There can be a substantial disconnect between the requirements of proper statistical inference and adaptive statistical procedures such as those central to statistical learning. Forecasting, which will play an important role in the pages ahead, is also a level II activity because projections are made from data on hand to the values of certain variables that are unobserved.

3. For a "level III" regression analysis, causal inference is overlaid on the results of a level I regression analysis, sometimes coupled with level II results. There can be demanding conceptual issues such as specifying a sensible "counterfactual." For example, one might consider the impact of the death penalty on crime; states that have the death penalty are compared to states that do not. But what is the counterfactual to which the death penalty is being compared? Is it life imprisonment without any chance of parole, a long prison term of, say, 20 years, or probation? In many states the counterfactual is life in prison with no chance of parole. Also, great care is needed to adjust for the possible impact of confounders. In the death penalty example, one might want to control for average clearance rate in each of the state's police departments. Clearance rates for some kinds of homicides are very low, which means that it is pretty easy to get away with murder, and the death penalty is largely irrelevant.[4] Level III regression analysis will not figure significantly in the pages ahead because of a reliance on algorithmic methods rather than model-based methods (Breimen 2001b).

In summary, a focus on conditional distributions will be a central feature in all that follows. One does not require generative models, statistical inference, or causal inference. On the one hand, a concentration on conditional distribution may seem limiting. On the other hand, a concentration on conditional distributions may seem liberating. In practice, both can be true and be driven substantially by the limitations of conventional modeling to which we now briefly turn.

Of necessity, the next several sections are more technical and more conceptually demanding. Readers with a substantial statistical background should have no problems, although some conventional ways of thinking will need to be revised. There may also need to be an attitude adjustment. Readers without a substantial statistical background may be best served by skimming the material primarily to see the topics addressed, and then returning to the material as needed when in subsequent chapters those topics arise.

[4] A crime is "cleared" when the perpetrator is arrested. In some jurisdictions, a crime is cleared when the perpetrator has been identified, even if there has been no arrest.

1.3 Revisiting the Ubiquitous Linear Regression Model

Although conditional distributions are the foundation for all that follows, linear regression is its most common manifestation in practice and needs to be explicitly addressed. For many, linear regression is the canonical procedure for examining conditional relationship, or at least the default. Therefore, a brief review of its features and requirements can be a useful didactic device to highlight similarities to and differences from statistical learning.

When a linear regression analysis is formulated, conventional practice combines a level I and level II perspective. Important features of the data are conceptually embedded in how the data were generated. Y is an $N \times 1$ numerical response variable, where N is the number of observations. There is an $N \times (p + 1)$ "design matrix" $\mathbf{X}$, where p is the number of predictors (sometimes called regressors). A leading column of 1s is usually included in $\mathbf{X}$ for reasons that will clear momentarily. Y is treated as a random variable. The p predictors in $\mathbf{X}$ are taken to be fixed. Whether predictors are fixed or random is not a technical detail, but figures centrally in subsequent material.

The process by which the values of Y are realized then takes the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \tag{1.1}$$

where

$$\varepsilon_i \sim \text{NIID}(0, \sigma^2). \tag{1.2}$$

β_0 is the y-intercept associated with the leading column 1s. There are p regression coefficients, and a random perturbation ε_i. One might say that for each case i, nature sets the values of the predictors, multiplies each predictor value by its corresponding regression coefficient, sums these products, adds the value of the constant, and then adds a random perturbation. Each perturbation, ε_i, is a random variable realized as if drawn at random and independently from a single distribution, often assumed to be normal, with a mean of 0.0. In short, nature behaves as if she adopts a linear model.

There are several important implications. To begin, the values of Y can be realized repeatedly for a given case because its values will vary solely because of ε. The predictor values do not change. Thus, for a given high school student, one imagines that there could be a limitless number of scores on the mathematics SAT, solely because of the "noise" represented by ε_i. All else in nature's linear combination is fixed: the number of hours spent in an SAT preparation course, motivation to perform well, the amount of sleep the night before, the presence of distractions while the test is being taken, and so on. This is more than an academic formality. It is a substantive theory about how SAT scores come to be. For a given student, nature requires that an observed SAT score could have been different by chance alone, but not because any of variation in the predictors.[5]

[5] If on substantive grounds one allows for nature to set more than one value for any given predictor and student, a temporal process is implied, and there is systematic temporal variation to build into the regression formulation. This can certainly be done, but the formulation is more complicated,

From Eqs. 1.1 and 1.2, it can be conceptually helpful to distinguish between the mean function and the disturbance function (also called the variance function). The mean function is the expectation of Eq. 1.1. When in practice a data analyst specifies a conventional linear regression model, it will be “first-order correct” when the data analyst (a) knows what nature is using as predictors, (b) knows what transformations, if any, nature applies to those predictors, (c) knows that the predictors are combined in a linear fashion, and (d) has those predictors in the dataset to be analyzed. For conventional linear regression, these are the first-order conditions. The only unknowns in the mean function are the values of the y-intercept and the regression coefficients. Clearly, these are daunting hurdles.

The disturbance function is Eq. 1.2. When in practice the data analyst specifies a conventional linear regression model, it will be “second-order correct” when the data analyst knows that each perturbation is realized independently of all other perturbations and that each is realized from a single distribution that has an expectation of 0.0. Because there is a single disturbance distribution, one can say that the variance of that distribution is “constant.” These are the usual second-order conditions. Sometimes the data analyst also knows the functional form of the distribution. If that distribution is the normal, the only distribution unknown whose value needs to be estimated is its variance σ^2.

When the first-order conditions are met and ordinary least squares is applied to the data, estimates of the slope and y-intercept are unbiased estimates of the corresponding values that nature uses. When in addition to the first-order conditions, the second-order conditions are met, and ordinary least squares is applied to the data, the disturbance variance can be estimated in an unbiased fashion using the residuals from the realized data. Also, conventional confidence intervals and statistical tests are valid, and by the Gauss–Markov theorem, each estimated β has the smallest possible sampling variation of any other linear estimator of nature’s regression parameters. In short, one has the ideal textbook results for a level II regression analysis. Similar reasoning properly can be applied to the entire generalized linear model and its multi-equation extensions, although usually that reasoning depends on asymptotics.

Finally, even for a conventional regression analysis, there is no need to move to level III. Causal interpretations are surely important when they can be justified, but they are an add-on, not an essential element. With observational data, moreover, causal inference can be in principle very controversial (Freedman 1987, 2004).

1.3.1 Problems in Practice

There are a wide variety of practical problems with the conventional linear model, many recognized well over a generation ago (e.g., Leamer 1978; Rubin 1986, 2008; Freedman 1987, 2004; Berk 2003). This is not the venue for an extensive review, and

(Footnote 5 continued)
requires that nature be even more cooperative, and for the points to be made here, adds unnecessary complexity.

David Freedman's excellent text on statistical models (2009a) can be consulted for an unusually cogent discussion. Nevertheless, it will prove useful later to mention now a few of the most common and vexing difficulties.

There is effectively no way to know whether the model specified by the analyst is the means by which nature actually generated the data. And there is also no way to know how close to the "truth" a specified model really is. One would need to know that truth to quantify a model's disparities from the truth, and if the truth were known, there would be no need to analyze any data to begin with. Consequently, all concerns about model specification are translated into whether the model is good enough.

There are two popular strategies addressing the "good enough" requirement. First, there exist a large number of regression diagnostics taking a variety of forms and using a variety of techniques including graphical procedures, statistical tests, and the comparative performance of alternative model specifications (Weisberg 2014). These tools can be useful in identifying problems with the linear model, but they can miss serious problems as well. Most are designed to detect single difficulties in isolation when in practice, there can be many difficulties at once. Is evidence of nonconstant variance a result of mean function misspecification, disturbances generated from different distributions, or both? In addition, diagnostic tools derived from formal statistical tests typically have weak statistical power (Freedman 2009b), and when the null hypothesis is not rejected, analysts commonly "accept" the null hypothesis that all is well. In fact, there are effectively a limitless number of other null hypotheses that would also not be rejected.[6] Finally, even if some error in the model is properly identified, there may be little or no guidance on how to fix it, especially within the limitation of the data available.

Second, claims are made on subject-matter grounds that the results make sense and are consistent with – or at least not contradicted by – existing theory and past research. This line of reasoning can be a source of good science and good policy, but also misses the point. One might learn useful things from a data analysis even if the model specified is dramatically different from how nature generated the data. Indeed, this perspective is emphasized many times in the pages ahead. But advancing a scientific or policy discourse does not imply that the model used is right, or even close.

If a model's results are sufficiently useful, why should this matter? It matters because one cannot use the correctness of the model to justify the subject-matter claims made. For example, interesting findings said to be the direct product of an elaborate model specification might have surfaced just as powerfully from several scatter plots. The findings rest on a very few strong associations easily revealed by simple statistical tools. The rest is pretense.

It matters because certain features of the analysis used to bolster substantive claims may be fundamentally wrong and misleading. For example, if a model is not first-order correct, the probabilities associated with statistical tests are almost certainly incorrect. Even if asymptotically valid standard errors are obtained with such tools as the sandwich estimator (White 1980a, b), the relevant estimate from the data will

[6] This is sometimes called "the fallacy of accepting the null" (Rozeboom 1960).

on the average be offset by its bias. If the bias moves the estimate away from the null hypothesis, the estimated p-values will be on the average too small. If the bias moves the estimate toward the null hypothesis, the estimated p-values will on the average be too large. In a similar fashion, confidence intervals will be offset in one of the two directions.

It matters because efforts to diagnose and fix model specification problems can lead to new and sometime worse difficulties. For example, one response to a model that does not pass muster is to re-specify the model and re-estimate the model's parameters. But it is now well known that model selection and model estimation undertaken on the same data (e.g., statistical tests for a set of nested models) lead to biased estimates even if by some good fortune the correct model happens to be found (Leeb and Pötscher 2005; 2006; 2008; Berk et al. 2010; 2014).[7] The model specification itself is a product of the realized data and a source of additional uncertainty — with a different realized dataset, one may arrive at a different model. As a formal matter, statistical tests assume that the model has been specified *before* the data are examined.[8] This is no longer true. The result is not just more uncertainty overall, but a particular form of uncertainty that can result in badly biased estimates of the regression coefficients and pathological sampling distributions.

And finally, it matters because it undermines the credibility of statistical procedures. There will be times when an elaborate statistical procedure is really needed that performs as advertised. But why should the results be believed when word on the street is that data analysts routinely make claims that are not justified by the statistical tools employed?

1.4 Working with Statistical Models that Are Wrong

Is there an alternative way to proceed that can be more satisfactory? The answer requires a little deeper look at conventional practice. Emphasis properly is placed on the word "practice." There are no fundamental quarrels with the mathematical statistics on which conventional practice rests.

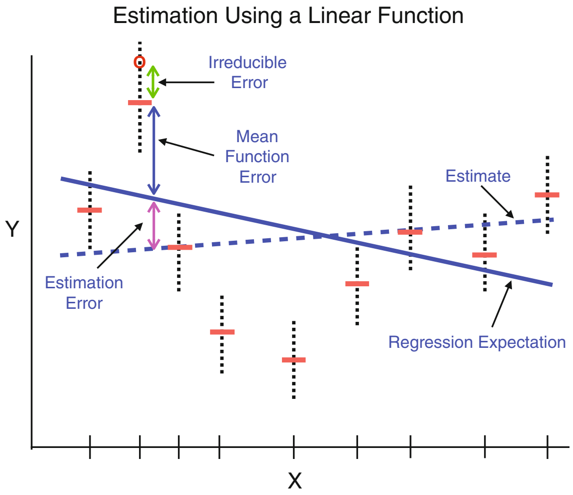
Model misspecification is hardly a new topic, and some very smart statisticians and econometricians have been working on it for decades. One tradition concentrates on patching up models that are misspecified. The other tradition tries to work constructively with misspecified models. We will work within the second tradition. For many statisticians and practitioners, this can require a major attitude adjustment.

Figure 1.4 is a stylized representation of the sort of practical problems that can follow for a level II analysis when for a linear model one assumes that the first-

[7] Model selection in some disciplines is called variable selection, feature selection, or dimension reduction.

[8] Actually, it can be more complicated. For example, if the predictors are taken to be fixed, one is free to examine the predictors. Model selection problems surface when the response variable is examined as well. If the predictors are taken to be random, the issues are even more subtle.

Fig. 1.4 Estimation of a nonlinear response surface under the true linear model perspective (The *broken line* is an estimate from a given dataset, *solid line* is the expectation of such estimates, the *vertical dotted lines* represent conditional distributions of Y with the *red bars* as each distribution's mean.)



and second-order conditions are met. The Figure is not a scatterplot but an effort to illustrate some key ideas from the relevant statistical theory. For simplicity, but with no important loss of generality for the issues to be addressed, there is a single predictor on the horizontal axis. For now, that predictor is assumed to be fixed.[9] The response variable is on the vertical axis.

The red, horizontal lines in Fig. 1.4 are the true conditional means that constitute nature's response surface. The vertical, black, dotted lines are meant to show the distribution of y-values around each conditional mean. Those distributions are also nature's work. No assumptions are made about what form the distributions take, but for didactic convenience each conditional distribution is assumed to have the same variance.

An eyeball interpolation of the true conditional means reveals an approximate U-shaped relationship but with substantial departures from that simple pattern. Nature provides a data analyst with realized values of Y by making independent draws from the distribution associated with each conditional mean. The red circle is one such y-value; the red circle is one output from nature's data generation process.

A data analyst assumes the usual linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. With a set of realized y values and their corresponding x values (not shown), estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$ are obtained. The broken blue line shows the estimated mean function. One can imagine nature generating many (formally, a limitless number) such datasets so that there are many mean function estimates that will naturally vary because the realized values y will change from dataset to dataset. The solid blue line represents the expectation of those many estimates.

[9] If one prefers to think about the issues in a multiple regression context, the single predictor can be replaced by the predictor adjusted, as usual, for its linear relationships with the other predictors.

Clearly, the assumed linear mean function is incorrect because the true conditional means do not fall on a straight line. The blue, two-headed arrow shows the bias at one value of x. The size and direction of the biases differ over the values of x because the disparities between regression expectation and the true conditional means differ.

The data analyst does not get to work with the expectation of the estimated regression lines. Usually, the data analyst gets to work with one such line. The random variation captured by one such line is shown with the magenta, double-headed error. Even if the broken blue line fell right on top of the solid blue line, and if both went exactly through the true conditional mean being used as an illustration, there would still be a gap between the observed value of Y (the red circle) and that conditional mean (the short red horizontal line). In Fig. 1.4, that gap is represented by the green, double-headed arrow. It is sometimes called "irreducible error" because it exists even if nature's response surface is known.

Summarizing the implications for the conventional linear regression formulation, the blue double-headed arrow shows the bias in the estimated regression line, the magenta double-headed arrow shows the impact of the variability of that estimate, and the green double-headed arrow shows the irreducible error. For any given estimated mean function, the distance between the estimated regression line and a realized y-value is a combination of mean function error (also called mean function misspecification), random variation in the estimated regression line caused by ε_i, and the variability in ε_i itself. Sometimes these can cancel each other out, at least in part, but all three will always be in play.

Some might claim that instrumental variables provide a way out. It is true that instrumental variable procedures can correct for some forms of bias if (a) a valid instrument can be found and if (b) the sample size is large enough to capitalize on asymptotics. But the issues are tricky (Bound et al. 1995). A successful instrument does not address all mean function problems. For example, it cannot correct for wrong functional forms. Also, it can be very difficult to find a credible instrumental variable. Even if one succeeds, an instrumental variable may remove most of the regression coefficient bias and simultaneously cause a very large increase in the variance of the regression coefficient estimate. On the average, the regression line is actually farther away from the true conditional means even through the bias is largely eliminated. One is arguably worse off.

It is a simple matter to alter the mean function. Perhaps something other than a straight line can be used to accurately represent nature's true conditional means. However, one is still required to get the first-order conditions right. That is, the mean function must be correct. Figure 1.5 presents the same kinds of difficulties as Fig. 1.4. All three sources of error remain: model misspecification, sampling variability in the function estimated, and the irreducible error. Comparing the two figures, the second seems to have on the average a less biased regression expectation, but in practice it is difficult know whether that is true or not. Perhaps more important, it is impossible to know how much bias remains.[10]

[10] We will see later that by increasing the complexity of the mean function estimated, one has the potential to reduce bias. But an improved fit in the data on hand is no guarantee that one is

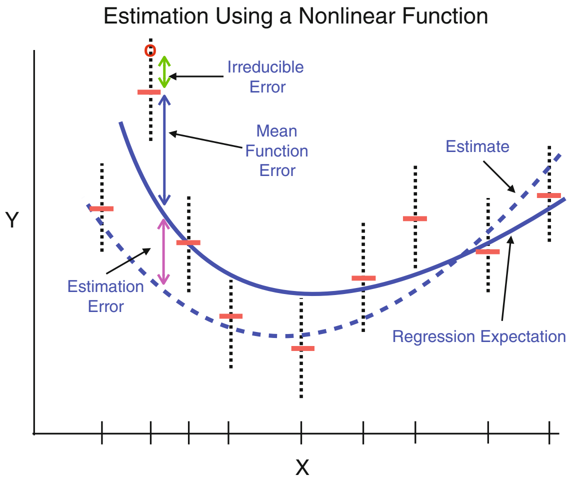


Fig. 1.5 Estimation of a nonlinear response surface under the true nonlinear model perspective (The *broken line* is an estimate from a given dataset, *solid line* is the expectation of such estimates, the *vertical dotted lines* represent conditional distributions of Y with the *red bars* as each distribution's mean.)

One important implication of both Figs. 1.4 and 1.5 is that the variation in the realized observations around the fitted values will not be constant. The bias, which varies across x-values, is captured by the least squares residuals. To the data analyst, this will look like heteroscedasticity even if the variation in ε_i is actually constant. Conventional estimates of σ^2 will likely be incorrect. Incorrect standard errors for the intercept and slope follow, which jeopardize statistical tests and confidence intervals.

When faced with non-constant variance, the "sandwich" estimator (White 1980b) can provide asymptotically valid standard errors. But the mean function must be correctly specified. The requirement of proper mean function specification too commonly is overlooked.

It seems that we are at a dead end. But we are not. All of the estimation difficulties are level II regression problems. If one can be satisfied with a level I regression analysis, these difficulties disappear. Another option is to reformulate conventional linear regression so that the estimation task is more modest. We turn to that next. Yet another option considered in later chapters requires living with, and even reveling in, at least some bias. Unbiased estimates of the nature's response surface are not a prerequisite if one can be satisfied with estimates that are as close as possible on the average to nature's response surface over realizations of the data. There can be bias if in trade, there is a substantial reduction in the variance; on the average, the regression line is then closer to nature's response surface. We will see that in practice, it is difficult to decrease both the bias and the variance, but often there will be ways which arrive at a beneficial balance in what is called the "bias–variance tradeoff." Still, as long as any bias remains, statistical tests and confidence intervals need to be reconsidered. As for the irreducible variance, it is still irreducible.

(Footnote 10 continued)
more accurately representing the mean function. One complication is that greater mean function complexity can foster overfitting.

1.4.1 An Alternative Approach to Regression

The material in this section can be conceptually demanding and has layers. There are also lots of details. It may be helpful, therefore, to make two introductory observations. First, in the words of George Box, "All models are wrong..." (Box 1976). It follows that one must learn to work with wrong models and not proceed as if they are right. This is a large component of what follows. Second, if one is to work with wrong models, the estimation target is also a wrong model. Standard practice has the "true" model as the estimation target. In other words, one should be making correct inferences to an incorrect model and not be making incorrect inferences to a correct model. Let's see how these two observations play out.

If a data analyst wants to employ a level II regression analysis, inferences from the data must be made to something. Within conventional conceptions, that something is the parameter of a linear model used by nature to generate the data. The parameters are the estimation targets. Given the values of those parameters and the fixed-x values, each y_i is realized by the linear model shown in Eqs. 1.1 and 1.2.[11]

Consider as an alternative what one might call the "joint probability distribution model." It has much the same look and feel as the "correlation model" formulated by Freedman (1981), and is very similar to a "linear approximation" perspective proposed by White (1980a). Both have important roots in the work of Huber (1967) and Eicker (1963, 1967). Angrist and Pischke (2008: Sect. 3.1.2) provide a very accessible introduction.

For the substantive or policy issues at hand, one imagines that there exists a materially relevant, joint probability distribution composed of variables represented by $\mathbf{Z}$. The joint probability distribution has familiar parameters such the mean (i.e., the expected value) and variance for each variable and the covariances between variables. No distinctions are made between predictors and responses. Nature can "realize" independently any number of observations from the joint probability distribution. This is how the data are generated. One might call the process by which observations are realized from the joint probability distribution the "true generative model." This is the "what" to which inferences are to be made in a level II analysis.

A conceptually equivalent "what" is to consider a population of limitless size that represents all possible realizations from the joint probability distribution. Inferences are made from the realized data to this "infinite population." In some circles, this is called a "superpopulation." Closely related ideas can work for finite populations (Cochran 1977: Chap. 7). For example, the data are a simple random sample from a well-defined population that is in principle observable. This is the way one usually thinks about sample surveys, such as well-done political polls. The population is all registered voters and a probability sample is drawn for analysis. In finite populations, the population variables are fixed. There is a joint distribution of all the variables in the population that is just a multivariate histogram.

[11] The next several pages draw heavily on Berk et al. (2014) and Buja et al. (2016).

Switching to matrix notation for clarity, from $\mathbf{Z}$, data analysts will typically distinguish between predictors $\mathbf{X}$ and the response $\mathbf{y}$. Some of $\mathbf{Z}$ may be substantively irrelevant and ignored. These distinctions have nothing to do with how the data are generated. They derive from the preferences of the individuals who will be analyzing the data.

For any particular regression analysis, attention then turns to a conditional distribution of $\mathbf{y}$ given some $\mathbf{X} = \mathbf{x}$. For example, $\mathbf{X}$ could be predictors of longevity, and $\mathbf{x}$ is the predictor values for a given individual. The distribution of $\mathbf{y}$ is thought to vary from one $\mathbf{x}$ to another $\mathbf{x}$. Variation in the mean of $\mathbf{y}$, $\mu(\mathbf{x})$, is usually the primary concern. But now, because the number of observations in the population is limitless, one must work with the $\mathrm{E}[\mu(\mathbf{x})]$.

The values for $\mathrm{E}[\mu(\mathbf{x})]$ constitute the "true response surface." The true response surface is the way the expected values of Y are actually related to $\mathbf{X}$ within the joint probability distribution. It is unknown. Disparities between the $\mathrm{E}[\mu(\mathbf{x})]$ and the potential values of Y are the "true disturbances" and necessarily have an expectation of 0.0 (because they are deviations around a mean – or more properly, an expected value)

The data analyst specifies a working regression model using a conventional, linear mean function meant to characterize *another* response surface within the same joint probability distribution. Its conditional expectations are equal to $\mathbf{X}\boldsymbol{\beta}$. The response $\mathbf{y}$ is then taken to be $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is an array of least squares coefficients. Because $\boldsymbol{\varepsilon}$ also is a product of least squares, it has by construction an expectation of 0.0 and is uncorrelated with $\mathbf{X}$. For reasons that will be clear later, there is no requirement that $\boldsymbol{\varepsilon}$ have constant variance. Nevertheless, thanks to least squares, one can view the conditional expectations from the working model as the *best linear approximation* of the true response surface. We will see below that it is the best linear approximation of the true response surface that we seek to estimate, not the true response surface itself.

This is a major reformulation of conventional, fixed-x linear regression. For the working model, there is no a priori determination of how the response is related to the predictors and no commitment to linearity as the truth. In addition, the chosen predictors share no special cachet. Among the random variables $\mathbf{Z}$, a data analyst determines which random variables are predictors and which random variables are responses. Hence, there can be no such thing as an omitted variable that can turn a correct model into an incorrect model. If important predictors are overlooked, the regression results are just incomplete; the results are substantively insufficient but still potentially very informative. Finally, causality need not be overlaid on the analysis. Although causal thinking may well have a role in an analyst's determination of the response and the predictors, a serious consideration of cause and effect is not required at this point. For example, one need not ponder whether any given predictor is actually manipulable holding all other predictors constant.

Still to come is a discussion of estimation, statistical tests and confidence intervals. But it may be important to pause and give potential critics some air time. They might well object that we have just traded one fictional account for another.

From an epistemological point of view, there is real merit in such concerns. However, in science and policy settings, it can be essential to make empirically based claims that go beyond the data on hand. For example, when a college admissions office uses data from past applicants to examine how performance in college is related to the information available when admission decisions need to be made, whatever is learned will presumably be used to help inform future admission decisions. Data from past applicants are taken to be realizations from the social processes responsible for academic success in college. Insofar as those social processes are reasonably consistent over several years, the strategy can have merit. A science fiction story? Perhaps. But if better admissions decisions are made as a result, there are meaningful and demonstrable benefits. To rephrase George Box's famous aphorism, all models are fiction, but some stories are better than others. And there is much more to this story.

1.4.1.1 Statistical Inference with Wrong Models

Figure 1.6 can be used to help understand estimation within the "wrong model" framework. It is a stylized rendering of the joint probability distribution. There is a single predictor treated as a random variable. There is a single response, also treated as a random variable. Some realized values of Y are shown as red circles. The solid back line represents nature's unknown, true response surface, the "path" of the conditional means, or more accurately, the path of the conditional expectations.

The true response surface is allowed to be nonlinear, although for ease of exposition, the nonlinearity in Fig. 1.6 is rather well behaved. For each location along the response surface, there is a conditional distribution represented in Fig. 1.6 by the dotted, vertical lines. Were one working with a conventional regression perspective, the curved black line would be the estimation target.

Under the wrong model perspective, the straight blue line in Fig. 1.6 represents the mean function implied by the data analyst's working linear model. Clearly, the linear mean function is misspecified. It is as if one had fitted a linear least squares regression within the joint probability distribution. The blue line is the new estimation target that can be interpreted as the best linear approximation of the true response surface. It can be called "best" because it is conceptualized as a product of ordinary least squares; it is best by the least square criterion. Although the best linear approximation is the estimation target, one also gets estimates of the regression coefficients responsible. These may be of interest for least squares regression applications and procedures that are a lot like them. By the middle of the next chapter, however, most of the connections to least squares regression will be gone.

Consider the shaded vertical slice of the conditional distribution toward the center of Fig. 1.6. The disparity between the true response surface and the red circle near the top of the conditional distribution results solely from the irreducible error. But when the best linear approximation is used as a reference, the apparent irreducible error is much smaller. Likewise, the disparity between the true response surface and the red circle near the bottom of the conditional distribution results solely from the

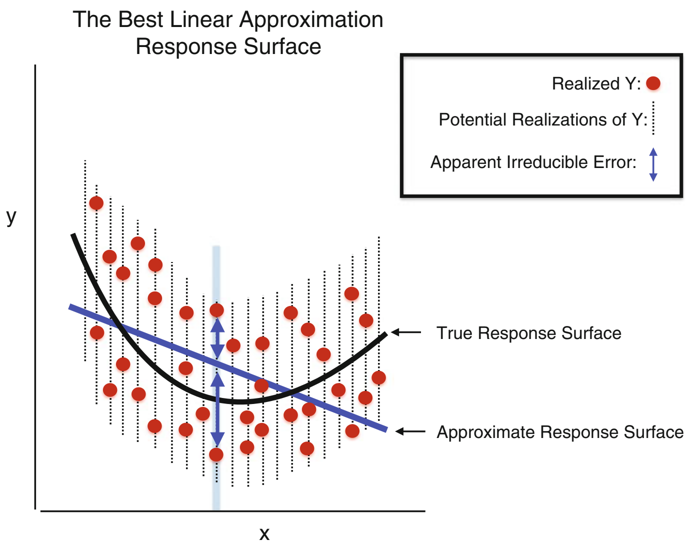


Fig. 1.6 Within the joint probability distribution, mean function error as a cause of nonconstant variance (The *black curved line* is the true response surface, and the *straight blue line* is the best linear approximation of that response surface.)

irreducible error. But when the best linear approximation is used as a reference, the apparent irreducible error is much larger. Both distortions result from the gap between the true response surface and the best linear approximation response surface. Because X is a random variable, mean function misspecification is a random variable captured as a component of the *apparent* irreducible error. Similar issues arise for the full range of x-values in the figure.

Suppose a data analyst wanted to estimate from data the best linear approximation of nature's true response surface. The estimation task can be usefully partitioned into five steps. The first requires making the case that each observation in the dataset was independently realized from a relevant joint probability distribution. Much more is required than hand waving. Required is usually subject-matter expertise and knowledge about how the data were collected. There will be examples in the pages ahead. Often a credible case cannot be made, which takes estimation off the table. Then, there will probably be no need to worry about step two.

The second step is to define the target of estimation. For linear regression of the sort just discussed, an estimation target is easy to specify. Should the estimation target be the true response surface, estimates will likely be of poor statistical quality. Should the estimation target be the best linear approximation of the true response surface, the estimates can be of good statistical quality, at least asymptotically. We will see in later chapters that defining the estimation target often will be far more difficult because there will commonly be no model in the conventional regression sense. One cannot sensibly proceed to step three unless there is clarity about what is to be estimated.

The third step is to select an estimator. Sometimes the best estimator will be apparent. The least squares estimator used in conventional regression is a good example. There are other relatively straightforward examples when the mean function is determined without any formal model selection or data snooping. But most of the procedures considered in later chapters capitalize on model selection, even if not quite in plain sight, and informal data snooping is a common practice. Getting the estimator right can then be challenging. The risk is that an inappropriate estimator is used by default, justified by performance claims that are incorrect.

Fourth, the estimator needs to be applied to the data. This is usually the easiest step because the requisite software is often easily found and easily deployed. But there are exceptions when the data have unusual properties or the questions being asked of the data are unusual. One hopes that appropriate software can be easily written, but sometimes the underlying statistical problems are unsolved.

In the final step, the estimates and any associated confidence intervals and tests are interpreted. The major risk is that the results of earlier steps are not properly taken into account. A common example is asserting asymptotic properties with far too few observations in the data.

Let us play this through for estimates of the best linear approximation. In the absence of real data not much can be said here about the first step. It will figure large

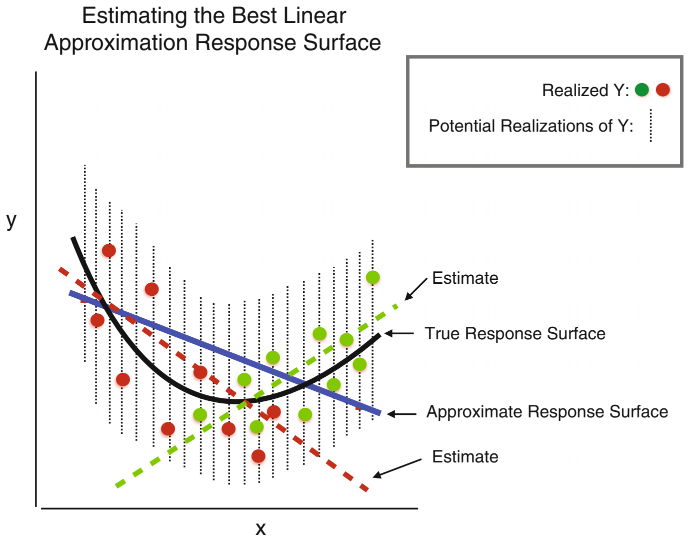


Fig. 1.7 Bias in the estimates of the best linear approximation (The *black curved line* is the true response surface, and the *straight blue line* is the best linear approximation of that response surface, and estimates are shown as *broken lines*.)

in real applications later. The second step has already been addressed. The estimation target is the best linear approximation of the true response surface.

Important technical complications materialize for steps three, four, and five. Consider Fig. 1.7. Suppose that in the data, the realized values of X tend to be concentrated at the smaller values. This is illustrated by the red filled circles in Fig. 1.7. Because the nonlinear true response surface is sloping downward where the red observations are more likely to be concentrated, estimated least squares lines tend to have a negative slope. In contrast, suppose that the realized values of X tend to be concentrated at the larger values. This is illustrated by the green filled circles in Fig. 1.7. Because the nonlinear true response surface is sloping upward where the green observations are more likely to be concentrated, estimated least squares lines will tend to have a positive slope. For a conventional fixed X regression, this leads to based estimates of the best linear approximation.

However, under the joint probability distribution approach, all observations are realized by the equivalent of random sampling, and all the predictors are random variables. One can show, therefore, that conventional estimates of the best linear approximation are asymptotically unbiased (Buja et al. 2016). And there is more good news. Despite the nonconstant variance described earlier, asymptotically valid standard errors may be obtained using a sandwich procedure or a nonparametric bootstrap. Asymptotically valid statistical tests and confidence intervals follow (Buja et al. 2016).

These conclusions apply to nonlinear parametric approximations as well. For example, one might choose to approximate the true response surface with a cubic polynomial function of X. One would have the best cubic approximation of the true response surface. The conclusions also apply to the entire generalized linear model (White 1980a). For example, the response might be binary and a form of binomial regression might be the estimation procedure. In short, the standard estimators can work well in large samples, and in those large samples, the sandwich or a bootstrap estimator of the regression coefficient standard errors can work well too. And there is readily available software for both. Asymptotically, valid statistical tests and confidence intervals follow.

It may seem that in the second step, the selection of the best linear approximation as the estimation target comes perilously close to a resurrection of conventional model-based procedures. But there is absolutely no reason to be limited to a linear model, and subject-matter knowledge may suggest a better mean function. There remains a very important role for subject-matter theory and the results from past research. One is still trying to learn about the true response surface, but that knowledge will come from working models that are, in conventional terms, misspecified. The way one learns from models that are wrong is not to pretend they are right. The way one learns from wrong models is to acknowledge their imperfections and exploit their instructive features nevertheless. There can be important complications to be sure, but they will be constructively addressed in the chapters ahead.

Given the goal of learning about the true response surface with a misspecified model, one can imagine using that model to forecast the response. That is, for a *new* vector of x-values realized from the same joint probability distribution, the misspecified model is used to produce a good guess about the response when the response value is not known. Ideally, that forecast will fall right on the true response surface, but in practice this will not happen very often. (And how would you know?) One hopes, therefore, to be close most of the time. But, if the forecasting target is the true response surface, one has reintroduced all of the estimation problems faced when working with wrong models assumed to be specified correctly. We seem to be back where we started.

Actually, we are not. In principle, one can obtain estimates of the best linear approximation that have good asymptotic properties. The approximation can provide useful information about how the response is related to the predictors, informed by asymptotically valid statistical tests and confidence intervals. None of this is available within the conventional approach to linear regression when the model's mean function is misspecified. Then, as an *additional* step, one can try to forecast accurately values of true response surface. This step necessarily reintroduces the problems just discussed. But, we will see that there are ways to make good progress here too. In the spirit of the approximation approach to estimation, one gives up the goal of unbiased forecasts and settles for trying to get forecasts that are close. How this can be done depends on the fitting procedure used and will be discussed in some detail in the pages ahead.

For many readers, the wrong model formulation addressed in the past few pages may seem odd and perhaps even heretical. But from a sampling perspective, our wrong model formulation was anticipated in the work of Willian Cochran over 50 years ago (1977: Chap. 7). One has a finite population, such as all students at a university. A simple random sample is drawn, and a conventional linear regression applied. The predictors are random variables because in new random samples, the students and their x-values will change. The estimation target is the response surface of the same regression specification were it applied in the population. It does not matter whether the mean function for population regression is specified properly; wrong models are fine. Estimates of the wrong model's response surface are biased. When the number of observations in the sample approaches the number of observations in the finite population, the bias disappears. Valid statistical inference can follow (Thompson 2002: Sect. 8.1).

1.4.1.2 Wrong Regression Models with Binary Response Variables

Up to this point, analyses with quantitative response variables have dominated the discussion, in part because of a desire to make connections to conventional linear regression. The majority of subsequent chapters will introduce procedures for the analysis of categorical response variables in part because that is where some of the most interesting and useful work on statistical learning can be found.

To help set the stage, in Fig. 1.8, there is a binary response coded as 1 or 0, and as before, a single numerical X. Dropping out of high school might be coded as 1, and graduating might be coded as 0. The fuzzy lines at y-values of 1 and 0 are the potential realized values of Y. There is, as before, a true response surface and an approximation. The latter is the estimation target although there can be interest in the regression coefficients responsible. Fitted values can range from 0 to 1. For a level I analysis, the fitted values often can be interpreted as proportions. For a level II analysis, the fitted values often can be interpreted as probabilities.

Typically, the fitted values are used to determine class membership. In Fig. 1.8, there are two classes: A and B. A might be dropping out of high school and B might be graduating. There is also a threshold shown at an illustrative value of .30. Fitted values at or above the threshold imply membership in class A. Fitted values below that threshold imply membership in class B. Notice that the true response surface and its approximation classify many cases differently. For the value of *X* at the gray vertical rectangle, the true response surface would classify a case as an A, and the approximation would classify a case as a B. The use of .30 as a threshold might seem strange, but we will see later that there can principled reasons for choosing threshold values other than .50.

Within a parametric perspective, one might apply logistic regression to the data. Typically, fitted values of .50 or larger imply membership in one class ("50–50 or better"). Fitted values smaller than .50 imply membership in the other class ("less than 50–50"). Hence, the threshold is .50. Within statistical learning traditions, there are several effective ways to estimate the nonlinear approximation in a nonparametric fashion. These will be discussed in later chapters.

1.4.1.3 Wrong Models in Practice

We are adopting a perspective for level II regression in which the data are generated as independent realizations from a joint probability distribution. All the realized

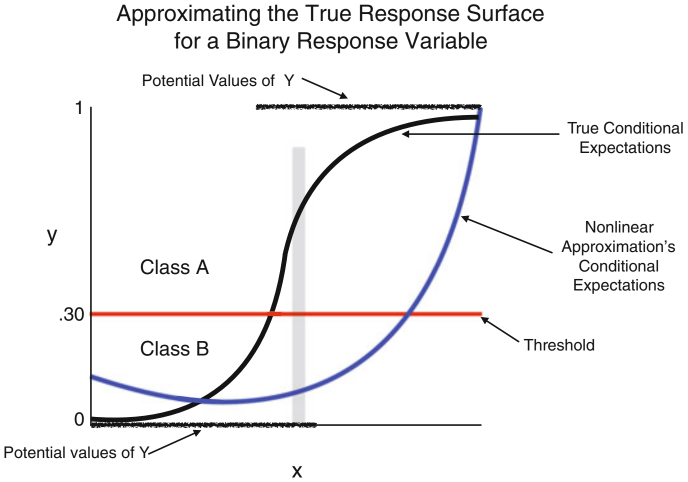


Fig. 1.8 Estimation with the best linear approximation for a binary response Class A or B

variables are random variables. Responses and predictors are determined by the data analyst, not by nature. There is a true response surface of conditional expectations of a response that is unknown. The true response surface is not the estimation target. The estimation target is an approximation of that response surface whose relation to the true response surface is undetermined. No restrictions are placed on the approximation's functional form, and it can be specified before looking at the data or, as we will see later, arrived at inductively. When specified in advance, the approximation can be estimated in an asymptotically unbiased fashion with valid standard errors, confidence intervals, and statistical tests. When the specification is determined as part of the data analysis, the issues are more complicated, and we will address them subsequently. With an estimate of the approximation in hand, one can compute fitted values that may be used as estimates of the true response surface. One should proceed assuming that these estimates are biased, even asymptotically. The goal is to get close to the true response surface on the average. Getting the response surface right on the average usually is too hard.

A superficial reading of the last several pages might suggest that in practice one could proceed with linear or logistic regression as usual by simply providing a more cautious interpretations of the results. This typically will not be a good idea. Because the estimation target is still the true response surface, estimates of the regression coefficients and fitted values will be biased, even asymptotically. It follows that even if valid asymptotic standard errors could be computed, statistical tests and confidence intervals will not be correct. There is also the likely prospect of substantial additional work trying to diagnose model misspecifications and their consequences. And in the end, justification of results will often look like just so much hand waving.

So, what should be done in practice? For each of the special cases of the generalized linear model, one computes as one ordinarily would, but uses either sandwich or nonparametric bootstrap standard error estimates. Then, proper asymptotic statistical tests and confidence intervals can be constructed. The regression coefficients can be interpreted as usual with one critical caveat: they are the product of *an incorrect model*. That is, they are covariance adjusted measures of association as usual, but the adjustments differ from those of a correct model. For example, one can obtain an estimate of the how the probability of a fatal car accident differs for 18 year olds compared to 25 year olds, which could be useful for insurance companies to know. But that estimate might well differ if covariance adjustments were made for average miles driven per month, the kind of vehicle driven, and use of alcohol. A more formal and complete discussion can be found in the paper by Buja and his colleagues (2016), but moving to a level III analysis will likely be ill advised.

1.5 The Transition to Statistical Learning

As a first approximation, statistical learning can be seen as a form of nonparametric regression in which the search for an effective mean function is especially data intensive. For example, one might want to capture how longevity is related to genetic and lifestyle factors. A statistical learning algorithm could be turned loose on a large dataset with no preconceptions about what the nature of the relationships might be. Fitted values from the exercise could then be used to anticipate which kinds of patients are more likely to face life-threatening chronic diseases. Such information would be of interest to physicians or actuaries. Likewise, a statistical learning algorithm could be applied to credit card data to determine which features of transactions are associated with fraudulent use. Again, there would be no need for any preconceptions about the relationships involved. With the key relationships determined, banks that issue credit cards would be able to alert their customers when their cards were likely being misused.

So far, this sounds little different from conventional regression analyses. And in fact, there has been considerable confusion in the applications literature about the nature of statistical learning and which procedures qualify (Baca-García et al. 2006; Kessler et al. 2015). In truth, the boundaries are fuzzy. The difference between models and algorithms is a good place to start clarifying the issues.

1.5.1 Models Versus Algorithms

Consider again the pair of equations for the conventional linear model.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{p_i} + \varepsilon_i, \tag{1.3}$$

where

$$\varepsilon_i \sim \text{NIID}(0, \sigma^2). \tag{1.4}$$

Equations 1.3 and 1.4 are a theory of how each case *i* came to be. They are generative models because they represent the data generation mechanisms. When a data analyst works with these equations, a substantial amount of thought goes into model specification. What predictors should be used? What, if any, transformations are needed? What might be done about possible dependence among the disturbances or about nonconstant disturbance variances? But once these decisions are made, the necessary computing follows almost automatically. Usually, the intent is to minimize the sum of the squared residuals, and there is a convenient closed-form solution routinely implemented. On occasion, a more robust fitting criterion is used, such as minimizing the sum of the absolute value of the residuals, and although there is no closed form solution, the estimation problem is easily solved with linear programming.

Statistical learning allocates a data analysts' effort differently. Equations 1.3 and 1.4 often can be replaced by

$$y_i = f(\mathbf{X}_i) + \varepsilon_i, \tag{1.5}$$

where $f(\mathbf{X}_i)$ is some unknown function of one or more predictors, and ε_i is a residual if the analysis is level I. In a level II analysis, ε_i is a random disturbance whose assumed properties can depend on the data and statistical procedure used.

There can be two level II interpretations of Eq. 1.5. Consistent with conventional practice in regression, Eq. 1.5 may be interpreted as how nature generated the data, and then ε_i has its usual properties. It is just that the form of the mean function is not specified. However, that interpretation is inconsistent with the joint probability distribution framework used in statistical learning and raises again all of the problems with conventional linear regression.

The second interpretation builds on the mean function approximation approach used above for working with incorrect models. The $f(\mathbf{X}_i)$ is the estimation target taken to be some approximation of the true response surface, and ε_i is an additive disturbance whose properties we will consider shortly. But Eq. 1.5 is not a model of the data generation process.

Very little is asked of the data analyst because no commitment to any particular mean function is required. Indeed, the only decision may be to introduce ε_i additively. There is effectively no concern about whether the mean function is right or wrong because for all practical purposes there is no model responsible for the data. In practice, the objective is to arrive at fitted values from a computed $\hat{f}(\mathbf{X}_i)$ that make subject-matter sense and that correspond as closely as possible to realized values of Y. What form the $\hat{f}(\mathbf{X}_i)$ takes to get the job done can be of little concern.

In contrast, there needs to be serious thought given to the algorithm through which the fitted values are computed. Hence, the methods often are called "algorithmic." A simple and somewhat stylized outline of one kind of algorithmic method proceeds as follows.

1. Specify a linear mean function of the form of Eq. 1.3 and apply least squares as usual.
2. Compute the fitted values.
3. Compute the residuals.
4. Compute a measure of fit.
5. Apply least squares again, but weight the data so that observations with larger residuals are given more weight.
6. Update the fitted values obtained from the immediately preceding regression with the new fitted values weighted by their measure of fit.
7. Repeat steps 2–6 until the quality of the fit does not improve (e.g., 1000 times).
8. Output the fitted values.

In a process that has some of the same look and feel as boosting, a regression analysis is repeated over and over, each time with altered data so that hard-to-fit values of the Y are given more weight. In that sense, the hard-to-fit observations are counted more heavily when the sum of squared residuals is computed. The final result is a single set of fitted values that is a weighted sum of the many sets of fitted values. Hard thought must be given to whether this algorithm is an effective way to link predictors to a response and whether other algorithms might do the job better.

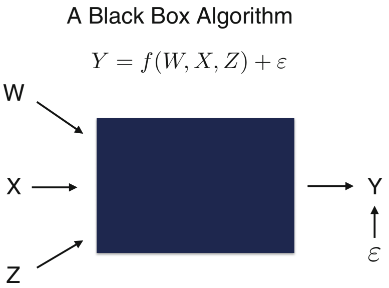


Fig. 1.9 Inputs W, X, and Z linked to the output Y through a black box algorithm

There are also interpretative issues. In this algorithm, there can be a very large number of regressions and an even larger number of regression coefficients. For example, if there are 1000 regressions and 10 predictors, there are 10,000 regression coefficients. It is effectively impossible to make subject matter sense of 10,000 regression coefficients. Moreover, each set is computed for data with different weights so that the fitting task is continually changing.

In the end, one has a "black box" algorithm of the form shown in Fig. 1.9. There are in Fig. 1.9 three inputs $\{W, X, Z\}$, and a single output Y, connected by complicated computations that provide no information of substantive importance.[12] One can get from the inputs to the final set of fitted values. But, there is no model. The many regressions are just a computational device. In other terms, one has a procedure not a model.

Thinking back to the college admissions example, one can use the results of the algorithm to forecast a college applicant's freshman GPA *even though one does not know exactly how the predictors are being used to make that forecast*. In a similar fashion, one can use such methods to determine the dollar value of insurance claims a given driver is likely to make over the course of a year or the total precipitation a city will receive in a particular month. When Y is binary, one can project which parolees are likely to be rearrested or which high school students are likely to drop out of school.

There can be more interpretative information when one is able to change what inputs are used and re-run the algorithm. One could determine, for example, how the fitted values for college GPA change whether or not gender is included as an input. One could determine how much any measures of fit change as well. We will see later that there are special algorithms operating in much the same spirit that allow one to at least peep into the black box.

But one must be clear about exactly what might be learned. Suppose the association between gender and GPA operates in concert with age. The association between gender and college GPA is stronger for younger students perhaps because male stu-

[12] Some academic disciplines like to call the columns of **X** "inputs," and Y an "output" or a "target." Statisticians typically prefer to call the columns of **X** "predictors" and Y a "response." By and large, the terms predictor (or occasionally, regressor) and response will be used here except when there are links to computer science to be made. In context, there should be no confusion.

dents do not mature as rapidly as female students. As a result, should the quality of the fit improve when gender is included, the improvement results from a main effect and an interaction effect. Moreover, the algorithm might have transformed age in a manner that is unknown to the data analyst. A claim that on the average gender improves the quality of the fit is technically correct, but how gender is related to college GPA remains obscure.

A metaphor may help fix these ideas. Suppose one wants to bake some bread. The recipe calls for the following:

1. 2 packages of yeast
2. 2 cups of warm water
3. 2 tablespoons of melted butter
4. 2 tablespoons of sugar
5. 1 tablespoon of salt
6. 4 cups of all-purpose flour.

These ingredients are mixed and stirred until the batter is stiff, adding more flour if needed. The stiff batter is then kneaded until it is smooth and elastic, put into a greased bowl and allowed to rise for about an hour. The dough is then punched down and divided in half, placed into two greased loaf pans and again allowed to rise for about 30 min. Finally, the two loaves are baked at 400 ° for about 25 min.

The bread baking begins with known ingredients in specified amounts. From that point onward — the knead and baking — complicated physical and chemical processes begin that change the molecular structure of the ingredients as they are combined so that a bland watery batter can be turned into a delicious solid punctuated by air holes. The baker knows little about such details, and there is no way for the baker to document exactly how the ingredients are turned into bread. But if bread is tasty, the recipe will be repeated in the future. Looking again at Fig. 1.9, the ingredients are $\{W, X, Z\}$. The bread is Y. The black box is all of the physics and chemistry in-between.

It is possible to alter the bread recipe. For example, one might use 1 teaspoon of sugar rather than 2. That would likely lead to changes in the bread that comes out of the oven. It might be more preferable or less. Or one might choose to substitute whole wheat flour for the all-purpose flour. It is possible therefore, to see how changing the ingredients and/or their proportions affects the quality of the bread. But the baker does not learn much about the processes by which those ingredients are transformed into bread.[13]

Why might one prefer black box algorithmic methods rather than a traditional parametric regression? If the primary goal of the data analysis is to understand how the predictors are related to the response, one would not. But if the primary goal of the data analysis is to make decisions based at least in part on information provided by fitted values, statistical learning really has no downside. It should perform at least as well as model-based methods, and often substantially better. The reasons will

[13] In later chapters, several procedures will be discussed that can help one consider the “importance” of each input and how inputs are related to outputs.

be considered in later chapters when particular statistical learning procedures are discussed.

What roles do estimation, statistical tests and confidence intervals play? As before, they are effectively irrelevant for a level I analysis. For a level II analysis, the broader issues are the same as those already discussed. Inferences are being made to an approximation of the unknown response surface using the fitted values constructed from the data. However, that approximation is not model-based and is derived directly from the data. In the algorithm just summarized, for instance, each new weighted regression is altered because of the results of the earlier regressions. The algorithm is engaged in a very extensive form of data snooping. The resulting problems are similar to those produced by model selection, but with additional complications. One important complication is that there are tuning parameters whose values need to be determined by the data, and the number of observations can make an important difference. For example, if the sample size is 1000, the algorithm might well be tuned differently from when the sample size is 10,000, and the joint probability distribution to which inferences are being made has a limitless number of observations. What does tuning mean in that setting? In later chapters, when particular statistical learning procedure is discussed, the conceptual issues will be revisited and potential solutions provided.

Forecasting remains a level II activity. The approximation is used to compute forecasts and consequently, the forecasts will contain bias. One hopes that the forecasts are close to the truth, but there is no way to know for sure. As before, we will see that real progress can be made nevertheless.

Finally, we need to briefly address what to call an algorithmic approach for linking inputs to outputs. Suppose again that we have a set of fitted values constructed from 1000 linear, residual-weighted regressions. Do we call the computed relationships between the inputs and the outputs a model? In statistics, the term “model” is often reserved for the “generative model.” The model conveys how the data were generated. But we are proceeding, at least for level II applications, assuming the data are generated as realizations from a joint probability distribution. That is not what is represented by each of the 1000 regressions. So, calling those 1000 regressions a model can be confusing.

Unfortunately, there seems to be no commonly accepted alternative term. We will proceed from here on with one of four terms: “algorithmic model,” “algorithmic procedure,” “statistical learning procedure,” or most simply, “procedure.” There should be no confusion in context.

1.6 Some Initial Concepts

Within a regression analysis framework, a wide variety of statistical learning procedures are examined in subsequent chapters. But, before going much farther down that road, a few key concepts need to be briefly introduced. They play central roles in the chapters ahead and, at this point, would benefit from some initial exposure. We return to these ideas many times, so nothing like mastery is required now. And that is a good thing. Important details can only be addressed later in the context of particular statistical learning procedures. For now, we consider what statistical learning looks like from 30,000 ft up.

1.6.1 Overall Goals of Statistical Learning

The range of procedures we examine have been described in several different ways (Christianini and Shawe-Taylor 2000; Witten and Frank 2000; Hand et al. 2001; Breiman 2001b; Dasu and Johnson, 2003; Bishop 2006; Hastie et al. 2009; Barber 2012; Marsland 2014; Sutton and Barto 2016), and associated with them are a variety of names: statistical learning, machine learning, supervised learning, reinforcement learning, algorithmic modeling, and others. The term "Statistical learning" as emphasized in the pages that follow, is based on the following notions.

The definition of regression analysis still applies, but as already noted, some statisticians like a function estimation framework. Thus,

$$Y = f(X) + \varepsilon \tag{1.6}$$

or

$$G = f(X) + \varepsilon. \tag{1.7}$$

For a quantitative response variable, the goal is to examine $Y|X$ for a response Y and one or more 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 a data generated from a joint probability distribution, X is a random variable, although it is sometimes convenient to condition on the values of X that happen to be realized in the sample.

In conventional level II regression models, ε is assumed to be IID, with a mean of 0 and a constant variance. Drawing on the best linear approximation perspective, ε now conflates, as before, irreducible error, estimation (sampling) error, and bias. The new wrinkle is that at this point, no structure is imposed on the mean function. For a level I analysis, ε is just a residual.

Many different features of $Y|X$ and $G|X$ can be examined with statistical learning procedures. Usually conditional means or proportions, respectively, are of primary interest. For a level I analysis, fitted values suffice with no additional conceptual scaffolding. For a level II analysis, the fitted values are taken to be estimates of the conditional expectations of the response surface approximation discussed earlier. But there are several different flavors of this undertaking depending on the statistical procedure and the nature of the data.

When the response is categorical, the conditional expectations are usually interpreted as probabilities. The categories of G are often called "classes," and the data analysis exercise is often called "classification." There are also some issues that are unique to classification.

As anticipated in our earlier discussion of linear regression, it is very important to distinguish between fitted values computed when the response values are known and fitted values computed when the response values are unknown. Suppose for the moment that the response values are known. One then hopes that a statistical learning procedure succeeds in finding substantial associations between a set of predictors and a response. But, what is the task when a new set of predictor values is provided without known response values? This motivates a very different, but related, enterprise. The goals differ, but one cannot undertake the second unless the first has been reasonably well accomplished.[14]

When the response is known, the usual intent is to characterize how well the procedure is performing and for some statistical learning approaches, to describe how the inputs are related to the outputs. The aims may be substantive, but how the procedure performs will often lead to tuning. In our illustrative algorithm shown earlier, one might decide to slow down the speed at which the fitted values are updated even if that means re-running the regressions many more times.[15] There can sometimes be additional analysis steps making the statistical learning results more understandable.

When the response is unknown, there are two different analysis activities that depend on why the response is not known. The first kind of analysis is imputation. The response values exist but are not in the data. For example, can the carbon emissions from a coal-powered energy plant be approximated from information about the kind of technology the plant uses and amount of coal burned over a day? Can a student's true score in a standardized test be inferred from a pattern of answers that suggests cheating? Can the fish biomass of a tropical reef be estimated from information about the kinds of coral of which the reef is composed, the size of the reef, water temperature, and the amount of fishing allowed?

The second kind of analysis is forecasting: an outcome of interest is not just unknown, it has not occurred yet. What is the likely return from a given investment? What will be the top wind speed when a particular hurricane makes shore? For a certain county in Iowa, how many bushels of corn per acre can be expected in September from information available in June?

For categorical responses, one might try to impute an unknown class. For example, does a pattern of expenditures indicate that a credit card is being used fraudulently? Does a DNA test place a given suspect at the crime scene? Do the injuries from a car crash indicate that the driver was not wearing a seat belt? But just as with quantitative responses, forecasts are commonly undertaken as well. Will a particular

[14] When there is no interest whatsoever in a response Y, and attention is limited exclusively to X, supervised learning is no longer on the table, but unsupervised learning can be. Some kinds of cluster analysis can be seen as examples of unsupervised learning. Supervised learning becomes unsupervised learning when there is no response variable. In computer science terms, there is no "labeled" response.

[15] For instance, one might divide the updating set of fitted values by 10 making their updating impact 10 times weaker.

prison inmate be rearrested within two years when later released on parole? Will a certain earthquake fault rupture in the next decade? Will a given presidential candidate win the overall popular vote when a year later the election is held?

Why does the difference between imputation and forecasting matter? There are usually operational matters such as for imputation trying to understand why the data are missing? The answers help determine what remedies might be appropriate. The deeper difference is that in contrast to imputation, forecasting involves the realization of new cases from a joint probability distribution. One has to consider whether the same joint probability distribution is the source and whether the new cases are random realizations.

In short, statistical learning focuses on fitted values as the primary algorithmic product. They may be used for description, especially when associations with predictors can be calculated or displayed. As important in practice, associations between predictors and a response are used to compute fitted values when the response values are unknown. The enterprise can be imputation or forecasting.

1.6.2 Data Requirements: Training Data, Evaluation Data and Test Data

It is well known that all statistical procedures are vulnerable to overfitting. The fitting procedure capitalizes on the noise in the dataset as well as the signal. This is an especially important problem for statistical learning because there can be many ways to conflate noise with signal.

For a statistical learning level I analysis, the result can be an unnecessarily complicated summary of relationships and the risk of misleading interpretations. For a level II analysis, the problems are more complex and troubling. An important consequence of overfitting is that when performance of the procedure is examined with new data, even if realized from the same joint probability distribution, predictive accuracy will often degrade substantially. If out-of-sample performance is meaningfully worse than in-sample performance, generalizing the results from original data is compromised.

Overfitting can be exacerbated when the statistical learning procedure is allowed to seek a complex $f(X)$. Recall that in conventional linear regression, the greater the number of non-redundant regression coefficients whose values are to be estimated, the better the in-sample fit, other things equal. There have been, therefore, many attempts to develop measures of fit that adjust for this source of overfitting, typically by taking the degrees of freedom used by the procedure into account. Mallows' Cp is an example (Mallows 1973). Adjusting for the degrees of freedom used can counteract an important contributor to overfitting. Unfortunately, that does not solve the problem because the adjustments are all in-sample.

Probably the most important challenge to out-of-sample performance stems from various kinds of data snooping. It is common for data analysts to fit a regression model, look at the results and revise the regression model in response. This has become a widely promoted practice over the past several decades as various forms of regression diagnostic techniques have been developed. Thus, if evidence of non-constant variance is found in the residuals, a transformation of the response variable may be undertaken and the regression run again. Sometimes there is automated data snooping. Stepwise regression is a popular example. Many different models are compared in-sample, and the "best" one is chosen. Statistical learning procedures also data snoop because typically they are tuned. Sometimes the tuning is automated, and sometimes tuning is done by the data analyst. For example, tuning can be used to determine how complex the $\hat{f}(X)$ should be.

Data snooping can have especially pernicious effects because the problems caused go far beyond overly optimistic measure of performance. Data snooping introduces *model* uncertainty that is not included in standard formulations of statistical inference. That is, with different realizations of the data, there can be different models chosen. Canonical frequentist statistical inference assumes a known, fixed model before the data analysis begins. When data snooping leads to revising a mean function specification, the new mean function will typically lead to biased estimates, even asymptotically, and all statistical tests and confidence intervals can be invalid (Leeb and Pötscher 2005, 2006, 2008; Berk et al. 2010).[16] For level II analyses, these problems apply to statistical learning as well as conventional modeling.

Several in-sample solutions have been proposed (Berk et al. 2014; Lockhart et al. 2014), but they are not fully satisfactory, especially for statistical learning (Dwork et al. 2015). When a sufficient amount of data is available, the problems of overfitting and model selection sometimes can be effectively addressed with an out-of-sample approach. The realized data to which the statistical learning procedure is initially applied are usually called "training data." Training data provide the information through which the algorithm learns. There is then a second dataset, sometimes called "evaluation data," realized from the same joint probability distribution, used in the tuning process. The statistical learning procedure is tuned not by its performance in the training data, but by its performance in the evaluation data. One uses the results from the training data to predict the response in the evaluation data. How well the predicted values correspond to the actual evaluation data outcomes provides feedback on performance. Finally, there is a third dataset, commonly called "test data," also realized from the same joint probability distribution, used to obtain an "honest" assessment of the procedure's performance. Once a statistical learning procedure has been satisfactorily tuned, there can be a proper measure of out-of-sample performance. Much as was done with the evaluation data, a fitting and/or prediction

[16] Data snooping can also begin before an initial model is specified, and the implications are much the same. For example, all bivariate correlations between predictors and a response can be used to determine which predictors are selected for use in a regression analysis (Fan and Lv 2008).

exercise can be undertaken with the test data. The new set of fitted values can then be compared to the actual outcomes. In practice, there may be extensions of this process and interpretative complications, both of which will be addressed in subsequent chapters. For example, a lot depends on exactly what one is trying to estimate. Also, it can be useful to represent the uncertainty in the test sample results .

Arguably the most defensible approach is to have three datasets of sufficient size: a training dataset, an evaluation dataset, and a test dataset. "Sufficient" depends on the setting, but a minimum of about 500 cases each can be effective. All three should be realizations from the same joint probability distribution. If there is only one dataset on hand that is at least relatively large (e.g., 1500 cases), a training dataset, an evaluation dataset, and a test dataset can be constructed as three random, disjoint subsets. Then, there can important details to consider, such as the relative sizes of the three splits (Faraway 2014).

In addition, the split-sample approach is only justified asymptotically, and it is not clear how large a sample has to be before one is close enough. For example, if one or more of the key variables are highly skewed, no observations from the tails may have been included in the data overall, or in each of the data splits. Thus, in a study of sexually transmitted diseases (STDs), the very few individuals who have unprotected sex with a very large number of different individuals, may not be in the data. Yet these are the individuals who are most responsible for STD transmission. The chances that such a problem will materialize get smaller as the sample size gets bigger. But how big is big enough is not clear, at least in the abstract.

Finally, randomly splitting the data introduces a new source of uncertainty. Were the data split again, the fitted values would be at least somewhat different. In principle, this could be addressed with resampling. A large number of different splits could be used, and the fitted values in the test data across the different splits averaged. The main obstacle would be an additional computational burden.

When only training data are available, and the dataset on hand has too few observations for data splitting, there are several procedures one can use to try to approximate the out-of-sample ideal. Perhaps the most common is cross-validation. Consider a training data set with, say, 500 training observations and no evaluation or test data. Suppose the most pressing need is to document a procedure's performance conditional on the values of tuning parameters.

One can divide the data into five random, disjoint subsamples of 100 each, perhaps denoted by one of five letters A through E. The fitting procedure is applied to the 400 cases in subsets A–D and evaluated using remaining 100 "hold-out" cases in E. The fitting process is repeated for the 400 cases in subsets A, C, D, and E, and evaluated with the 100 hold-out cases in B. One proceeds in the same fashion until each of the five splits is used once as the holdout subset, and each split has four opportunities to be included with three other splits as training data. There are then five measures of performance that can be averaged to obtain an approximation of true evaluation data performance. For example, one might compute the average of five AICs (Akaike 1973). The tuning parameter values that lead to the best performance in the evaluation samples are selected. One has tuned using a fivefold cross-validation.

Cross-validation is no doubt a clever technique for level II analyses that might seem straightforward. It is not straightforward, and its use depends heavily on how the data were generated and the kind of statistical procedures undertaken. To begin, splitting the data five ways is somewhat arbitrary. Occasionally, there are as few as three splits and often as many as ten. Also, one can treat the jackknife as "leave-one-out" cross validation, in which case there are N splits. A larger number of splits means that the size of the training data relative to the test data is greater. This improves performance in the data training splits, but makes performance in the hold-out data less stable. The converse follows from a smaller number of splits. But, there is no formal justification for any particular number of splits which will typically depend on the setting. Common practice seems to favor either fivefold or tenfold cross-validation.

Much as with a split sample approach, there is a more fundamental issue: because there are no evaluations or test data, the results of cross-validation are conditional on the training data alone. Suppose one is trying to study the correlates of blindness for individuals under 50 years of age and has training data with 1000 observations. By the luck of the draw, 100 of those observations have macular degeneration whereas in the generative joint probability distribution, such cases are very rare. (Macular degeneration is usually found in substantially older people.) Moreover, that 10 % has a significant impact on any regression analyses. In all of those random splits of the data, about 10 % of the cases will be suffering from macular degeneration and the cross-validated regression results will be shaped accordingly. Again, defensible results depend on asymptotics.

In short, a lot can depend on having training data for which its distributional features are much like those of the generative joint probability distribution. And there is no way of knowing if this is true from the training data alone. Put another way, cross-validation is more likely to provide generalizable results from training data with a large number of observations. The empirical joint distribution in the data is more likely to correspond well to the joint probability distribution from which the data were realized. Formally, one needs to capitalize on asymptotics.

Finally, cross-validation is a popular procedure in part because single datasets that are too small to subdivide are common. But cross-validation shares with data splitting reductions in sample size. We will see later that many statistical learning procedures are sample size dependent. Smaller training datasets can increase bias in estimates of the true response surface. In addition, if the statistical learning results are sample size dependent, what does it mean to estimate features of a joint probably distribution for which the number of observations can be seen as limitless? We will address this matter in the next chapter.

In summary, for level II analyses one should try to avoid in-sample determination of tuning parameters and assessments of procedure performance. Having legitimate training data, evaluation data, and test data is probably the best option. Split samples are one fallback position, and cross-validation is another. They provide somewhat different challenges. Split samples provide separately for tuning and for honest performance assessments. Cross-validation forces a choice between proper tuning and honest performance assessments. Split samples do not immediately allow for representations of uncertainty, although there are extensions that can provide that information. Cross-validation can provide measures of uncertainty for many applications. There are also tradeoffs in how observations are allocated. Cross-validation uses most of the cases as training data. Split samples allow for a wide range of different data allocations. Sometimes, it effectively will not matter which one is used, and when it does matter, the choice will often be determined by the application. Later chapters have examples of both situations.

1.6.3 Loss Functions and Related Concepts

Loss functions can be used to quantify how well the output of a statistical procedure corresponds to certain features of the data. As the name implies, one should favor small loss function values. A very general expression for a loss function can be written as $L(Y, \hat{f}(X))$, where Y represents some feature of the data, and $\hat{f}(X)$ represents some empirical approximation of it. Often, Y is a response variable, and $\hat{f}(X)$ is the fitted values from some statistical procedure.[17]

In conventional treatments of estimation, there are loss functions that the estimators minimize with respect to the data on hand. Least squares regression, for example, minimizes the sum of the squared residuals. Least absolute residual regression, minimizes the sum of the absolute values of the residuals. For a level I analysis, these loss functions can be sensible ways to justify the summary statistics computed. Thus, least squares regression leads to fitted values for conditional means. Least absolute residual regression leads to fitted values for conditional medians. For a level II analysis, the resulting estimates have well-known and desirable formal properties as long as the requisite statistical assumptions are met. An absence of bias is a popular instance. But a key conceptual point is that when a loss function is minimized for the data on hand, performance is being addressed solely in-sample.

As already noted, there is a rich tradition of in-sample measures of fitting error that attempt to correct for the degrees of freedom used by the procedure. Recall that other things equal, a regression model that uses more degrees of freedom will automatically fit the data better. This is undesirable because a good fit should result from predictors that are strongly related to the response, not just because there are a lot of them. Measures that try to correct for the degrees of freedom used include the adjusted R^2, AIC, BIC, and Mallows Cp. The adjusted R^2 is popular because it seems easy to interpret, but it lacks a rigorous formal rationale. The other three measures each have good, but different, formal justifications (Hastie et al. 2009: Sect. 7.5).

[17] Loss functions are also called "objective functions" or "cost functions."

Unfortunately, they can still provide a falsely optimistic assessment of performance. The fitting is undertaken in-sample and capitalizes on idiosyncratic features of the data that undermine generalization. We need a better way.

For statistical learning, a better way can be especially important inasmuch as attention typically is directed to fitted values. A natural level II question, therefore, is how well the $\hat{f}(X)$ performs out-of-sample. One better way is to use *generalization error*, also called *test error*, as a performance metric (Hastie et al. 2009: 228). Suppose one has a test data observation denoted by (X^*, Y^*), where X^* can be a set of predictors. For the random variable Y, random variables X, and a statistical learning result constructed from the training data $\mathcal{T}$, generalization error is defined as

$$\text{Err}_{\mathcal{T}} = \text{E}_{(X^*,Y^*)}[L(Y^*, \hat{f}(X^*))|\mathcal{T})]. \tag{1.8}$$

The training data are treated as fixed once they are realized. A statistical procedure is applied to the training data in an effort minimize some loss function in-sample. The result is a set of parameter estimates. For example, in conventional regression, one obtains estimates of the regression coefficients and the intercept. At that point, the results of the minimization become fixed as well. Just as in many real forecasting settings, the training data and the parameters estimates are in the past and now do not change. Then, one may want to know how well the procedure performs with a test data observation. One can compare Y^* to $\hat{f}(X^*)$ within a loss function such as squared error.[18] But, $\text{E}_{(X^*,Y^*)}$ means that generalization error is the average loss over a limitless number of realized test observations, not a single observation. The result is an average loss in the test data. If one actually has test data, this is easy to operationalize. Cross-validation can be a fallback approach (Hastie et al. 2009: Sect. 7.12)

There can also be interest in the average generalization error, if in theory the training data $\mathcal{T}$ can also be realized over and over. With each realization of $\mathcal{T}$, the entire procedure represented in Eq. 1.8 is repeated. Then, *expected prediction error* (EPE) is defined as

$$\text{Err} = \text{E}_{\mathcal{T}}\text{E}_{(X^*,Y^*)}[L(Y^*, \hat{f}(X^*))|\mathcal{T})]. \tag{1.9}$$

If one has test data, the entire procedure can be wrapped in a resampling procedure such as a bootstrap. More will be said about the bootstrap in later chapters. Cross-validation once again can be a fallback position.

Whether one uses generalization error or expected prediction error depends on how much the X distribution matters. Because X and Y are treated as random variables and a response surface approximation is the estimation target, it can make sense to consider how the distribution of X can affect the results. Recall the earlier discussion of estimation with misspecified mean functions. But a lot depends on how the results of the estimation procedure will be used. With forecasting applications, for instance,

[18] In R, many estimation procedures have a *predict()* function that can easily be used with test data to arrive at test data fitted values.

generalization error may be more relevant than expected prediction error. Examples are provided in later chapters.

For categorical responses, generalization error can be written as

$$\mathrm{Err}_{\mathcal{T}} = \mathrm{E}_{(X^*,G^*)}[L(G^*, \hat{G}(X^*))|\mathcal{T})]. \tag{1.10}$$

As before, $\mathrm{E}[\mathrm{Err}_{\mathcal{T}}]$ is expected prediction error, and both $\mathrm{Err}_{\mathcal{T}}$ and $\mathrm{E}[\mathrm{Err}_{\mathcal{T}}]$ can be of interest.

A look back at Fig. 1.8 will provide a sense of what generalization error and expected prediction error are trying to capture with $\hat{G}|X^*$. Suppose there are two actual response categories coded as 1 or 0. There are fitted values in the metric of proportions. A natural in-sample measure of fit is the deviance. There are also fitted classes depending on where a given proportion falls with respect to the classification threshold. A natural in-sample measure of fit is the proportion of cases for which the fitted value is the same as the actual class. Some very powerful statistical learning procedures classify without the intermediate step of computing proportions, but the correspondence between the actual class and the fitted class is still central.

The rationale for choosing generalization error or expected prediction error does not change for categorical response variables. The estimation options are also much the same. There will be examples later.

The loss functions considered so far are symmetric. For a numerical Y, a fitted value that is too large by some specific amount makes the same contribution to the loss function as a fitted value that is too small by that same amount. Consider, 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 policy implications from underestimating the number of homeless individuals in a census tract (Berk et al. 2008). In the first instance, costly social services may be unnecessarily allocated to certain census tracts. In the second instance, those services may not be provided in census tracts that really need them. Yet, a symmetric loss function would assume that in the metric of costs, their consequences are exactly the same. One needs a loss function that properly takes the asymmetric costs into account so that the homeless estimates are responsive to how they will be used in practice.

Symmetric loss functions also dominate when the response variable is categorical. Suppose there are K exhaustive and mutually exclusive classes. Any misclassification — the fitted class is the wrong class — is given the same weight of 1.0. In a forecasting application, for instance, the error of predicting that a high school student will dropout when that student will not is given the same weight as predicting that a high school student will not dropout when that student will. (Correct classifications are given a value of 0.0.)

Once again, one must ask if symmetric costs are reasonable. Are the costs really the same, or even close? In the case of the potential high school dropout, are the costs of interventions for a student who needs no interventions the same as failing to provide interventions for a student who needs them? A lot depends on the content of those interventions (e.g., academic tutoring, counseling). Especially in policy settings where decision makers will be using the statistical learning results, symmetric costs may not be reasonable. Some mistakes are much worse than others, and the asymmetric costs must be allowed to affect how the statistical learning procedure performs. In later chapters, this will be a central concern.

1.6.4 The Bias-Variance Tradeoff

Before we move into more of the nuts and bolts of estimation, we need to revisit a bit more the bias–variance tradeoff. Recall that the bias–variance tradeoff is a level II problem that arises when the true response surface is explicitly the estimation target. The goal is to produce an estimate of the true response surface that is as close as possible to the truth. In principle, this can be achieved by a judicious tradeoff between the bias of the estimates and the variance in those estimates.

If the estimation target is the approximate response surface, there is no bias, at least asymptotically, but a closely related tradeoff can be in play. When the focus is on generalization error, for example, the goal is to impute or forecast as accurately as possible even though one explicitly is using an approximation of the true response surface. That is, the actual estimation target is the potential y-values in the joint probability distribution responsible for the data.

To illustrate, using an asymptotically unbiased estimate of the mean function approximation linking daily ozone concentrations in a city and emergency room visits for respiratory distress, one might want to forecast a day or two in advance how many such visits there might be. Many factors affect emergency room visits, so one is clearly working with a mean function approximation. Bias in the projected number of emergency room visits might be reduced by using a different approximation; the approximation could be made more complex. Thinking in parametric terms for the moment, a 4th degree polynomial might be used instead of a second-degree polynomial. But with a more complex parametric approximation, the effective degrees of freedom will be larger (more on that shortly). Other things equal, an increase in the effective degrees of freedom will increase the variance in fitted values as estimates. Hence, there is a potential tradeoff.

The tradeoff can be especially dicey with statistical learning because of the inductive nature of the procedures and the routine use of tuning. One problem is that the meaning and calculation of the effective degrees of freedom can be a challenge (Janson et al. 2015; Kauffman and Rosset 2014). Nevertheless, the split sample approach can work satisfactorily, and there are fallback positions that can also lead to useful results.

1.6.5 Linear Estimators

Level II statistical learning can capitalize on a wide variety of estimators. Linear estimators are often preferred because they can be seen as variants of conventional linear regression and are easily shown to have good statistical properties. Recall the hat matrix from conventional, fixed-x linear regression:

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}. \tag{1.11}$$

The hat matrix $\mathbf{H}$ transforms the y_i in a linear fashion into $\hat{y}_i$.

A smoother matrix is a generalization of the hat matrix. Suppose there is a training dataset with N observations, a single fixed 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

$$\hat{y}_0 = \sum_{j=1}^{N} \mathbf{S}_{0j} y_j. \tag{1.12}$$

$\mathbf{S}$ is an N by N matrix of fixed weights and is sometimes called a “smoother matrix.” $\mathbf{S}$ can be a product of a statistical learning procedure. The subscript 0 denotes the row corresponding to the case whose fitted value of y is to be computed. 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.

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.

$$\begin{pmatrix} 1.0 & 0 & 0 & 0 & 0 \\ .25 & .50 & .25 & 0 & 0 \\ 0 & .25 & .50 & .25 & 0 \\ 0 & 0 & .25 & .50 & .25 \\ 0 & 0 & 0 & 0 & 1.0 \end{pmatrix} \begin{pmatrix} 3.0 \\ 5.0 \\ 6.0 \\ 9.0 \\ 10.0 \end{pmatrix} = \begin{pmatrix} 3.00 \\ 4.75 \\ 6.50 \\ 8.50 \\ 10.00 \end{pmatrix}. \tag{1.13}$$

The leftmost matrix is $\mathbf{S}$. It is post multiplied by the vector $\mathbf{y}$ 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.

For a level II analysis, one has a linear estimator of the conditional means of Y. It is a linear estimator because with $\mathbf{S}$ 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. Linearity can make it easier to determine the formal properties of an estimator, which are often highly desirable. Unbiasedness is a primary example.[19]

When one views the data as generated from a joint probability distribution, X is no longer fixed.[20] Linear estimators with fixed values of X become nonlinear estimators with random values of X. The use of turning parameters also can lead to nonlinear estimators. One has to rely on asymptotics to derive an estimator's formal statistical properties. Asymptotic unbiasedness can sometimes be demonstrated depending on precisely what features of the joint probability distribution are being estimated. One also will need to rely on asymptotics to arrive at, when appropriate, proper standard errors, statistical tests, and confidence intervals.

1.6.6 Degrees of Freedom

Woven through much of the discussion of level II analyses has been the term "degrees of freedom." It will prove useful to expand just a bit on the several related conceptual issues. 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. 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 linear regression analysis. The number of degrees of freedom used 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 to be estimated (Weisberg 2014: 26).

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 **H** and **S** (Hastie et al. 2009: Sect. 5.4.1) imply that for linear estimators, the trace of **S** can be interpreted as the degrees of freedom used. Its value is sometimes called the "effective degrees of freedom" and can

[19] "Linear" in regression can be used in two different ways. An estimator may or may not be linear. The relationship between Y and X may or may not be linear. For example, an estimator may be linear, and the relationship between Y and X may be highly nonlinear. In short, a linear estimator does not necessarily imply a linear relationship.

[20] Although as before, one can treat the random x-values as fixed once they are realized in the data, and as before, one is then limited to generalizing only to joint probability distributions with the exact same x-values.

roughly be interpreted as the “equivalent number of parameters” (Ruppert et al. 2003: Sect. 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**. One can also think of the trace as qualifying “*optimism* of the residual sum of squares as an estimate of the out-of-sample prediction error” (Janson et al. 2015: 3)[21] As already noted several times, when more degrees of freedom are used (other things equal), in-sample fit will provide an increasingly unjustified, optimistic impression of out-of-sample performance.[22]

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

$$df_{\mathbf{S}} = 2\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{S}\mathbf{S}^T). \tag{1.14}$$

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. 2009: Sect. 5.4.1). Consequently, for linear estimators we use the trace of **S** as the smoother degrees of freedom.

Unfortunately, there are complications. When there is model selection, more degrees of freedom are being used than the number of non-redundant regression parameters in the final model chosen. This is no less true when tuning parameters are employed. It makes intuitive sense that tuning requires data snooping, even if automated, and degrees of freedom are spent in the process.

Efron’s early work on prediction errors (1986) allows us to takes a step back to reformulate the issues. Effective degrees of freedom used boils down to how well the data are fit in training data compared to how well the data are fit in test data. Other things equal, the larger the gap, the larger the effective degrees of freedom used. Drawing from Efron (1986), Kaufman and Rosset (2014) and Janson and colleagues (2015), the effective degrees of freedom can be defined as

$$\mathrm{EDF} = \mathrm{E}\left[\sum_{i=1}^{N}(y_i^* - \hat{y}_i)^2\right], \tag{1.15}$$

where y_i^* is a realized y-value in test data, and $\hat{y}_i$ is a fitted value computed from the training data. The vector of x-values for case i does not change from realization to realization. Thus, one imagines two, fixed-x realizations of the response for each

[21] Emphasis in the original.

[22] The residual degrees of freedom can then be computed by subtraction (see also Green and Silverman 1994: Sect. 3.3.4).

case. One is included in the training data and used to construct a fitted value. The other is included in the test data. The effective degrees of freedom is the expectation of the summed (over N), squared disparities between the two. The greater the average squared disparities between the fitted values from the training data and the new, realized values of Y, the greater the EDF. The EDF captures how much the degrees of freedom used by the fitting procedure by itself inflates the quality of the fit.

When the fitted values are constructed from a procedure with IID finite variance disturbances, as discussed in Sect. 1.6.1, Eq. 1.15 becomes

$$\text{EDF} = \frac{1}{\sigma^2}\sum_{i=1}^{N}\text{Cov}(y_i, \hat{y}_i). \tag{1.16}$$

The covariance for each case i is defined over realizations of Y with the predictor values fixed, and σ^2 is the variance of the disturbances as usual. Equation 1.16 is a standardized representation of similarity between the realized values of Y and the fitted values of Y. The greater the standardized linear association between the two, the larger the effective degrees of freedom.

In practice, neither definition is operational. But there are important special cases for which estimates of the EDF can be obtained. One of the most useful is when the estimator for the fitted values is linear (e.g., for a smoother matrix **S**). However, current thinking about the EDF appears to be limited to the fixed-x case, whereas statistical learning usually conceives both Y and X as random variables. How to formulate the EDF with random X is apparently unsolved. Indeed, the concept of EDF might usefully be abandoned and replaced by formulations for unjustified optimism. In a very wide set of circumstance, this could be addressed with training and test data, as suggested above.

1.6.7 Basis Functions

Another consideration in thinking about the effective degrees of freedom is that the procedures discussed in subsequent chapters commonly do not work directly with the given set of predictors. Rather, the design matrix in a level I or level II analysis can be comprised of linear basis expansions of X. Linear basis expansions allow for a more flexible fitting function, typically by increasing the dimensionality of the design matrix. A set of p predictors becomes a set of predictors much greater than p. This can make the fitted values 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 in a linear fashion as follows:

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X). \tag{1.17}$$

There are M transformations of X, which can include the untransformed predictor, and typically a leading column of 1s is included (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.

One common transformation employs polynomial terms such as $1, x, x^2, x^3$. Each term does not have to be a linear transformation of x, but the transformations are *combined* in a linear fashion. Then, Eq. 1.17 takes the form

$$f(X) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3. \tag{1.18}$$

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.

$$f(Z) = \beta_0 + \beta_1(I[z > 5]) + \beta_2(I[z > 8|z > 5]) + \beta_3(I[z < 2]). \tag{1.19}$$

As before, fitting by least squares leads to a conventional hat matrix from which the fitted values may be constructed.[23]

Equation 1.17 can be generalized so that $p > 1$ predictors may be included:

$$f(X) = \sum_{j=1}^{p} \sum_{m=1}^{M_j} \beta_{jm} h_{jm}(X). \tag{1.20}$$

There are p predictors, each denoted by j, and each with its own M_j transformations. All of the transformations for all predictors, each with its weight β_{jm}, are combined in a linear fashion. For example, one could combine Eqs. 1.18 and 1.19 with both X and Z as predictors. It is also possible, and even common in some forms of machine learning, to define *each* basis function as a complicated function of two or more predictors. For example, recall that the usual cross-product matrix so central to linear regression is $\mathbf{X}^T\mathbf{X}$. As we will see later, "kernels" broadly based on $\mathbf{XX}^T$ can be constructed that serve is very effective linear basis expansions.

Linear basis expansions are no less central to many forms of classification. Figure 1.10 provides a visual sense of how. Decisions boundaries are essentially fitted values from some procedure that separate one class from another, and can then be used to decide in which class a new case belongs. In Fig. 1.10, there are two classes

[23] The symbol I denotes an indicator function. The result is equal to 1 if the argument in brackets is true and equal to 0 if the argument in brackets is false. The 1s and 0s constitute an indicator variable. Sometimes indicator variables are called a dummy variables.

represented by either a red circle or a blue circle. There are two predictors, X_1 and X_2. Figure 1.10 is a 3-dimensional scatterplot.

The upper plot shows in 2-D predictor space a linear decision boundary. All cases falling above the linear boundary are classified as red circles, because red circles are the majority. All cases falling below the linear boundary are classified as blue circles because blue circles are the majority. The linear decision boundary produces three classification errors. There is one blue circle above the decision boundary and two red circles below the decision boundary. Separation between the two classes is imperfect, and in this illustration, no linear decision boundary can separate the two classes perfectly. However, also shown is a nonlinear decision boundary that can. The trick would be to find transformations of the two predictors from which such a decision boundary could be constructed.

Sometimes there is an effective way to proceed. The lower plot in Fig. 1.10 shows the same binary outcomes in 3-D space. A third dimension has been added. The two curved arrows show how the locations for two illustrative points are moved. As just noted, new dimensions can result from transformations when there is a basis expansion. Here, the three transformation functions are shown as h_1, h_2 and h_3. Within this 3-D predictor space, all of the blue circles are toward the front of the figure, and all of the red circles are toward the back of the figure. The plane shown is a *linear* decision boundary that leads to perfect separation. By adding a dimension, perfect separation can be obtained, and one can work in a more congenial linear world. In 3-D, one has in principle an easier classification problem. Then if one wishes, the 3-D predictor space can be projected back to the 2-D predictor space to view the results as a function of the two original predictors. Back in 2-D predictor space, the decision boundary can then be nonlinear, often very nonlinear.

But if Eq. 1.20 is essentially multiple regression, where does statistical learning come in? The answer is that statistical learning procedures often "invent" their own linear basis expansions. That is, the linear basis expansions are inductively

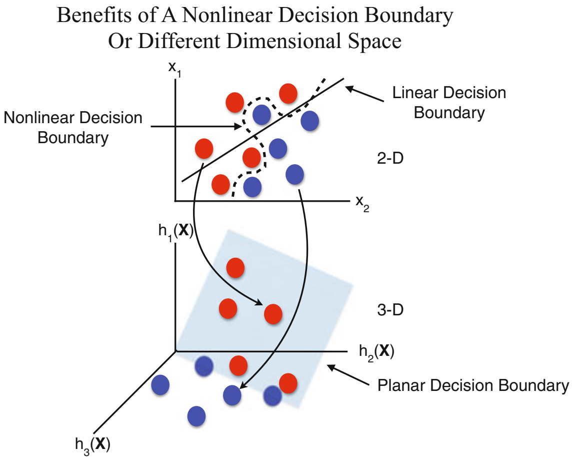

Benefits of A Nonlinear Decision Boundary
Or Different Dimensional Space
x1
Nonlinear Decision Boundary
Linear Decision Boundary
2-D
x2
h1(X)
3-D
h2(X)
Planar Decision Boundary
h3(X)

Fig. 1.10 Reductions in classification errors under linear basis expansions (The *red filled circles* and the *blue filled circles* represent different classes. The *top figure* shows how classification errors can be reduced with a nonlinear decision boundary. The *bottom figure* shows how classification errors can be reduced by including a third predictor dimension.)

constructed as a product of how the algorithmic procedure “learns.” Alternatively, a data analyst may provide the algorithm with far too many basis expansions terms, sometimes more terms than there are observations, and the algorithm decides inductively which are really needed.

Figure 1.11 provides an illustration. The observational units are all 50 states each year from 1978 to 1998, for a total of 1000 observations. For each state each year, the homicide rate and the number of executions for capital crimes were recorded. Data such as these have been central in debates about the deterrent value of the death penalty (Nagin and Pepper 2012).

Executions lagged by one year is on the horizontal axis. It is the only predictor and is included as a single vector of values. No position is taken by the data analyst about the nature of its relationship with the response; the values “as is” are input to the algorithm. The homicide rate per 1000 people is on the vertical axis. The blue, dashed line shows fitted values centered at zero. They are arrived at inductively though a linear basis expansion of the number of executions. The residuals around the fitted values are shown with small blue dots. Error bands around the fitted values are shown light blue. For reasons that will be discussed in the next chapter, the error bands only capture the variability in the fitted values; they are not confidence intervals. Still, if the error bands are used, one has a level II regression analysis.

Within a level I perspective, 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) and in this region of the figure, the data are very sparse.

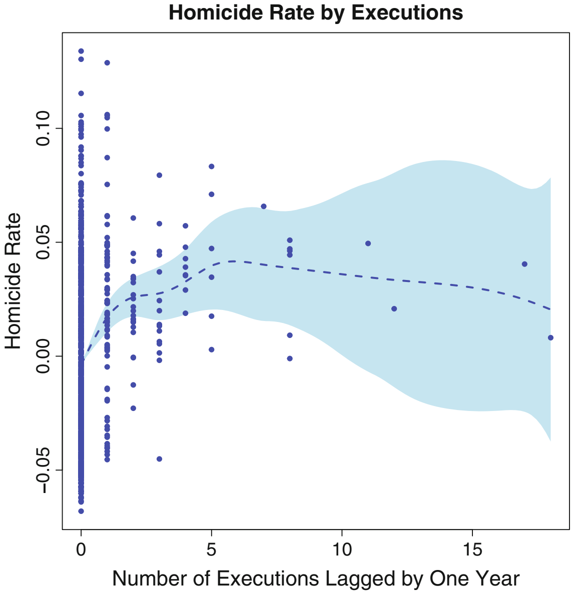


Fig. 1.11 The homicide rate per 1,000 as a function of the number of executions (The homicide rate is on vertical axis, and the number of executions one year earlier is on the horizontal axis. The *broken line* represents the fitted values, and the *light blue region* shows the error bands.)

When there are 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 association for 99 % of the data. When a given state in a given year executes six or more individuals, the relationship appears to turn negative. With more executions, there are fewer homicides one year later. But there are almost no data supporting the change in sign, and from a level II perspective, the error bands around that portion of the curve show that the relationship could easily be flat and even positive.[24] 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 2005.)

Figure 1.11 provides a visualization of how a response of great policy interest and a single predictor of great policy interest are related. There is no model in the conventional regression sense. The fitted values shown were arrived at inductively by a statistical learning algorithm. Had a linear mean function been imposed, the few influential points to the far right of the figure would have produced a negative regression coefficient. One might incorrectly conclude that on the average, there is evidence for the deterrent effect of executions. In practice, of course, the potential role of confounders would need to be considered.

In summary, linear basis expansions can be an important, and even essential, feature of statistical learning. Statistical learning algorithms can be seen as instruments in service of finding linear basis expansions that facilitate prediction. Where the algorithms differ is in exactly how they do that.

1.6.8 The Curse of Dimensionality

Linear basis expansions increase the dimensionality of a dataset. As just described, this is often a good thing. In this era of "Big Data" it is also increasingly common to have access to data not just with a very large number of observations, but with a very large number of variables. For example, the IRS might merge its own records with records from Health and Human Services and the Department of Labor. Both the number of observations (N) and number of dimensions (p) could be enormous. Except for data management concerns, one might assume that bigger data are always better than smaller data. But it is not that simple.

One common and vexing complication is called "the curse of dimensionality." The number of variables exceeds the number of observations that can be effectively exploited. Figure 1.12 shows two especially common difficulties that can arise in practice. The cube at the top illustrates that as the number of dimensions increases linearly, the volume of the resulting space increases as a power function of the number of dimensions. Hence, the 3 by 3 square has 9 units of space to fill with data, whereas

[24] To properly employ a level II framework, lots of hard thought would be necessary. For example, are the observations realized independently as the joint probability distribution approach requires? And if not, then what?

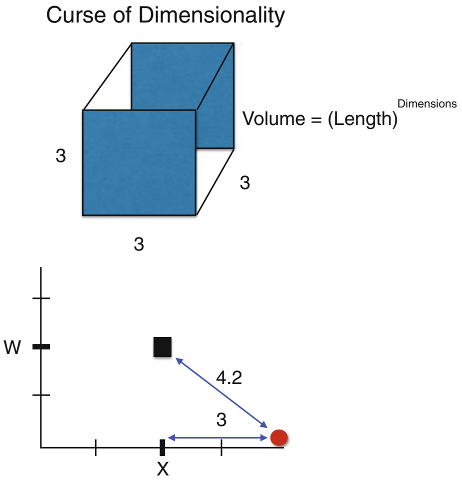


Fig. 1.12 Two consequences of the curse of dimensionality (The *top figure* shows how volume increases as a power function, the *bottom figure* shows how observations move farther away from the *center*.)

the 3 by 3 by 3 cube has 27 units of space to fill with data. The result for a dataset with a certain number of observations is that the distribution of the observations can become very sparse. The space is less adequately covered, so that the sample size per unit of space decreases. A data analysis one might like to do can become impractical. In particular, key regions of a nonlinear $f(\mathbf{X})$ may be very poorly estimated for lack of sufficient data.

Unless more observations can be obtained, some simplifications need to be imposed by the data analyst. A popular approach is to make the function some linear combination of the predictors. Sometimes that can improve the quality of the fitted values, and sometimes that can make the quality worse. This is another manifestation of the bias–variance tradeoff. Alternatively, one can try to reduce the dimensionality of the data using model selection procedures, shrinkage, an incomplete Cholesky decomposition, or principle components. But each of these comes with their own challenges. For example, if principle components analysis is used, one must determine how many of the principle components to include, which introduces a form of model selection.

The bottom figure illustrates a second complication. With an increase in the number of dimensions, the data move farther from the center of the space. For a very large p, the data can be concentrated toward the edges of the space. In the figure, a distance of 3 units in one dimension can become a distance of 4.2 in two dimensions. Thus, a hypercube with 5 sides of 3 units each has a maximum Euclidian distance of 6.7 units, whereas a hypercube with 10 sides of 3 units each has a maximum Euclidian distance of 22.3 units. The problem for a data analysis is that the region toward the center of the space becomes especially sparse. In that region, it will be very difficult to estimate effectively a response surface, especially if it is complicated. Once again, the data analyst has to simplify how the estimation is undertaken or reduce the number of dimensions.

In short, higher dimensional data can be very useful when there are more associations in the data that can be exploited. But at least ideally, a large p comes with a large N. If not, what may look like a blessing can actually be a curse.

1.7 Statistical Learning in Context

Data analysis, whatever the tools, takes place in a professional setting that can influence how the analysis is undertaken. Far more is involved than formal technique. Although we cannot consider these matters at any length, a few brief observations are probably worth mention.

- *It is Important to Keep Humans in the Loop* — With the increasing processing power of computers, petabytes of storage, efficient algorithms, and a rapidly expanding statistical toolbox, there are strong incentives to delegate data analyses to machines. At least in the medium term, this is a huge mistake. Humans introduce a host of critical value judgements, intuitions and context that computers cannot. Worse than a statistical bridge to nowhere is a statistical bridge to the wrong place. A better formulation is a structured computer–human partnership on which there is already interesting work in progress (Michelucci and Dickinson 2016).
- *Sometimes There is No Good Solution* — The moment one leaves textbook examples behind, there is a risk that problems with the data and/or the statistical procedures available will be insurmountable. That risk is to be expected in the real world of empirical research. There is no shame in answering an empirical question with "I can't tell." There is shame in manufacturing results for appearance's sake. Assume-and-proceed statistical practice can be a telling example. In later chapters, we will come upon unsolved problems in statistical learning where, in the words of Shakespeare's Falstaff, "The better part of valor is discretion" (Henry the Fourth, Part 1, Act 5, Scene 4).
- *The Audience Matters* — Results that are difficult to interpret in subject matter terms, no matter how good the statistical performance, are often of little use. This will sometimes lead to another kind of tradeoff. Algorithmic procedures that perform very well by various technical criteria may stumble when the time comes to convey what the results mean. Important features of the data analysis may be lost. It will sometimes be useful, therefore, to relax the technical performance criteria a bit in order to get results that effectively inform substantive or policy matters. One implication is that an effective data analysis is best done with an understanding of who will want to use the results and the technical background they bring. It can also be important to anticipate preconceptions that that might make it difficult to "hear" what the data analysis has to say. For example, it can be very difficult to get certain academic disciplines to accept the results from algorithmic procedures because those disciplines are so steeped in models.
- *Decisions That Can Be Affected* — Knowing your audience can also mean knowing what decisions might be influenced by the results of a data analysis. Simply put,

if one's goal is to bring information from a data analysis to bear on real decisions, the data analysis must be situated within the decision-making setting. This can mean making sure that the inputs and outputs are those that decision-makers deem relevant and that the details of the algorithmic procedures comport well with decision-maker needs. For example, if forecasting errors lead to asymmetric losses, asymmetric costs should be built into the algorithmic procedure.

- *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, although it is important to keep up with important developments, the newest are not necessarily important.
- *Software That Makes No Difference (or is actually worse)* — The hype can apply to software as well. While this edition is being written, the world is buzzing with talk of "data mining," "big data" and "analytics." Not surprisingly, there are in response a substantial number of software purveyors claiming to offer the very latest and very best tools, which perform substantially better than the competition. *Caveat Emptor*. Often, information on how the software runs is proprietary and no real competitive benchmarks are provided. Much like for the Wizard of Oz, there may be little behind a slick user interface. That is one reason why in this book we exclusively use the programming language R. It is free, so there are no sales incentives. The source code can be downloaded. If one wants to make the effort, it is possible to determine if anyone is hiding the ball. And with access to the source code, changes and enhancements in particular procedures can be written.
- *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. A closely related point is that a substantial fraction of the time invested in a given data analysis will be spent cleaning up the data and getting it into the requisite format. These tasks can require substantial skill only tangentially related to conventional statistical expertise.
- *The Role of Subject-Matter Expertise* — Subject-matter expertise can be very important in the following:

  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; and
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.

Demonstrations and Exercises

The demonstrations and exercises in the book emphasize data analysis, not the formalities of mathematical statistics. The goal is to provide practice in learning from data. The demonstrations and exercises for this chapter provide a bit of practice doing regression analyses by examining conditional distributions without the aid of conventional linear regression. It is an effort to get back to first principles unfiltered by least squares regression. Another goal is to show how data snooping can lead to misleading results. Commands in R are shown in italics. However, as already noted several times, R and the procedures in R are moving targets. What runs now may not run later, although there will almost certainly be procedures available that can serve as adequate substitutes. Often, examples of relevant code in R be found in the empirical applications provided in each chapter.

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 *summary()* take a look at some summary statistics for the data frame. Note that there are some missing data and that all of the variables are numeric.
2. Using *pairs()*, construct of a scatterplot matrix including all of the variables in the dataset. These will all be joint (bivarate) distributions. Describe the relationships between each pair of variables. Are there associations? Do they look linear? Are there outliers?
3. Using *boxplot()*, construct separate side-by-side boxplots for ozone concentrations conditioning on month and ozone concentrations conditioning on day. Does the ozone distribution vary by month of the year? In what way?
4. Construct a three-dimensional scatterplot with ozone concentrations as the response and temperature and wind speed as predictors. This will be a joint distribution. Try using *cloud()* from the *lattice* package. There are lots of slick options. What patterns can you make out? Now repeat the graphing but condition on month. What patterns do you see now? (For ease of readability, you can make the variable month a factor with each level named. For really fancy plotting, have a look at the library *ggplot2*.)

5. From the *graphics* library, 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?

    (a) 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?
    (b) How do the patterns differ across the conditioning scatter plots? What does that tell you about interaction effects: how do the relationship between ozone concentrations and temperature differ for different wind speeds?

6. Construct an indicator variable for missing data for the variable Ozone. (Using *is.na()* is a good way.) Applying *table()*, 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? (If you want to percentage the table, *prop.table()* is a good way.)
7. Write out the conventional 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.
8. Implement your regression model in R using *lm()* and examine the results. Look at the regression diagnostics using *plot()*. What do the four plots tell you about your model? 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?

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 how test data can help.

1. Construct the training data as follows. For your predictor: *x = rep(1:30, times = 5)*. This will give you 150 observations with values 1 through 30. For your response: *y = rnorm(150)*. This will give you 150 random draws from the standard normal distribution. As such, they are on the average independent of x. This is the same as letting $y = 0 + 0x + \varepsilon$, which is nature's data generation process.
2. Plot the response against the predictor and describe what you see. Is what you see consistent with how the data were generated?
3. Apply a bivariate regression using *lm()* Describe what overall conclusions you draw from the output. The linear functional form is the "smoothest" possible relationship between a response and a predictor.
4. Repeat the linear regression with the predictor as a factor. Apply the same R code as before but use *as.factor(x)* instead of x. This is a linear basis expansion of x. The set of indicator variables for x (one for each value of x) when used as predictors, leads to the "roughest" possible relationship between a response and a predictor. (Technically, you are now doing a multiple regression.) Each value

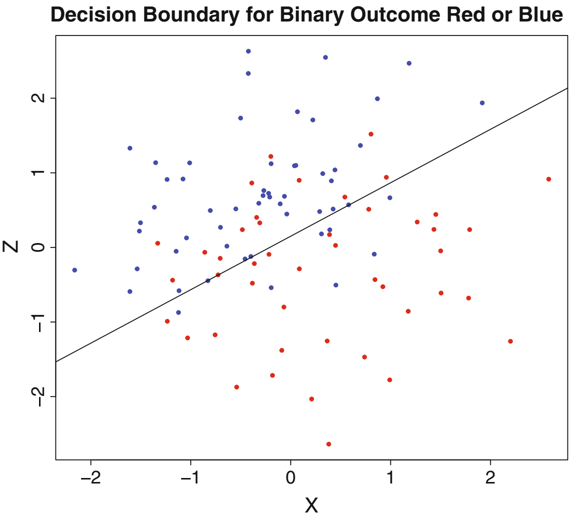


Fig. 1.13 For predictors X and Z, and a binary response coded as *blue* or *red*, an overlaid decision boundary derived from a logistic regression (N = 100: 45 *reds* and 55 *blues*.)

of the predictor can have its own estimate of the conditional mean. (In this case, you know that those conditional means are 0.0 in nature's "generative process.") Compare the R^2 and the adjusted R^2 from the *lm()* output and to the output from #3. What can you say about overfitting. Is there evidence of it here?

5. Construct 1/0 indicator variables for x-indicator variables whose t-values are greater than 1.64. (The *ifelse()* command is a good way to do this.) Apply *lm()* again including only these indicator variables as predictors. What do you find? (By chance, it is possible — but unlikely — that there is still nothing that is "statistically significant." If so, go back to step #1 and regenerate the data. Then pick up at step #3.)
6. Construct test data by repeating step #1. Because x is treated as fixed, you only need to regenerate y. Regress the new y on the subset of indicator variables you used in the previous step. What do you find? The results illustrate the important role of test data.

Set 3

The purpose of this exercise is to get you thinking about decision boundaries for classification problems. Figure 1.13 shows a decision boundary for a binary response coded as red or blue. The predictors are X and Z. The overlaid straight line is a decision boundary based on a logistic regression and values of X and Z for which response odds are $.5/(1 - .5) = 1.0$.

1. Should the observations above the decision be classified as blue or red? Why?
2. Should the observations below the decision be classified as blue or red? Why?
3. Suppose there were an observation with a z-value of 1 and an x-value of -1, but with an unknown response. What would be your best guess: red or blue? Why?

4. Suppose there were an observation with a z-value of -1.5 and an x-value of .5, but with an unknown response. What would be your best guess: red or blue? Why?
5. Why do you think the decision boundary was located at odds of 1.0?
6. How many red observations are misclassified? (For purposes of this exercise, points that seem to fall right on the decision boundary should not be considered classification errors. They are a little above or a little below, but you cannot really tell from the plot.)
7. How many blue observations are misclassified? (For purposes of this exercise, points that seem to fall right on the decision boundary should not be considered classification errors. They are a little above or a little below, but you cannot really tell from the plot.)
8. What fraction of the blue observations is classified correctly?
9. What fraction of the red observations is classified correctly?
10. Which outcome is classified more accurately?
11. What fraction of all of the observations is classified correctly?