Chapter 11
WHY MODEL?

Do not worry. No more chapter introductions around the effect of your looking up on other people's looking up. We squeezed that example well beyond what seemed possible. In Part II of this book, most examples involve real data. The data sets can be downloaded from the book's web site.

Part I was mostly conceptual. Calculations were kept to a minimum, and could be carried out by hand. In contrast, the material described in Part II requires the use of computers to fit regression models, such as linear and logistic models. Because this book cannot provide a detailed introduction to regression techniques, we assume that readers have a basic understanding and working knowledge of these commonly used models. Our web site provides links to computer code in R, SAS, Stata, and Python to replicate the analyses described in the text. The CODE margin notes specify the portion of the code that is relevant to the analysis described in the text.

This chapter describes the differences between the nonparametric estimators used in Part I and the parametric (model-based) estimators used in Part II. It also reviews the concept of smoothing and, briefly, the bias-variance trade-off involved in any modeling decision. The chapter motivates the need for models in data analysis, regardless of whether the analytic goal is causal inference or, say, prediction. We will take a break from causal considerations until the next chapter. Please bear in mind that the statistical literature on modeling is vast; this chapter can only highlight some of the key issues.

11.1 Data cannot speak for themselves

Consider a study population of 16 individuals infected with the human immunodeficiency virus (HIV). Unlike in Part I of this book, we will not view these individuals as representatives of 1 billion individuals identical to them. Rather, these are just 16 individuals randomly sampled from a large, possibly hypothetical super-population: the target population.

At the start of the study each individual receives a certain level of a treatment A (antiretroviral therapy), which is maintained during the study. At the end of the study, a continuous outcome Y (CD4 cell count, in cells/mm^3) is measured in all individuals. We wish to consistently estimate the mean of Y among individuals with treatment level $A = a$ in the population from which the 16 individuals were randomly sampled. That is, the *estimand* is the unknown population parameter $\mathrm{E}[Y|A = a]$.

As defined in Chapter 10, an *estimator* $\widehat{\mathrm{E}}[Y|A = a]$ of $\mathrm{E}[Y|A = a]$ is some function of the data that is used to estimate the unknown population parameter. Informally, a consistent estimator $\widehat{\mathrm{E}}[Y|A = a]$ meets the requirement that "the larger the sample size, the closer the estimate to the population value $\mathrm{E}[Y|A = a]$." Two examples of possible estimators $\widehat{\mathrm{E}}[Y|A = a]$ are (i) the sample average of Y among those receiving $A = a$, and (ii) the value of the first observation in the dataset that happens to have the value $A = a$. The sample average of Y among those receiving $A = a$ is a consistent estimator of the population mean; the value of the first observation with $A = a$ is not. In practice we require all estimators to be consistent, and therefore we use the sample average to estimate the population mean.

See Chapter 10 for a rigorous definition of a consistent estimator.

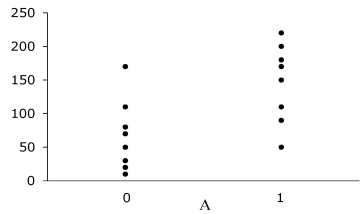

Figure 11.1

Suppose treatment A is a dichotomous variable with two possible values: no treatment ($A = 0$) and treatment ($A = 1$). Half of the individuals were treated ($A = 1$). Figure 11.1 is a scatter plot that displays each of the 16 individuals as a dot. The height of the dot indicates the value of the individual's outcome Y. The 8 treated individuals are placed along the column $A = 1$, and the 8 untreated along the column $A = 0$. As defined in Chapter 10, an *estimate* of the mean of Y among individuals with level $A = a$ in the population is the numerical result of applying the estimator—in our case, the sample average—to a particular data set.

Our estimate of the population mean in the treated is the sample average 146.25 for those with $A = 1$, and our estimate of the population mean in the untreated is the sample average 67.50 in those with $A = 0$. Under exchangeability of the treated and the untreated, the difference $146.25 - 67.50$ would be interpreted as an estimate of the average causal effect of treatment A on the outcome Y in the target population. However, this chapter is not about making causal inferences. Our current goal is simply to motivate the need for models when trying to estimate population quantities like the mean $\mathrm{E}[Y|A = a]$, irrespective of whether the estimates do or do not have a causal interpretation.

Now suppose treatment A is a polytomous variable that can take 4 possible values: no treatment ($A = 1$), low-dose treatment ($A = 2$), medium-dose treatment ($A = 3$), and high-dose treatment ($A = 4$). A quarter of the individuals received each treatment level. Figure 11.2 displays the outcome value for the 16 individuals in the study population. To estimate the population means in the 4 groups defined by treatment level, we compute the corresponding sample averages. The estimates are 70.0, 80.0, 117.5, and 195.0 for $A = 1$, $A = 2$, $A = 3$, and $A = 4$, respectively.

CODE: Program 11.1

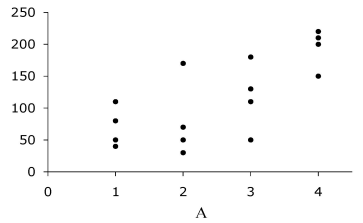

Figure 11.2

Figures 11.1 and 11.2 depict examples of discrete (categorical) variables with 2 and 4 categories, respectively. Because the number of study individuals is fixed at 16, the number of individuals per category decreases as the number of categories increases. The sample average in each category is still an exactly unbiased estimator of the corresponding population mean, but the probability that the sample average is close to the corresponding population mean decreases as the number of individuals in each category decreases. The length of the 95% confidence intervals (see Chapter 10) for the category-specific means will be greater for the data in Figure 11.2 than for the data in Figure 11.1.

Finally, suppose that A represents the dose of treatment in mg/day, and that it takes integer values from 0 to 100 mg. Figure 11.3 displays the outcome value for each of the 16 individuals. Because the number of possible values of treatment is much greater than the number of individuals in the study, there are many values of A that no individual received. For example, there are no individuals with treatment dose $A = 90$ in the study population.

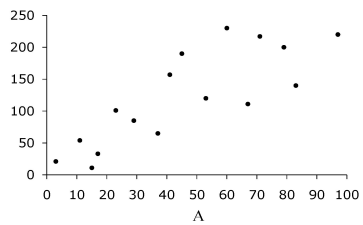

Figure 11.3

This creates a problem: how can we estimate the mean of the outcome Y among individuals with treatment level $A = 90$ in the target population? The estimator we used for the data in Figures 11.1 and 11.2—the treatment-specific sample average—is undefined for treatment levels for which there are zero individuals in Figure 11.3. If treatment A were a truly continuous variable, then the sample average would be undefined for nearly all treatment levels. (A continuous variable A can be viewed as a categorical variable with an uncountably infinite number of categories.)

The above description shows that we cannot always let the data "speak for themselves" to obtain a meaningful estimate. Rather, we often need to supplement the data with a model, as we describe in the next section.

11.2 Parametric estimators of the conditional mean

We want to estimate the mean of Y among individuals with treatment level $A = 90$, i.e., $\mathrm{E}[Y|A = 90]$, from the data in Figure 11.3. Suppose we expect the mean of Y among individuals with treatment level $A = 90$ to lie between the mean among individuals with $A = 80$ and the mean among individuals with $A = 100$. In fact, suppose we knew that the treatment-specific population mean of Y is a linear function of the value of treatment A throughout the range of A. More precisely, we know that the mean of Y, $\mathrm{E}[Y|A]$, increases (or decreases) from some value θ_0 for $A = 0$ by θ_1 units per unit of A. Or, more compactly,

$$\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$$

More generally, the restriction on the shape of the relation is known as the *functional form* and, by some authors, as the *dose-response curve*. We do not use the latter term because it suggests that the dose of treatment causally effects the response, which could be false in the presence of confounding.

This equation is a restriction on the shape of conditional mean function $\mathrm{E}[Y|A]$. This particular restriction is referred to as a *linear mean model*, and the quantities θ_0 and θ_1 are referred to as the *parameters of the model*. Models that describe the conditional mean function in terms of a finite number of parameters are referred to as parametric conditional mean models. In our example, the parameters θ_0 and θ_1 define a straight line that crosses (intercepts) the vertical axis at θ_0 and that has a slope θ_1. That is, the model specifies that all conditional mean functions are straight lines, though their intercepts and slopes may vary.

We are now ready to combine the data in Figure 11.3 with our parametric mean model to estimate $\mathrm{E}[Y|A = a]$ for all values a from 0 to 100. The first step is to obtain estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ of the parameters θ_0 and θ_1. The second step is to use these estimates to estimate the mean of Y for any value $A = a$. For example, to estimate the mean of Y among individuals with treatment level $A = 90$, we use the expression $\widehat{\mathrm{E}}[Y|A = 90] = \hat{\theta}_0 + 90\hat{\theta}_1$. The estimate $\widehat{\mathrm{E}}[Y|A]$ for each individual is referred to as the *predicted value*.

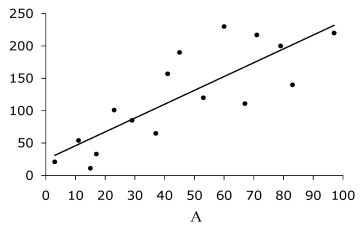

Figure 11.4

An exactly unbiased estimator of θ_0 and θ_1 can be obtained by the method of *ordinary least squares*. A nontechnical motivation of the method follows. Consider all possible candidate straight lines for Figure 11.3, each of them with a different combination of values of intercept θ_0 and slope θ_1. For each candidate line, one can calculate the vertical distance from each dot to the line (the *residual*), square each of those 16 residuals, and then sum the 16 squared residuals. The line for which the sum is the smallest is the "least squares" line, and the parameter values $\hat{\theta}_0$ and $\hat{\theta}_1$ of this "least squares" line are the "least squares" estimates. The values $\hat{\theta}_0$ and $\hat{\theta}_1$ can be easily computed using linear algebra, as described in any statistics textbook.

CODE: Program 11.2
Under the assumption that the variance of the residuals does not depend on A (homoscedasticity), the Wald 95% confidence intervals are $(-21.2, 70.3)$ for θ_0, $(1.28, 2.99)$ for θ_1, and $(172.1, 261.6)$ for $\mathrm{E}[Y|A = 90]$.

In our example, the parameter estimates are $\hat{\theta}_0 = 24.55$ and $\hat{\theta}_1 = 2.14$, which define the straight line shown in Figure 11.4. The predicted mean of Y among individuals with treatment level $A = 90$ is therefore $\widehat{\mathrm{E}}[Y|A = 90] = 24.55 + 90 \times 2.14 = 216.9$. Because ordinary least squares estimation uses all data points to find the best line, the mean of Y in the group $A = a$, i.e., $\mathrm{E}[Y|A = a]$, is estimated by borrowing information from individuals who have values of treatment A not equal to a.

So what is a model? A model is defined by an a priori restriction on the joint distribution of the data. Our linear conditional mean model says that the conditional mean function $\mathrm{E}[Y|A]$ is a straight line, which restricts its shape. For example, the model restricts the mean of Y for $A = 90$ to be between the mean of Y for $A = 80$ and the mean of Y for $A = 100$. This restriction is encoded by the parameters θ_0 and θ_1. A parametric conditional mean model

is, through its a priori restrictions, adding information to compensate for the lack of sufficient information in the data.

Parametric estimators—those based on a parametric conditional mean model—allow us to estimate quantities that cannot be estimated otherwise, e.g., the mean of Y among individuals in the target population with treatment level $A = 90$ when no such individuals exist in the study population. But this is not a free lunch. When using a parametric model, the inferences are correct only if the restrictions encoded in the model are correct, i.e. if the model is correctly specified. Thus model-based causal inference—to which a large fraction of the remainder of this book is devoted—relies on the condition of (approximately) *no model misspecification*. Because parametric models are rarely, if ever, perfectly specified, a certain degree of model misspecification is almost always expected. This can be at least partially rectified by using nonparametric estimators, which we describe in the next section.

11.3 Nonparametric estimators of the conditional mean

Let us return to the data in Figure 11.1. Treatment A is dichotomous and we want to consistently estimate the mean of Y in the treated $\mathrm{E}[Y|A = 1]$ and in the untreated $\mathrm{E}[Y|A = 0]$. Suppose we have become so enamored with models that we decide to use one to estimate these two quantities. Again we proposed a linear model

$$\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$$

where $\mathrm{E}[Y|A = 0] = \theta_0 + 0 \times \theta_1 = \theta_0$ and $\mathrm{E}[Y|A = 1] = \theta_0 + 1 \times \theta_1 = \theta_0 + \theta_1$. We use the least squares method to obtain estimates of the parameters θ_0 and θ_1. These estimates are $\hat{\theta}_0 = 67.5$ and $\hat{\theta}_1 = 78.75$. We therefore estimate $\widehat{\mathrm{E}}[Y|A = 0] = 67.5$ and $\widehat{\mathrm{E}}[Y|A = 1] = 146.25$. Note that our model-based estimates of the mean of Y are identical to the sample averages we calculated in Section 11.1. This is not a coincidence but an expected finding.

CODE: Program 11.2

In this book we define "model" as an a priori mathematical restriction on the possible states of nature (Robins, Greenland 1986). Part I was entitled "Causal inference without models" because it only described saturated models.

Let us take a second look at the model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ with a dichotomous treatment A. If we rewrite the model as $\mathrm{E}[Y|A = 1] = \mathrm{E}[Y|A = 0] + \theta_1$, we see that the model simply states that the mean in the treated $\mathrm{E}[Y|A = 1]$ is equal to the mean in the untreated $\mathrm{E}[Y|A = 0]$ plus a quantity θ_1, where θ_1 may be negative, positive or zero. But this statement is of course always true! The model imposes no restrictions whatsoever on the values of $\mathrm{E}[Y|A = 1]$ and $\mathrm{E}[Y|A = 0]$. Therefore $\mathrm{E}[Y|A = a] = \theta_0 + \theta_1 A$ with a dichotomous treatment A is not a model because it lets the data speak for themselves, just like the sample average does. "Models" which do not impose restrictions on the distribution of the data are *saturated models*. Because they formally look like models even if they do not fit our definition of model, saturated models are ordinarily referred to as models too.

Generally, the model is saturated whenever the number of parameters in a conditional mean model is equal to the number of unknown conditional means in the population. For example, the linear model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ has two parameters and, when A is dichotomous, there exist two unknown conditional means: the means $\mathrm{E}[Y|A = 1]$ and $\mathrm{E}[Y|A = 0]$. Since the values of the two parameters are not restricted by the model, neither are the values of the means. As a contrast, consider the data in Figure 11.3 where A can take values from 0 to 100. The linear model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ has two parameters but estimates 101 quantities, i.e., $\mathrm{E}[Y|A = 0], \mathrm{E}[Y|A = 1], ..., \mathrm{E}[Y|A = 100]$. The only hope

A saturated model has the same number of unknowns on both sides of the equal sign.

Fine Point 11.1

Fisher consistency. Our definition of a nonparametric estimator in the main text coincides with what is known in statistics as a *Fisher consistent estimator* (Fisher 1922). That is, an estimator of a population quantity that, when calculated using the entire population rather than a sample, yields the true value of the population parameter. By definition, a Fisher consistent estimator lacks any model restrictions but, as discussed in the text, a Fisher consistent estimate may not exist for many population quantities. Technically, Fisher consistent estimators, when they exist, are the nonparametric maximum likelihood estimators of population quantities under a saturated model.

In the statistical literature, the term nonparametric estimator is sometimes used to refer to estimators that are not Fisher consistent but that impose very weak restrictions, such as kernel regression models. See Technical Point 11.1 for details.

for unbiasedly estimating 101 quantities with these two parameters is to be fortunate to have all 101 means $\mathrm{E}[Y|A = a]$ lie along a straight line. When a model has only a few parameters but it is used to estimate many population quantities, we say that the model is *parsimonious*.

Identifiability assumptions are the assumptions that we have to make to compute the parameter even if we had an infinite amount of data. Modeling assumptions are the additional assumptions that we have to make to estimate the parameter because we do not have an infinite amount of data. Formally, identifiability assumptions make the parameter a unique function of the joint distribution of the observed data.

Here we define nonparametric estimators of the conditional mean function as those that produce estimates from the data without any a priori restrictions on the conditional mean function (see Fine Point 11.1 for a more rigorous definition). An example of a nonparametric estimator of the population mean $\mathrm{E}[Y|A = a]$ for a dichotomous treatment is its empirical version, the sample average or, equivalently, the saturated model described in this section. When A is discrete with 100 levels and no individual in the sample has $A = 90$, no nonparametric estimator of $\mathrm{E}[Y|A = 90]$ exists. All methods for causal inference that we described in Part I of this book—standardization, IP weighting, stratification, matching—were based on nonparametric estimators of population quantities under a saturated model because they did not impose any a priori restrictions on the value of the effect estimates. In contrast, most methods for causal inference described in Part II of this book rely on estimators that are parametric estimators of some part of the distribution of the data. Parametric estimation is one approach used to borrow information when, as is often the case, data are unable to speak for themselves.

11.4 Smoothing

Caution: Often the term "linear" is used with two different meanings. A model is *linear* when it is expressed as a linear combination of parameters and functions of the variables, even if the latter are nonlinear functions (e.g., higher powers or logarithms) of the covariates.

Consider again the data in Figure 11.3 and the linear model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$. The parameter θ_1 is the difference in mean outcome per unit of treatment dose A. Because θ_1 is a single number, the model specifies that the difference in mean outcome Y per unit of treatment A must be constant throughout the entire range of A, i.e., the model requires the conditional mean outcome to follow a straight line as a function of treatment dose A. Figure 11.4 shows the best-fitting straight line.

But one can imagine situations in which the difference in mean outcome is larger for a one-unit change at low doses of treatment, and smaller for a one-unit change at high doses. This would be the case if, once the treatment dose reaches certain level, higher doses have an increasingly small effect. Under this scenario, the model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ is incorrect. However, linear models can be made more flexible.

For example, suppose we fit the model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$, where $A^2 = A \times A$ is A-squared, to the data in Figure 11.3. This is still referred to as a linear model because the conditional mean is expressed as a linear combination, i.e., as the sum of the products of each covariate (A and A^2) with its associated coefficient (the parameters θ_1 and θ_2) plus an intercept (θ_0). However, whenever θ_2 is not zero, (θ_0, θ_1, θ_2) now define a curve—a parabola—rather than a straight line. We refer to θ_1 as the parameter for the linear term A, and to θ_2 as the parameter for the quadratic term A^2.

The curve under the 3-parameter linear model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ can be found via ordinary least squares estimation applied to the data in Figure 11.3. The estimated curve is shown in Figure 11.5. The parameter estimates are $\hat{\theta}_0 = -7.41$, $\hat{\theta}_1 = 4.11$, and $\hat{\theta}_2 = -0.02$. The predicted mean of Y among individuals with treatment level $A = 90$ is obtained from the expression $\widehat{\mathrm{E}}[Y|A = 90] = \hat{\theta}_0 + 90\hat{\theta}_1 + 90 \times 90\hat{\theta}_2 = 197.1$.

CODE: Program 11.3
Under the homoscedasticity assumption, the Wald 95% confidence interval for $\widehat{\mathrm{E}}[Y|A = 90]$ is $(142.8, 251.5)$.

We could keep adding parameters for a cubic term ($\theta_3 A^3$), a quartic term ($\theta_4 A^4$)... until we reach a 15th-degree term ($\theta_{15} A^{15}$). At that point the number of parameters in our model equals the number of data points (individuals). The shape of the curve would change as the number of parameters increases. In general, the more parameters in the model, the more inflection points will appear.

That is, the curve generally becomes more "wiggly," or less smooth, as the number of parameters increase. A linear model with 2 parameters—a straight line—is the smoothest model. A linear model with as many parameters as data points is the least smooth model because it has as many possible inflection points as data points. In fact, such model interpolates the data, i.e., each data point in the sample lies on the estimated conditional mean function.

Often modeling can be viewed as a procedure to transform noisy data into more or less smooth curves. This smoothing occurs because the model borrows information from many data points to predict the outcome value at a particular combination of values of the covariates. The smoothing results from $\mathrm{E}[Y|A = a]$ being estimated by borrowing information from individuals with A not equal to a. All parametric estimators incorporate some degree of smoothing.

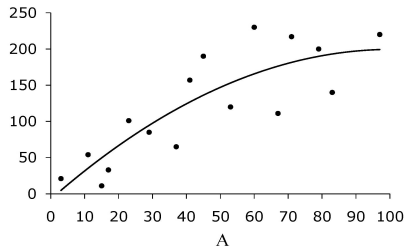

Figure 11.5

The degree of smoothing depends on how much information is borrowed across individuals. The 2-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ estimates $\mathrm{E}[Y|A = 90]$ by borrowing information from all individuals in the study population to find the least squares straight line. A model with as many parameters as individuals does not borrow any information to estimate $\mathrm{E}[Y|A]$ at the values of A that occur in the data, though it borrows information (by interpolation) for values of A that do not occur in the data.

We used a model for continuous outcomes as an example. The same reasoning applies to models for dichotomous outcomes such as logistic models (see Technical Point 11.1)

Intermediate degrees of smoothing can be achieved by using an intermediate number of parameters or, more generally, by restricting the number of individuals that contribute to the estimation. For example, to estimate $\mathrm{E}[Y|A = 90]$ we could decide to fit a 2-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ restricted to individuals with treatment doses between 80 and 100. That is, we would only borrow information from individuals in a 10-unit window of $A = 90$. The wider the window around $A = 90$, the more smoothing would be achieved.

In our simplistic examples above, all models included a single covariate (with either a single parameter for A or two parameters for A and A^2) so that the curves can be represented on a two-dimensional book page. In realistic applications, models often include many different covariates so that the curves are really hyperdimensional surfaces. Regardless of the dimensionality of the problem, the concept of smoothing remains invariant: the fewer parameters in the model, the smoother the prediction (response) surface will be.

Fine Point 11.2

Model dimensionality and the relation between frequentist and Bayesian intervals. In frequentist statistical inference, probability is defined as frequency. In Bayesian inference, probability is defined as degree-of-belief—a concept very different from probability as frequency (de Finetti 1972). Chapter 10 described the confidence intervals used in frequentist statistical inference. Bayesian statistical inference uses credible intervals, which have a more natural interpretation: A Bayesian 95% credible interval means that, given the observed data, "there is a 95% probability that the estimand is in the interval". However, in part because of the requirement to specify the investigators' degree of belief, Bayesian inference is less commonly used than frequentist inference.

Interestingly, in simple, low-dimensional parametric models with large sample sizes, 95% Bayesian credible intervals are also 95% frequentist confidence intervals, whereas in high-dimensional or nonparametric models, a Bayesian 95% credible interval may not be a 95% confidence interval as it may trap the estimand much less than 95% of the time. The underlying reason for these results is that Bayesian inference requires the specification of a prior distribution for all unknown parameters. In low-dimensional parametric models the information in the data swamps that contained in reasonable priors. As a result, inference is relatively insensitive to the particular prior distribution selected. However, this is no longer the case in high-dimensional models. Therefore if the true parameter values that generated the data are unlikely under the chosen prior distribution, the center of Bayes credible interval will be pulled away from the true parameters and towards the parameter values given the greatest probability under the prior.

11.5 The bias-variance trade-off

In previous sections we have used the 16 individuals in Figure 11.3 to estimate the mean outcome Y among people receiving a treatment dose of $A = 90$ in the target population, $\mathrm{E}[Y|A = 90]$. Since nobody in the study population received $A = 90$, we could not let the data speak for themselves. So we combined the data with a linear model. The estimate $\widehat{\mathrm{E}}[Y|A = 90]$ varied with the model. Under the 2-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$, the estimate was 216.9 (95% confidence interval: 172.1, 261.6). Under the 3-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$, the estimate was 197.1 (95% confidence interval: 142.8, 251.5). We used two different parametric models that yielded two different estimates. Which one is better? Is 216.9 or 197.1 closer to the mean in the target population?

If the relation is truly curvilinear, then the estimate from the 2-parameter model will be biased because this model assumes a straight line. On the other hand, if the relation is truly a straight line, then the estimates from both models will be valid. This is so because the 3-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ is correctly specified whether the relation follows a straight line (in which case $\theta_2 = 0$) or a parabolic curve (in which case $\theta_2 \neq 0$). One safe strategy would be to use the 3-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ rather than the 2-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$. Because the 3-parameter model is correctly specified under both a straight line and a parabolic curve, it is less likely to be biased. In general, the larger the number of parameters in the model, the fewer restrictions the model imposes; the less smooth the model, the more protection afforded against bias from model misspecification.

Although less smooth models may yield a less biased estimate, they also result in a larger variance, i.e., wider 95% confidence intervals around the estimate. For example, the estimated 95% confidence interval around $\widehat{\mathrm{E}}[Y|A = 90]$ was much wider when we used the 3-parameter model than when we used the 2-parameter model. However, when the estimate $\widehat{\mathrm{E}}[Y|A = 90]$ based on the 2-parameter model is biased, the standard (nominal) 95% confidence interval

Fine Point 11.2 discusses the implications of model dimensionality for frequentist and Bayesian intervals.

is not calibrated, i.e., it does not cover the true parameter $\mathrm{E}[Y|A = 90]$ 95% of the time.

This bias-variance trade-off is at the heart of many data analyses. Investigators using models need to decide whether some protection against bias—by, say, adding more parameters to the model—is worth the cost in terms of variance. Though some formal procedures exist to aid these decisions, in practice many investigators decide which model to use based on criteria like tradition, interpretability of the parameters, and software availability. In this book we will usually assume that our parametric models are correctly specified. This is an unrealistic assumption, but it allows us to focus on the problems that are specific to causal analyses. Model misspecification is, after all, a problem that can arise in any sort of data analysis, regardless of whether the estimates are endowed with a causal interpretation. In practice, careful investigators will always question the validity of their models and will conduct alternative analysis under different model specifications that are compatible with existing expert knowledge. Their goal is to assess the sensitivity of their estimates to model specification.

We are now ready to describe the use of models for causal inference.

Technical Point 11.1

A taxonomy of commonly used models. The main text describes linear conditional mean models of the form $\mathrm{E}[Y|X] = \theta X \equiv \sum_{i=0}^{p} \theta_i X_i$ where X is a vector of covariates $X_0, X_1, ...X_p$ with $X_0 = 1$ for all n individuals. These models are a subset of larger class of conditional mean models (McCullagh and Nelder, 1989; McCulloch, Searle, and Neuhaus, 2008) which have two components: a linear functional form or predictor $\sum_{i=0}^{p} \theta_i X_i$ and a link function $g\{\cdot\}$ such that $g\{\mathrm{E}[Y|X]\} = \sum_{i=1}^{p} \theta_i X_i$.

The linear conditional mean models described in the main text uses the identity link function. Conditional mean models for outcomes with strictly positive values (e.g., counts, the numerator of incidence rates) often use the log link function to ensure that all predicted values will be greater than zero, i.e., $\log\{\mathrm{E}[Y|X]\} = \sum_{i=0}^{p} \theta_i X_i$ so $\mathrm{E}[Y|X] = \exp\left(\sum_{i=0}^{p} \theta_i X_i\right)$. Conditional mean models for dichotomous outcomes (i.e., those that only take values 0 and 1) often use a logit link i.e., $\log\left\{\frac{\mathrm{E}[Y|\mathbf{X}]}{1-\mathrm{E}[Y|\mathbf{X}]}\right\} = \sum_{i=0}^{p} \theta_i X_i$, so that $\mathrm{E}[Y|X] = \mathrm{expit}\left(\sum_{i=0}^{p} \theta_i X_i\right)$. This link ensures that all predicted values will be greater than 0 and less than 1. Conditional mean models that use the logit function, referred to as logistic regression models, are widely used in this book. For these links (referred to as canonical links) we can estimate θ by maximum likelihood under a normal working model for the identity link, a Poisson working model for the log link, and a logistic regression model for the logit link. These estimates are consistent for θ as long as the conditional mean model for $\mathrm{E}[Y|X]$ is correct. Generalized estimating equation (GEE) models, often used to deal with repeated measures, are a further example of a conditional mean model (Liang and Zeger, 1986).

Conditional mean models only specify a parametric form for $\mathrm{E}[Y|X]$ but do not otherwise restrict the distribution of $Y|X$ or the marginal distribution of X. Therefore, when X or Y are continuous, a parametric conditional mean model is a semiparametric model for the joint distribution of the data (X, Y) because parts of the distribution are modeled parametrically whereas others are left unrestricted. The model is semiparametric because the set of all unrestricted components of the joint distribution cannot be represented by a finite number of parameters.

Conditional mean models themselves can be generalized by relaxing the assumption that $\mathrm{E}[Y|X]$ takes a parametric form. For example, a kernel regression model does not impose a specific functional form on $\mathrm{E}[Y|X]$ but rather estimates $\mathrm{E}[Y|X = x]$ for any x by $\sum_{i=1}^{n} w_h(x - X_i) Y_i / \sum_{i=1}^{n} w_h(x - X_i)$ where $w_h(z)$ is a positive function, known as a kernel function, that attains its maximum value at $z = 0$ and decreases to 0 as $|z|$ gets large at a rate that depends on the parameter h subscripting w. As another example, generalized additive models (GAMs) replace the linear combination $\sum_{i=0}^{p} \theta_i X_i$ of a conditional mean model by a sum of smooth functions $\sum_{i=0}^{p} f_i(X_i)$. The model can be estimated using a backfitting algorithm with $f_i(\cdot)$ estimated at iteration k by, e.g., kernel regression (Hastie and Tibshirani 1990).

In the text we discuss smoothing with parametric models which specify an a priori functional form for $\mathrm{E}[Y|X = x]$, such as a parabola. In estimating $\mathrm{E}[Y|X = x]$, the model may borrow information from values of X that are far from x. In contrast, kernel regression models do not specify an a priori functional form and borrow information only from values of X near to x when estimating $\mathrm{E}[Y|X = x]$. A kernel regression model is an example of a "non-parametric" regression model. This use of the term "nonparametric" differs from our previous usage. Our nonparametric estimators of $\mathrm{E}[Y|X = x]$ only used those individuals for whom X equalled x exactly; no information was borrowed even from close neighbors. Here "nonparametric" estimators of $\mathrm{E}[Y|X = x]$ use individuals with values of X near to x. How near is controlled by a smoothing parameter referred to as the bandwidth h.