Chapter 19
What You Can Learn from Wrong Causal Models

Richard A. Berk, Lawrence Brown, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao

Abstract It is common for social science researchers to provide estimates of causal effects from regression models imposed on observational data. The many problems with such work are well documented and widely known. The usual response is to claim, with little real evidence, that the causal model is close enough to the "truth" that sufficiently accurate causal effects can be estimated. In this chapter, a more circumspect approach is taken. We assume that the causal model is a substantial distance from the truth and then consider what can be learned nevertheless. To that end, we distinguish between how nature generated the data, a "true" model representing how this was accomplished, and a working model that is imposed on the data. The working model will typically be "wrong." Nevertheless, unbiased or asymptotically unbiased estimates from parametric, semiparametric, and nonparametric working models can often be obtained in concert with appropriate statistical tests and confidence intervals. However, the estimates are not of the regression parameters typically assumed. Estimates of causal effects are not provided. Correlation is not causation. Nor is partial correlation, even when dressed up as regression coefficients. However, we argue that insights about causal effects do not require estimates of causal effects. We also discuss what can be learned when our alternative approach is not persuasive.

> *What I am trying to say throughout the book is that "doing" the models consists largely in thinking about the kind of model one wants and can justify in the light of the ideas whose validity one is prepared to take responsibility for* (Duncan 1975: viii).

Introduction

Perhaps the most widely known aphorism in the discipline of statistics is "All models are wrong, but some are useful" (Box 1979: 202). There are several possible reasons for its celebrity. For an enterprise highly dependent on models, it asserts that no model can be correct. At an epistemological level, all models are by design abstract simplifications of some reality. Without simplification, scientific progress can be very difficult. At a practical level, actually more consistent with the setting in which the aphorism was coined, even if models could be correct in principle, researchers can never have

R.A. Berk (✉) • L. Brown • E. George • E. Pitkin • M. Traskin • K. Zhang • L. Zhao
Departments of Statistics and Criminology, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: berkr@sas.upenn.edu

S.L. Morgan (ed.), *Handbook of Causal Analysis for Social Research*,
Handbooks of Sociology and Social Research, DOI 10.1007/978-94-007-6094-3_19,

them. There are all of the well-known discrepancies between what a model formally requires and what the data can deliver. Still, for many the take-home message is that a wrong model may be of no special concern as long as it is useful. This is surely comforting.

But there are complications. Far more clarity is needed on what is meant by "wrong." In addition, usefulness is multidimensional. A model may be useful along one dimension and worse than useless along another. There is often controversy, moreover, because one person's use can be another person's abuse. At a deeper level, there are many kinds of models. The social sciences are dominated by causal models. Other types of models can have different strengths and weaknesses from which it can follow different bundles of uses.

In the pages ahead, we confront these issues within the causal modeling tradition long popular in the social sciences (Goldberger and Duncan 1973; Duncan 1975; Greene 2003). From this perspective, a causal model is a quantitative theory of how the data were generated in which a statistical formalization for random variables is combined with a causal account derived from subject-matter knowledge (Kaplan 2009: Sect. 10.5). But, models of data generation actually do not have to be causal. Causal mechanisms can be seen as an interpretive overlay. We proceed in this fashion, but for continuity with past methodological discussions in the social sciences, on occasion we retain the term "causal model" when referring to social science practice.

We begin with conventional linear regression before examining semiparametric and nonparametric formulations. Readers will find no fundamental quarrel with Box's view that all models are wrong. But we argue that the best response is not to simply soldier on or try to patch things up around the margins. The best response is to rethink the enterprise. With that done, we will see that some wrong regression models can be useful, but not in the ways often favored by conventional practice. This is a general lesson that can apply beyond the particular models we discuss.

Sections "Regression Analysis Defined" and "A Regression Causal Model Defined" are devoted to clearing away some conceptual clutter. Regression analysis is defined along with what it means for a regression model to be "right." Section "A Regression Causal Model Defined" elaborates on what it means for a regression model to be "wrong." Section "What Can be Properly Estimated from a Working Regression Model?" addresses the properties of estimates from a conventional linear regression when the model is wrong. Sections "Nonparametric Regression" and "Summary and Conclusions" broaden the range of regression models considered to include semiparametric and nonparametric specifications. Section "Summary and Conclusions" offers some broad conclusions.

Regression Analysis Defined

Cook and Weisberg (1999: 27) offer a definition of regression analysis that corresponds well with much statistical thinking: "[to understand] as far as possible with the available data how the conditional distribution of the response y varies across subpopulations determined by the possible values of the predictor or predictors." The entire conditional distribution of y is considered, although in practice, attention is usually directed at the conditional mean and/or conditional variance.[1]

The definition may be interpreted in two ways. Regression analysis can be solely a descriptive tool for the data on hand (Berk 2003). The data are treated as a population. A bit more will be said about this conception shortly. Alternatively, regression analysis applied to the data on hand can be used for estimating properties of conditional distributions in the population from which the data came or for

[1]This definition can apply to categorical response variables, manifest or latent response variables, and response variables whose conditional distributions are related to one another. So, for example, the generalized linear model is covered as well as multiple equation models.

estimating properties of conditional distributions implied by the processes by which nature generated the data. This is the more common and more ambitious perspective that will be emphasized in this chapter.

There is nothing in either conception about hypothesis tests, confidence intervals, or causal inference, and often researchers want more than description or estimation. They want to properly represent the role of uncertainty in any estimates using confidence intervals and/or statistical tests. They want to make causal inferences as well; how would the response variable's distribution change if one or more explanatory variables were manipulated independently of all other explanatory variables? These are all reasonable aspirations.

At what point does one need to think about a model? There is no mention of a model, let alone a causal model, in the Cook and Weisberg definition of regression analysis. Models become relevant when one attempts to draw inferences beyond the data on hand. The issues raised can be subtle.

A Regression Causal Model Defined

We begin with the ubiquitous linear regression model that can be written as

$$y_i|\mathbf{x}_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \tag{19.1}$$

where y_i is the value of the response for case i, $\mathbf{x}_i$ is a vector of p predictor variables for case i, there are p regression coefficients and an intercept β_0, and $\varepsilon_i \sim NIID(0, \sigma^2)$. This is a statistical formalization for the conditional distribution of the random variable $\mathbf{Y}$ given $\mathbf{X}$. There is nothing causal.

To arrive at a causal model, practitioners introduce a "response schedule," at least implicitly, that "...says how one variable would respond, if you intervened and manipulated other variables..." (Freedman 2009: 87). Response schedules can be seen as mathematical, counterfactual formulations of causal effects (Berk 2003: 84–90). A causal model marries a statistical formalization with a response schedule so that causal interpretations can be made (Kaplan 2009: Sect. 10.8).

In this instance, one might say that nature sets for each case the values of the predictors. These values are fixed in a statistical sense. Nature then, in effect, multiplies each predictor times its regression coefficient and adds those products to the value of the intercept. Finally, nature draws for each case independently a disturbance from a normal distribution with a mean of zero and some variance equal to σ^2, and adds that disturbance to the linear combination of predictors. The result is y_i.[2]

Nature can repeat the last two steps a limitless number of times leading in principle to a population of all possible realizations of the data. For each case, the values of the regression coefficients, intercept,

[2]The properties of ε_i can be formulated other ways. For example, the causes of $\mathbf{Y}$ can be organized into two groups: regressors with large causal effects and regressors with small causal effects. In an early treatment that is representative, Hanushek and Jackson (1977: 82) distinguish similarly between "important" predictors and others. The variables with small causal effects are taken to be far more numerous than the variables with large causal effects and to be independent of one another. Nature sets the values of the many small causal variables too, but in the aggregate, the result is disturbances that are *effectively* independent of the causal variables with large effects. For formal results, this account is too imprecise. Rather, it is common to assume "sparsity." Sparsity requires that some predictors have true regression coefficients exactly equal to zero (not just small) after conditioning on all other predictors in the model. Then as a theoretical matter, a common question is whether a given model selection procedure will correctly identify which predictors have such regression coefficients (e.g., Leeb and Pötscher 2008b). The "real-world" sources of the disturbances are not addressed.

and predictors do not change. Nor does the disturbance distribution. What changes over realizations of the data is the value of the randomly drawn disturbance that is added to the linear combination of the predictors. The result is a realized conditional distribution for y_i. One important implication is that the uncertainty in y_i comes exclusively from ε_i. Another important implication is that the distributional properties of the realized ε_i are characterized over a limitless number of independent draws of ε_i.[3]

Suppose for the moment that the data on hand are treated as a population. For example, the data may be a full enumeration of all students enrolled in a given university, the inventory of a large warehouse, or a year's worth of financial transactions from a brokerage firm. For each, interest centers on what can be learned about the data on hand. For example, is a particular university in compliance with Title IX?

For a population, it is then no longer clear what conceptual benefits a data generation model confers. The value of y_i is generated only once, and then the requisite properties of the realized ε_i on which the realized y_i depends are not defined. For example, there is no $E(\varepsilon_i)$ because there is but one realization of the disturbances for a given case i.

Under these circumstances, data analysis can only be descriptive. There can be no statistical inference: estimation, confidence intervals, or statistical tests. Causal inference is also ruled out because the conventional counterfactual framework cannot apply. One cannot work with "potential" (i.e., hypothetical) outcomes because there are none. The data on hand are all that matter.[4]

We will, therefore, proceed in the rest of the chapter assuming the data are a sample in the sense that the values observed could have been different, and one could in principle see many independent realizations. At this point in the discussion, the data need not be a probability sample from a real, finite population. Rather, the data are what Thompson calls a "model-based sample" (Thompson 2002: Sect. 2.7). The model represents the mechanisms through which the data are "sampled" (i.e., generated) by nature.

Under model-based sampling, statistical inference is conventionally undertaken with respect to the data generation process. The parameter values to be estimated are those employed in that process. Of greatest interest are usually inferences about the regression coefficients employed by nature when the predictors are linearly combined. But there is sometimes interest in the conditional means of the response as well.

Statistical inference can follow directly from the formal properties of the disturbances. Causal inference depends on how nature sets the values of the predictors. If, for a given predictor, nature could set its values differently and independently of all other predictors, causal effects can at least be defined within the usual potential outcomes (i.e., counterfactual) framework. Equation (19.1) can be used to make causal statements.

[3] In practice, this summary of how nature functions would need to be fleshed out with specific subject-matter knowledge. For example, why does nature work with a linear combination of predictors, and how exactly does it do that? Still, at least a bit of mathematical license (e.g., a limitless number of independent realizations of the data) will be required so that theorems of interest can be proved.

[4] In mathematical statistics, data on hand that might be seen as a population are sometimes treated as a random realization from a population of all possible realizations of the data that nature could generate. Such populations are sometimes called "superpopulations." Although this formulation allows certain mathematical operations to play through, the scientific payoff is obscure unless one has a credible theory for how the superpopulation is generated and why the data to be analyzed are a random realization from that superpopulation. But if one has such a theory, and if it is of the same form as Eq. (19.1), the approach is essentially the same as the one just described.

How Regression Causal Models Can Go Wrong

We now allow for a "working model" that represents what a researcher actually employs with the data. From a working model, a researcher tries to learn about key features of the causal model. It follows that the relationship between the working model and the causal model is critical. Ideally, the two should perform in the same manner with respect to the causal model parameters being estimated. Researchers commonly proceed as if the working model *is* the causal model or that any differences do not materially affect the conclusions reached.

Inferences from a working model can be compromised by two related difficulties. First, the working model is wrong in a very obvious sense if it does not accurately represent credible understandings, built into the causal model, of how the data were generated. To take a simple instance, if one of the predictors nature is supposed to use is not included in the working model, the working model is by construction wrong. Thus, if a credible claim is that educational attainment is a cause of earnings and educational attainment is not in the working model, the working model is wrong. Moreover, if nature is said to linearly employ the log of a given predictor and the working model includes the unlogged form, the working model is by construction wrong.

There are certainly deeper epistemological issues such as whether the idea of a "true model" is an oxymoron. Yet, if there is no such thing as a true model, it is difficult to see how one would act on Duncan's (1975: viii) call to take responsibility for its validity. There is also ample precedent in the social sciences for the idea of a true model: "A coherent relationship between economic and statistical aspects of models seems very desirable in order to reduce the possibility of inconsistent and unclear implications of analyses" (Zellner 1984: 30).

For present purposes, we sidestep such issues. They would take us into difficult territory that is peripheral to the goals of this chapter. What matters for the discussion to follow is simply whether the regression equation estimated is consistent with existing claims of how nature generated the data.

Second, it is sometimes unappreciated that Eq. (19.1) determines the meaning of each regression coefficient. That is, the mathematical expression for any single regression coefficient depends on how all of the predictors and the disturbances are combined. For example, with a different set of predictors, or a nonlinear transformation of any predictor, the mathematical expression for each regression coefficient changes. Thus, if the working model is wrong, so are the regression coefficients. The regression coefficients are by construction not those that nature is supposed to employ.

In both cases, however, there can be an escape clause of sorts. The properties of any regression coefficient estimates depend substantially on whether for the working model each of the disturbances meet the conventional regression assumptions shown in Eq. (19.1), and in particular, whether $E(\varepsilon_i) = 0$. When for the working model, $E(\varepsilon_i) = 0$, the regression coefficients for the causal model's predictors included in the working model can be estimated in an unbiased manner by applying least-squares to the working model.[5]

This is, of course, a well-known property of least-squares regression, and some might claim that when the assumed properties of the disturbances are met, the working model is not wrong. For this chapter, we think the "escape-clause" characterization is more instructive, but in any case, it is usually

[5]Consider a simple example. Suppose for a response variable $\mathbf{Y}$ there are in the causal model two predictors that enter additively: $\mathbf{X}$ and $\log(\mathbf{Z})$. Because this is the correct model, $E(\varepsilon_i) = 0$. Therefore, the disturbances and the regressors are unrelated. Now suppose that the researcher does not know about $\mathbf{Z}$ and it is not included in the working model. If it is still true that $E(\varepsilon_i) = 0$, the working model least-squares regression coefficient for $\mathbf{X}$ will be unbiased. Somewhat different reasoning applies if the researcher mistakenly employs, say, $\mathbf{Z}$ instead of $\log(\mathbf{Z})$. Even if $E(\varepsilon_i) = 0$, the working model least-squares coefficient for $\mathbf{X}$ will be affected unless both $\mathbf{Z}$ and $\log(\mathbf{Z})$ are uncorrelated with $\mathbf{X}$ (i.e., mean independent). Still other reasoning applies if $\mathbf{X}$ is measured imperfectly. Even random measurement errors with a mean of zero imply that $E(\varepsilon_i) \neq 0$. For example, education may be measured in years of schooling. But years of schooling is but a proxy from what may really matters: increases in human capital. Biased estimates follow.

very difficult with observational data to persuasively argue that all variables omitted from the working model are uncorrelated with all the included predictors and/or that the functional forms employed should be treated as those used by nature. It then follows that the disturbances of the working model do not have an expectation of zero. The assumption of a common disturbance variance can be compromised as well. Because the causal model means of Y are systematically underestimated or overestimated, statistical inference for those means and the associated regression coefficients is in serious jeopardy.

Causal inference is also compromised when the working model is wrong. If $E(\varepsilon_i) \neq 0$, the disturbances are confounded with one or more predictors. Nature is assumed to make a clear distinction between the predictors and the disturbances. A wrong working model does not.[6]

In summary, one can make a useful distinction between a causal regression model meant to represent how nature generated the data and a working regression model applied to the data by a researcher. Sometimes researchers proceed as if a working regression model is the same as a causal regression model. At the very least, a strong justification should be provided grounded in the particulars of the research being undertaken. Some researchers, with the assistance of regression diagnostics, proceed as if the correspondence between a regression causal model and a regression working model is close enough. In practice, it is hard to know what "close enough" means, and faith in regression diagnostics can be misplaced (Freedman 2009). Perhaps a better strategy is to think about what can be learned from working models assumed to be substantially wrong.

What Can Be Properly Estimated from a Working Regression Model?

The approach we favor is to apply an alternative to the regression causal model that can be more appropriate for observational data. This alternative is called a "joint probability distribution model." It has much in common with the "correlation model" proposed by Freedman (1981) and is very similar to a "linear approximation" approach formulated by White (1980) that, in turn, has important roots in the work of Huber (1967) and Eicker (1963, 1967). Angrist and Pischke (2009: Sect. 3.1.2) provide very accessible and persuasive motivation. In short, much of what we propose has been around in various forms for quite some time.

Suppose one claims that nature generates the data for each case as a realization from some joint probability distribution composed of random variables $\mathbf{Z}$. That joint probability distribution can be characterized by the usual sorts of parameters such the mean and variance for each variable and the covariances between variables. There is no distinction between predictors and responses. For each case, nature can independently generate a limitless number of independent realizations of the random variables. Some might wish to call the joint probability distribution formulation the "true model."

For the random variables constituting $\mathbf{Z}$, researchers will often distinguish between predictors $\mathbf{X}$ and responses $\mathbf{Y}$. Some of $\mathbf{Z}$ may be ignored because it is not relevant for the substantive or policy issues at hand. Such decisions have nothing to do with how the data were generated. They have everything to do with the preferences of researchers.

The distinction predictors and responses are usually motivated by interest in the conditional distribution of some $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ because the distribution of $\mathbf{Y}$ is thought to change from one $\mathbf{x}$ to another $\mathbf{x}$. A change in its mean $\mu(\mathbf{x})$ is typically the primary concern, and it follows from how

[6]There are statistical procedures, such as instrumental variables, that under ideal conditions can overcome the confounding of predictors and disturbances. These ideal conditions are difficult to meet with observational data. In effect, an auxiliary model is required that has to be right. So, this escape clause too can be hard to exercise.

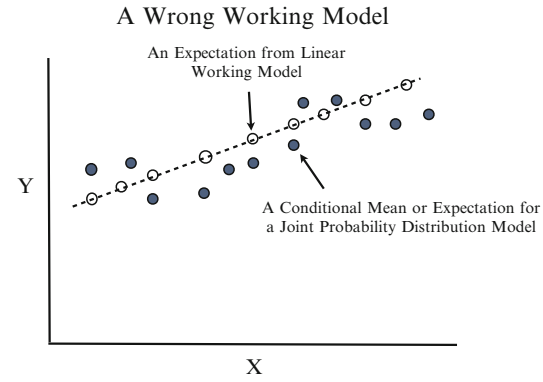

Fig. 19.1 Bias in a linear working model

nature generates the data that $E(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = \mu(\mathbf{x})$.[7] But here, $\mathbf{X}$ is random, so $\mu(\mathbf{x})$ is random. Therefore, researchers might also be interested in $E[\mu(\mathbf{x})] = E(\mathbf{Y})$.[8]

An interest in how the mean of the response varies depending on the values of predictors is shared with conventional regression models. Beyond that common goal, the regression model and the joint probability distribution model part company. First, under the joint probability distribution model, there is no a priori commitment to how the response is related to the predictors and certainly no linearity requirement. Second, the predictors have no special cachet. Among the random variables that nature can generate, the researcher decides to designate some as predictors. Third, there is, therefore, no such thing as an omitted variable. Finally, there is nothing causal whatsoever.

Suppose now that a working model assumes the conventional form of linear regression. The set of conditional means over cases, $\boldsymbol{\mu}$, is assumed to be related to $\mathbf{X}$ by $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. $\mathbf{Y}$ is then taken to be $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $\varepsilon_i \sim NIID(0, \sigma^2)$. One might at this point choose to treat the random predictors as fixed, although then the regression results cannot be generalized beyond the particular x-values in the realized data.

For all of the reasons mentioned earlier, such a working model will usually be wrong. In particular, it is compromised if the alternative joint probability distribution model is credible. For example, it is almost certain that $E(\varepsilon_i) \neq 0$. There is absolutely no guarantee, therefore, that $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. Indeed, the two will likely differ, often substantially. Nature did not use the equivalent of $\mathbf{X}\boldsymbol{\beta}$ to generate the conditional means of the response, but the researcher is proceeding as if nature did.

Figure 19.1 illustrates some potential consequences. For a single fixed regressor, the conditional means from the joint probability distribution model (i.e., the gray-filled circles) are plotted along with the expectation of the conditional means from a linear working regression (i.e., the unfilled circles). With a random regressor, the conditional means from the joint probability distribution become conditional expectations. Note that the figure is a representation of underlying statistical theory, not a conventional scatterplot of data.

As a description, the linear working model shown in Fig. 19.1 provides a good sense of the relationship. However, the linear fit is biased at every predictor value. By definition, when a conditional expectation from the model is not the same as a conditional mean (or expectation) from nature's joint probability distribution, there is bias. As a result, conventional statistical tests and

[7] Implied is that if one denotes the disparities over realizations between any μ_i and its y_i by ε_i, $E(\varepsilon_i) = 0$.

[8] If the predictors are treated as fixed, one cannot formally generalized the results to values of the predictors not found in the data. There is also a problem with forecasts because with fixed predictor values, there is no account for how the new predictor values were generated.

confidence intervals also will not perform as they should. Still, some researchers might find a linear approximation useful, so it is worth considering its properties in more depth.[9]

In matrix notation, $\mathbf{X}$ denotes a full-rank $n \times (p + 1)$ design matrix with a leading column of 1s, and $\mathbf{Y}$ denotes the $n \times 1$ response variable. Both are taken at face value. They are not analyzed as indicators, indices, or proxies for latent constructs.

For the moment, we will proceed conventionally treating $\mathbf{X}$ as fixed. We denote the regression coefficients from the linear working model by $\boldsymbol{\Gamma}$ and the conditional means of the response that follow from $\mathbf{X}\boldsymbol{\Gamma}$ by $\boldsymbol{\nu}$. One can understand $\boldsymbol{\nu}$ as the best linear approximation of $\boldsymbol{\mu}$ by a least-squares criterion.[10]

The vector of working regression coefficients can be estimated in the usual least-squares manner:

$$\hat{\boldsymbol{\Gamma}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{19.2}$$

As one would expect, it is highly unlikely that $\mathrm{E}(\hat{\boldsymbol{\Gamma}}) = \boldsymbol{\beta}$ and $E(\mathbf{X}\hat{\boldsymbol{\Gamma}}) = \boldsymbol{\mu}$. The effort to obtain unbiased estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ for the model $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ stumbles.

But there is more to the story. Consider first the fitted values

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}, \tag{19.3}$$

where $\mathbf{H}$ is the usual hat matrix. Taking the expectation,

$$E(\hat{\mathbf{Y}}) = \mathbf{H}E(\mathbf{Y}). \tag{19.4}$$

Equation (19.4) defines one "target" of the estimation. Just as in Fig. 19.1, $\hat{\mathbf{Y}}$ estimates $E(\hat{\mathbf{Y}}) = \boldsymbol{\nu}$. In short, it can be shown that the conditional means $\boldsymbol{\nu}$ for the working regression model can be estimated in an *unbiased* manner by the usual least squares procedures. When $\mathbf{X}$ is random, it can be shown that the estimates are asymptotically unbiased (Berk et al. 2011).[11] In contrast to usual social science practice, no assumptions are being made about the properties of the disturbances from the working model. For instance, they may be correlated with one or more of the predictors. One implication is that there is no need to disentangle the disturbances from the predictors so that procedures using instrumental variables, for instance, are unnecessary. Indeed, the usual econometric obsession with $E(\varepsilon_i)$ is no longer relevant.

In a similar fashion for fixed $\mathbf{X}$,

$$E(\hat{\boldsymbol{\Gamma}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{Y}). \tag{19.5}$$

Thus, $\hat{\boldsymbol{\Gamma}}$ estimates $E(\hat{\boldsymbol{\Gamma}}) = \boldsymbol{\Gamma}$. The least-squares estimates are unbiased with respect to the working model's regression coefficients $\boldsymbol{\Gamma}$. As before, when $\mathbf{X}$ is random, it can be shown that the estimates are unbiased asymptotically (Berk et al. 2011). Once again, no assumptions are being made about the properties of the working model's disturbances. Researchers should find this quite liberating, but it

[9] If the conditional means of the joint distribution really do have a linear relationship with the predictors in the working model, the linear approximation is no longer an approximation. There is, then, no bias in the least-squares estimates with respect to the joint probability distribution. This is an unrealistic scenario in practice because even if the linear approximation were actually correct, there would be no way to definitively know it. All one has is a realization from the joint probability distribution.

[10] We change the notation for regression model to underscore that we are no longer trying to estimate the "true" conditional means or "true" regression coefficients. Our estimates are for the linear approximation.

[11] All one requires is that $[E(\mathbf{X}^T\mathbf{X})]^{-1}$ and $E(\mathbf{X}^T\mathbf{Y})$ exit. The asymptotics assume that the number of predictors is fixed as the number of observations increases without limit.

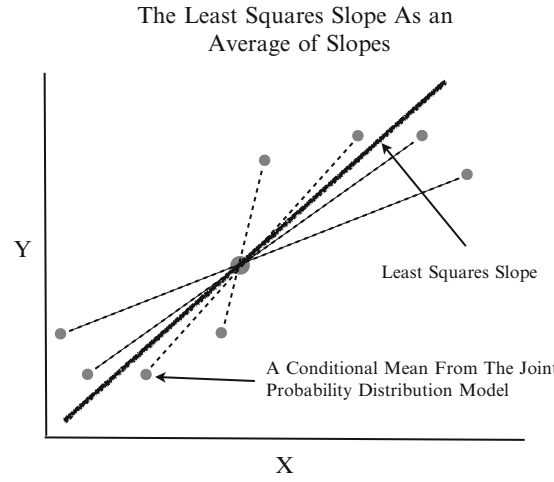

Fig. 19.2 Interpretation of a linear approximation

means that there will be a reconsideration of the meaning and usefulness of some popular regression diagnostics. For example, added variable plots may suggest ways in which the working model can better approximate the conditional means of nature's joint probability distribution, but not whether the working model can become the true model in the usual social science sense.

The regression coefficients from the linear approximation have a handy interpretation. Recall that a slope may be interpreted as the change in the mean of the response as the value of the predictor varies. For a single predictor, the usual estimator for a least squares regression coefficient can be rewritten as

$$\hat{\boldsymbol{\gamma}} = \frac{\sum_i \frac{y_i - \bar{y}}{x_i - \bar{x}}(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}, \tag{19.6}$$

where the fraction in the numerator is the ratio for a given observation of the mean-deviated value of the response and the mean-deviated value of the predictor. The rest of the expression serves as a weight. Observations farther from the mean of the predictor are weighted more heavily.[12]

Figure 19.2 illustrates what is being estimated. Eight of nature's conditional means are shown by the gray-filled circles. The large black circle shows the mean of the predictor and response from nature's joint distribution. The broken lines are the slopes for all pairs of conditional means that also pass through the mean of **Y** and the mean of **X**. Finally, the solid line is the slope of a linear approximation. It is the weighted average of the four other slopes. It is too flat for two of the pairs and too steep for two of the pairs. Both the individual slopes and the average slope are wrong. Still, it is perhaps a useful summary of how **Y** and **X** are related. It carries much the same information as a partial correlation coefficient, but because the original units of **Y** and **X** are retained, it may be more easily informed by subject-matter theory.

The same basic reasoning applies when there is more than one predictor. The main difference is that for any predictor, its values have been adjusted for all other predictors. The covariance adjustments are undertaken *within the linear approximation*. For that approximation, the usual results and interpretations apply. But for the "true" model, they do not. The regression coefficients may be altered too much or too little, and the conceptual parallels to post-stratification no longer hold. It follows that the slopes of the linear approximation cannot properly be given a causal interpretation.

[12]The subscripts i and k differ because the denominator is calculated first as a normalizing constant. Gelman and Park (2008: 3) have an expression that is similar to Eq. (19.6).

The model in which the linear approximation is embedded is not a causal model, and the role of each predictor cannot be separated from the role of the disturbances. In short, it is difficult to know from a linear approximation regression coefficient what would happen if its predictor were manipulated.

Although there is usually no substantive interest in σ^2, one typically needs an estimate of it for conventional, fixed $\mathbf{X}$, standard errors. If $\hat{\sigma}^2$ is obtained using the least-squares approximation of $\boldsymbol{\mu}$, the $\hat{\sigma}^2$ will be estimated incorrectly. The estimates of $\hat{\sigma}^2$ will capture not just the random variation in the disturbances but the disparities between the conditional means from nature and the conditional expectations from the linear approximation. The result for the approximation regression coefficients will be confidence intervals that are wrong for the stated coverage probabilities and statistical tests that are wrong as well. Both results can be misleading but may be acceptable for some researchers.

In summary, if a linear approximation can be descriptively useful, some helpful statistical properties can follow. In particular, unbiased estimates or asymptotically unbiased estimates of the working model's regression coefficients and conditional means may be obtained. This holds even though the disturbances from the working model do not have to meet the usual regression assumptions. Incorrect, but perhaps useful, confidence intervals and statistical tests can directly follow. At the very least, the variance of the estimates can be properly represented. Yet, the regression coefficients cannot properly be given causal interpretations. They have much the same conceptual status as partial correlation coefficients. And correlation, even partial correlation, is not causation.

More on Statistical Tests and Confidence Intervals

Resampling statistical inference is available within a framework in which the data are a realization from a joint probability distribution. All of the variables are random, even those selected to be predictors. When $\mathbf{X}$ is treated as random, recall that least-squares procedures produce for the linear approximation asymptotically unbiased estimates of the regression coefficients and the conditional means of the response. Then, the nonparametric bootstrap in which rows of the data are sampled randomly with replacement, can be used with real data to provide asymptotically appropriate statistical tests and confidence intervals (Mammen 1993). In finite samples of modest size, it is difficult to know how much credibility any inferential claims might have, but the bootstrap can at least provide estimates of the variability of parameter estimates from realization to realization. One can have stability intervals should confidence intervals not be appropriate.

The same bootstrap works if the data are a random sample from a well-defined population (Freedman 1981).[13] There is now no necessary role for nature. Humans generate the data on hand. The estimation targets are the finite population versions of $\boldsymbol{\nu}$ and $\boldsymbol{\Gamma}$.

A Further Fallback Position

Although the joint probability distribution model and the conditional probability distribution model are not as demanding as the linear regression model, they too can be wrong in the sense discussed earlier. Perhaps most fundamentally, one must explain how nature manages to generate independent random realizations of the data from the same joint probability distribution or at least, why it is helpful to think about nature's actions in this manner. How is the science being advanced? For example, a claim that the data for each case are realized independently can be a problem, especially for spatial

[13] The papers by Freedman and Mammen were in a general way anticipated by Fisher in 1924.

or temporal phenomena. Another challenge, if a key goal is causal inference, is to take account of causal processes explicitly in the means by which nature is supposed to have generated the data. One approach might be to envision realizations from conditional distributions defined by different values for the (fixed) causal variables. The response and the covariances would remain random variables.

A second option, discussed extensively elsewhere (Berk 2003), is to give up on the goals of statistical inference and causal inference and focus on description. Much of the empirical work currently undertaken in the social sciences is primarily descriptive, despite causal modeling claims. And good science can begin with good description. The examples presented shortly can be seen as illustrating a descriptive approach if a parametric modeling framework offered does not seem plausible.

Nonparametric Regression

The linear approximation approach has the advantages of simplicity and tractability. But there can be situations in which a nonlinear approximation is preferred. The conditional means of **Y** in a data set may appear to have strong nonlinear relationships with the predictors, and a nonlinear approximation may make more subject-matter sense as well.

There is sometimes accepted substantive theory that dictates the particular nonlinear functions required. One can then proceed much as when the approximation is linear. But, in many applications, there will be no such guidance. Under these conditions, can the data be used to arrive at a reasonable nonlinear approximation? The answer is a qualified yes, which opens the door to semiparametric and nonparametric regression.

Indicator Variable Regression

Assume, as before, that nature generates the data as if by random sampling from the joint probability distribution. Random variables **X** and **Y** are designated by a researcher. For the joint distribution, there is again $\boldsymbol{\mu}$, the conditional means of **Y** given **X**. These are nature's conditional means whose values are to be estimated. So far, there is nothing new.

Rather than assuming a linear function by which **Y** is related to **X**, a more flexible approach is taken. The conditional mean function is not specified. The "true" model now can be written as

$$y_i = \boldsymbol{\mu} + \varepsilon_i, \tag{19.7}$$

where $\boldsymbol{\mu}$ represents, as before, the conditional means from nature's joint probability and ε is a disturbance term about which no assumptions are made. It follows from the definition of a conditional mean that $E(\varepsilon_i) = 0$.

Equation (19.7) is not a statement about how the data were generated. That matter is already resolved in a different fashion. Equation (19.7) is a statement about particular relationships in nature's joint probability distribution. Indeed, it may be easier to think about the disturbances as population-level residuals. Equation (19.7) is not, therefore, a conventional regression model, and causality has no explicit role. Put another way, we are replacing a linear approximation with a potentially nonlinear approximation. The catch is that we don't know what form the nonlinear approximation takes. We need to learn that from the data.

There are a number of ways to empirically proceed. Primarily for didactic purposes, we begin with a variant of the conventional linear regression model in part to introduce some important ideas in a

familiar setting and in part to stay within a regression framework that we will carry through subsequent material. Can linear regression be used to provide good estimates of $\boldsymbol{\mu}$ and the unknown $f(\mathbf{X})$?

Suppose a researcher is, on subject-matter grounds, interested in the phenomenon captured by Eq. (19.7). For ease of discussion, assume that $\mathbf{X}$ is a single, quantitative predictor that for purposes of analysis is treated as fixed. One simple way to construct estimates of $\boldsymbol{\mu}$ is to replace the values of the single predictor with one indicator variable for each observed value of $\mathbf{X}$. Least-squares procedures can then be properly applied to the *multiple* regression specification. The systematic part of that multiple regression is a weighted sum of step functions.[14] Each $\hat{y}_i$ from the regression is an unbiased estimate of μ_i and when paired with the corresponding x_i (i.e., the predictor in its original form), provides a description of how the response is related to the predictor. Often the relationship is shown within a scatterplot format as an interpolation of adjacent conditional means.

The same approach and happy results might seem to apply when there is more than one predictor as long as each is represented by indicator variables in a similar fashion.[15] For example, if there are 20 values for years of age and 10 values for years of education, there are 19 indicator variables for age and nine indicator variables for education. We seem to have a solution.

Unfortunately, we do not. First, if in the realized data any values for the predictors are not present, the conditional means for those values cannot be estimated. For example, suppose in the joint distribution, the predictor age is measured by year from 18 to 60 years old. But suppose that in the realized data, there are no 20-year-olds. The mean of the response variable for 20-year-olds cannot be estimated. Thus, even for the single predictor case, some conditional means may not be estimated, which implies that the estimate of $f(\mathbf{X})$ is incomplete – there are some holes.

Second, because the number of observations with the same predictor values will usually be small (even just one), the distribution of the 1/0 indicator will be highly skewed and, consequently, have very little variance. The result is substantial instability and large standard errors for an estimate of the conditional mean.

Third, to guarantee unbiased estimates of $\boldsymbol{\mu}$, one needs to include indicator variables so that the regression specification is saturated. Even with a modest number of predictors, the number of indicator variables can become unmanageable. If a subset is to be used, which subset?

All three problems are exacerbated by the "curse of dimensionality" (Hastie et al. 2009: 22–26). As the number of predictors increases, the predictor space that must be filled by the data increases very quickly in a multiplicative fashion. In principle, every possible crossing of predictor values requires observations. For example, if there is a single predictor with 10 values, there are 10 locations that need data. If there are two predictors with 10 values each, there is 100 locations that need data. If there are three predictors with 10 values each, there are 1,000 locations that need data. In the same fashion, if there are four predictors, there are 10,000 locations and so on.

One might think that a good solution is to construct the indicator variables over ranges of the predictor values. For example, rather than having an indicator for each year of age, one might have an indicator for age in 5-year intervals. Then, estimates of the conditional means would be obtained for each age *interval*. But grouping the data in this fashion introduces a trade-off between bias and variance. By making the indicator variables more coarse, the variance of each may be increase (i.e., the distribution is more balanced), and each conditional mean will likely be estimated with greater precision. This is good.

[14] This actually is a little tricky. If there is one observation for each x-value and if there is an intercept in the model, one of the indicator variables must be deleted. Otherwise, the predictor cross-product matrix cannot be inverted in the usual manner. The problem disappears if there is no intercept but then the regression coefficients do not have their usual meaning.

[15] Categorical predictors would already be included as one or more indicator variables.

But in exchange, there will likely be a decrease in accuracy. Suppose, for example, that an age indicator is defined for ages 21–25. The observations for each year of age from 21 to 25 get an indicator code of "1." Unless each year of age has the same conditional mean for the response, there will be bias.

The bias-variance tradeoff is quite general. Many popular methods used to reduce the variance of estimates will increase the bias (and vice versa). A key implication to which we will return is that to obtain estimates that are on the average as close as possible to their estimation targets, one should try to minimize the *combined* impact of the mean and the variance. More formally, the goal is to minimize mean squared error in the estimate, which is equal to the sum of the squared bias and the variance. Biased estimates can be desirable if they also have relatively little variance.

Another general point is that indicator variable regression is related to a number of procedures that will generally perform better. In particular, indicator variables represent fixed, disjoint predictor intervals. There are estimation procedures that allow for intervals that can vary in size, sometimes depending on how the response is related to the predictors. The intervals also do not have to be disjoint. A popular and effective illustration is "locally weighted scatterplot smoothing" (LOWESS). An excellent discussion of such matters can be found in Hastie and Tibshirani (1990, Chaps. 2 and 3). In the pages ahead, however, we will follow a different path more directly related to the issues raised in this chapter, more closely linked to conventional regression approaches, and more easily extended recent regression-like advances, such as the LASSO (Tibshirani 1996).

Smoothing Splines

The data are once again a realization from nature's joint probability distribution. As before, both **Y** and **X** are random variables. For nature's joint probability distribution, we impose a new requirement. For the conditional distribution of **Y**|**X**, bounded second derivatives exist over the range of **X**. In that sense, $f(\mathbf{X})$ is smooth. In practice, this is not an especially restrictive assumption because a smooth function can still be highly nonlinear. What we get in return is the ability to more systematically address the bias-variance trade-off. In particular, a penalty term is appended to the usual least-squares procedure. For a single predictor treated as fixed, this leads to

$$\text{PSS}(\hat{f}, \lambda) = \sum_{i=1}^{N} [y_i - \hat{f}(x_i)]^2 + \lambda \int [\hat{f}''(t)]^2 dt. \tag{19.8}$$

PSS stands for penalized sum of squares, which is to be minimized conditional a penalty parameter λ. The first term on the right-hand side is just the usual residual sum of squares. The $\hat{f}(x_i)$ in Eq. (19.8) plays the same role as the $\hat{y}_i$ one would normally expect but is used to emphasize that the requisite function of the predictor is to be determined as part of the minimization process.

The second term imposes a cost for the complexity of the fit. The integral of the second derivatives over **X** defines the complexity penalty. It produces a summary of how sharply the slope of the fitted values changes over the values of the predictor.[16] A larger value means that the $\hat{f}(\mathbf{X})$ is more "rough." A smaller value means that the $\hat{f}(\mathbf{X})$ is more "smooth."

Once the summary measure of roughness is computed, the penalty parameter λ determines the weight given to that penalty in the fitting process. As λ increases, the usual least-squares line is more

[16] In $f''(t)$, the t is just a placeholder because when there is more than one predictor, there can be several sensible ways to represent the fitted values (Hastie et al. 2009: 165–167).

closely approximated. In the limit, no second derivatives are permitted because $\hat{f}(\mathbf{X})$ is a straight line. As λ approaches zero, the fitted values more closely approximate the interpolation results.

The bias-variance trade-off is clearly evident in Eq. (19.8). When λ is larger, the fitted values are forced to be smoother. The likely consequence is more bias and less variance. When λ is smaller, the fitted values are allowed to be rougher. The likely consequence is less bias and more variance.

The value of λ is usually determined empirically. One tries to minimize an estimate of the integrated squared prediction error, which is essentially an out-of-sample sum of squared residuals. However, it is important to apply substantive information as well. If the fitted values are too smooth or too rough, given what is known about the phenomenon, the value of λ should be adjusted accordingly.

Computational strategies for Eq. (19.8), based on B-splines, are discussed in Hastie et al. (2009: 189). They lead to a "smoother matrix" conditional on the value of λ and denoted by $\mathbf{S}_\lambda$ from which fitted values are constructed.[17] Like the usual regression hat matrix, $\mathbf{S}_\lambda$ is $N \times N$. Fitted values are produced in an analogous fashion: $\hat{\mathbf{Y}} = \mathbf{S}_\lambda \mathbf{Y}$. Consequently, the fitted values are a linear combination of the $\mathbf{Y}$, and Eq. 19.8 is one of a class of linear smoothers.[18]

Much as for the indicator variable linear regression discussed earlier, $\hat{\mathbf{Y}}$ can be paired with $\mathbf{X}$ to approximate nature's $f(\mathbf{X})$. As we show next, plots can be very instructive. But the estimation details are tricky. It can be shown that as $N \rightarrow \infty$ and $\lambda \rightarrow 0$, estimates of $f(\mathbf{X})$ converge to nature's $f(\mathbf{X})$. However, in finite samples, bias remains. We are constructing a particular nonlinear (rather than linear) approximation of nature's conditional means.

One might think that, just as in the linear case, the expectation of the nonlinear approximation is being estimated in an unbiased fashion. The estimation target would then be the nonlinear approximation for nature's joint probability distribution, not $\boldsymbol{\mu}$.[19] But there are significant complications. Because the function is nonlinear, the function estimated depends on the particular set of predictor values that are realized. Also, the tuning parameter λ is usually determined from the data. As a result, there is a model selection that also can introduce bias (Berk et al. 2010; Leeb and Pötscher 2006, 2008a, b). These and other factors raise estimation questions that are unresolved and beyond the scope of this chapter. How to proceed in practice will be addressed shortly.

An Example

Figure 19.3 shows a smoothing spline in action. The solid line represents $\hat{f}(\mathbf{X})$, and there is a rug plot along the horizontal axis. For a large American city, the log of the number of homeless individuals in a census tract has been regressed on the log of the proportion of housing units in a tract that are vacant.

The mass of the data falls between values of about 0.02 (i.e., e^{-4}) and 0.14 (i.e., e^{-2}) for the proportion of dwellings that are vacant. With λ determined by the generalized cross-validation statistic, $\hat{f}(\mathbf{X})$ is S shaped. It is essentially flat from proportions near zero to a proportion of about 0.05, then steeply positive up to a proportion of about 0.15, and then flat once again. The average number of homeless in a census tract increases from essentially zero to about 4 (i.e., $e^{1.4}$), but only for vacant dwelling proportions between about 0.05 and 0.15.

---

[17] The trace of the smoother matrix is the "effective degrees of freedom" used by the smoothing procedure, which plays the same role as the model degrees of freedom in conventional regression.

[18] There are many kinds of linear smoothers including local means, local linear fits, and local polynomials that can be employed within kernel functions. The LOWESS procedure (Cleveland 1979) is one popular example. We focus on smoothing splines here because it is a natural extension of least-squares regression, commonly available, and effective in practice. Readers seeking a more extensive treatment of smoothing should consult Hastie and colleagues (2009: Chaps. 3, 5, and 6).

[19] The estimation target is the nonlinear *approximation* within nature's joint probability distribution.

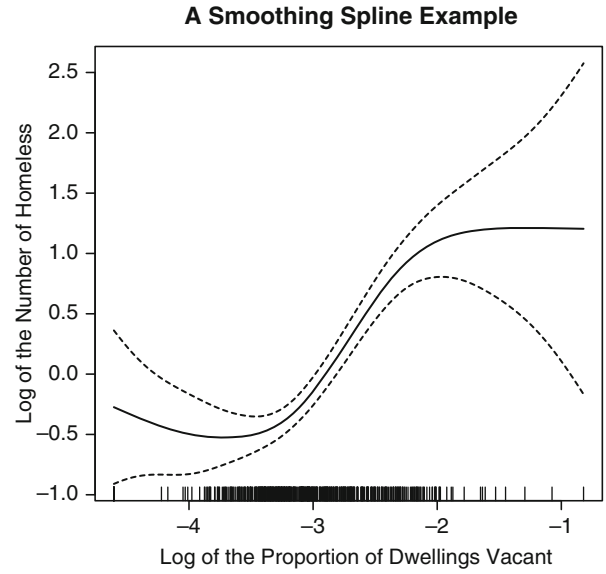

Fig. 19.3 Homelessness as a function of vacant dwellings

Figure 19.3 is not the product of a causal model. Yet, there are perhaps causal interpretations. For example, the S-shaped relationship may represent a tipping point process with an upper constraint. A certain concentration of vacant housing is required before homeless individuals or families begin to move in. Above that concentration, there is a clear signal that there are opportunities for squatters. But there are also constraints that keep the number of homeless in a census tract in check even when a relatively large proportion of the housing is vacant. Perhaps when the homeless are numerous enough to be deemed a nuisance, the police are notified. At the same time, the variation in the fitted values is small so that the tipping phenomenon is not strong. We will return to this issue later.

Also included in Fig. 19.3 is what is commonly promulgated as the point-by-point 95 % confidence interval. The interval widens dramatically with X-values less than about -3.5 and greater than about -2. It is difficult to get a good fix on the functional form in the tails of the predictor where the data are sparse. But to understand what a point-by-point interval really means, we need to consider in somewhat more depth statistical inference for smoothing splines.

Statistical Inference for Smoothing Splines

Equation (19.7) is meant to approximate the relationship between $\boldsymbol{\mu}$ and $\mathbf{X}$. Even with a very large sample, however, the approximation will be imperfect, and the use of penalized regression implies that estimates of the fitted values will be biased. The estimation procedure implicitly trades variance against bias. Moreover, the value of λ was determined empirically, which introduces model selection biases even for estimates of the expectations of the nonlinear approximation (Berk et al. 2010; Leeb and Pötscher 2006, 2008a, b). In short, $\hat{f}(\mathbf{X})$ is surely biased, whether for $f(\mathbf{X})$ or its linear approximation, perhaps substantially, so that conventional statistical tests and confidence intervals do not perform as they should.

One might think that the nonparametric bootstrap would once again be helpful. But there is apparently nothing that can be done about predictor values not in the realized data set and the other sources of bias. Consequently, one would be bootstrapping biased estimates of $f(\mathbf{X})$. Confidence intervals would not have their stated coverage, and test statistics would not produce accurate probabilities under the null hypothesis (e.g., 0.02 might really be 0.18).

The key point is that under current practice, point-by-point confidence intervals are constructed – and the nonparametric bootstrap is certainly a good way – so that they are actually "stability intervals" that capture only the variance, not the bias, in the fitted values. They convey how much the fitted values will likely vary over realizations of the data, but they say little about how often $\boldsymbol{\mu}$ falls within the stability band.

Nevertheless, the intervals shown in Fig. 19.3 are helpful. They suggest that $\hat{f}(\mathbf{X})$ should not be taken very seriously toward the tails of $\mathbf{X}$. What appears to be the absence of a relationship might actually be positive or negative.

Causal Inference for Smoothing Splines

Causal inference for smoothing splines is inherently problematic. There is no causal model within the joint probability distribution framework. And even if Eq. (19.7) were reinterpreted as such, the estimation procedures typically introduce bias. In the end, one cannot isolate the role of the predictor from the role of the disturbances.

The Multivariate Case

Smoothing splines can be effective when there is more than one predictor. The regressors $\mathbf{X}$ become a conventional $n \times (p + 1)$ matrix. In principle $f(\mathbf{X})$ can be more than two dimensional – with two predictors, for instance, it would be a surface not a line. But the curse of dimensioning intrudes once again.

A popular fallback position is to make the $f(\mathbf{X})$ additive. The additive form is familiar, relatively easy to work with, and like additive models more generally, performs surprisingly well in a variety of settings. In this spirit, we proceed with

$$y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, \tag{19.9}$$

where no assumptions need be made about ε_i. The expectation of Eq. (19.9) can be seen as a nonlinear approximation of $\boldsymbol{\mu}$ derived from an additive approximation of $f(\mathbf{X})$.

Equation (19.9) has an intercept represented by α followed by the sum of p functions, one for each predictor. There are no regression coefficients. Their role is absorbed in each predictor's functional form – technically, there can be a limitless number of slopes as the first derivative of the function changes. Therefore, the substantive story for each predictor is primarily in visualizations of various kinds, as it was for one predictor.

The intercept is not identified, but under the assumption that the average of the functions over the data is zero, the intercept is the average of the response variable (Hastie et al. 2009: 298). The mean of Y thus serves as a baseline. This seems to be a harmless constraint, much like the identifying restrictions used in analysis of variance.

We will continue to emphasize quantitative response variables, but Eq. (19.9) can be generalized in the spirit of the generalized linear model (GLM), in which case, it is called the generalized additive model, or GAM for short (Hastie and Tibshirani 1990). There are, for example, formulations for binary response variables and count response variables leading to generalizations of binomial regression and Poisson regression, respectively.

Estimation for Eq. (19.9) at first seems daunting. The functions need to be "partialed" in the same manner that the regression coefficients are in conventional linear regression. But the requisite residualizing process cannot be undertaken with unknown functions. The backfitting algorithm provides a solution (Hastie and Tibshirani 1990: 91) by cycling back and forth between smoothing for each predictor in turn and "partialing" for the dependence between predictors.[20]

Each nonparametric term in Eq. (19.9) requires a value for its λ or some other penalty parameter. Within the backfitting algorithm, therefore, the function of each nonparametric term is estimated largely as described for the single predictor case.[21] Clearly, there is a lot of heavy computing required. Somewhat surprisingly, current implementations of GAM (e.g., in R) usually run quickly except when the predictors are highly correlated. Then, convergence can be a problem.[22]

Equation (19.9) is considered "nonparametric." Although an additive form is required, no particular function for each predictor is imposed. Within this nonparametric approach, one can also include functions of predictor pairs so that one fits a surface rather than a line. In other words, functions of individual predictors and predictor pairs can be specified in a single regression equation.

The backfitting algorithm works in the same manner if particular functions are assumed for some of the predictors. For example, one predictor may be assumed to have a logarithmic relationship with the response. One can combine a weighted sum of smoothers and a weighted sum of conventional linear functions of predictors. For the former, the weights are assumed to be 1.0. For the latter, the weights are the usual regression coefficients. The result is a "semiparametric" regression. Finally, if explicit functions are imposed a priori on all predictors, one has returned to the generalized linear model, a form of parametric regression. One can still employ the backfitting algorithm or return to the usual GLM estimation procedures.

For each of the three GAM variants, categorical predictors are permitted. However, smoothing categorical predictors makes little sense. They perform, therefore, just as they do in conventional linear regression.[23] Interactions can be addressed by including products of the relevant predictors. Measures of fit (e.g., the AIC) can be computed, and just as in the single predictor case, standard confidence intervals and statistical tests are usually offered. Just as in the bivariate case, however, conventional statistical inference is problematic. All of the earlier issues reappear. Statistical tests and confidence intervals no longer have their stated properties. At this point, the best one can do is compute stability intervals.

20

1. Initialize: $\alpha = \text{ave}(y_i)$, and $f_j = f^0$, $j = 1, \ldots, p$ with linear functions.
2. Cycle: $j = 1, \ldots, p, 1, \ldots, p, \ldots$

$$f_j = \mathbf{S}_j \left(y = \alpha - \sum\nolimits_{k \neq j} f_k | x_j \right),$$

   where $\mathbf{S}_j$ is a smoother matrix.
3. Continue #2 until the individual functions don't change.

[21] For the backfitting binomial and Poisson variants, penalized maximum likelihood estimation is applied to each nonparametric regression term. In practice, this leads to the usual iteratively reweighted least-squares algorithm but with the penalty term included.

[22] There are two versions of GAM in R, one contained with the library *gam* and one contained within the library *mgcv*.

[23] Binary response variables are not a problem because the associated probabilities can be transformed into logits, which are quantitative.

A GAM Illustration

As before, the data come from Los Angeles county which arguably has the largest homeless population of any county in the country. The unit of analysis is the census tract, and there are 509 of them in the data set. Census tracts were selected by stratified random sampling from a population of 2,054 census tracts (Berk et al. 2008). The sampling was motivated substantially by the need to reduce data collection costs.

For the nonlinear approximation, the response variable is again the log of the number of homeless individuals in a census tract, obtained through a street count. The details need not trouble us here (see Berk et al. 2008). For this analysis, the predictors are (1) median household income, (2) the proportion of land used for residential purposes, (3) the log of the proportion of dwellings that are vacant, (4) the proportion of land used for commercial purposes, and (5) the proportion of residents that self-identify as a racial/ethnic minority. Past research was used to select these predictors, but there are no doubt important predictors being overlooked. For example, there are no measures of services and shelters available to the homeless that no doubt attract homeless individuals and families to certain census tracts. There are also no measures of police practices that can make some census tracts less attractive.

To illustrate the flexibility of the approach, the predictors are handled in three ways. There is a two-predictor smooth, a pair of one-predictor smooths, and a single predictor with the usual linear form imposed. The model is semiparametric.

For the two single predictors entered in a nonparametric fashion, smooths were estimated by smoothing splines. For the pair of predictors entered in a nonparametric fashion, thin plate splines was used.[24] Penalty parameters for each term on the right side were determined empirically using the generalized cross-validation statistic, but they were then evaluated for substantive credibility as well. About 30 % of the variance is accounted for.[25]

The key output can be seen in Fig. 19.4. The graph in the upper left is a perspective plot of the results for the two-predictor smooth. The two predictors are median income and proportion residential. There are some holes in the surface where there are no data.

Median income and the proportion residential should be negatively related to homelessness. In addition, median household income should matter less when the proportion of land that is residential land is lower, because there is a lower density of households to begin with. We used a two-predictor smooth to capture this interaction effect as well as any main effects. When there are no interaction effects, the predictors properly can be entered separately.

The vertical dimension represents the response variable. The label indicates that it is the smoothed fitted values that are plotted and that the smooth uses up 16.96 degrees of freedom. Smoothers can have fractional degrees of freedom but otherwise convey much the same information as degrees of freedom in linear regression models. In this example, a relatively large number of degrees of freedom is used up, indicating that the surface is very different from a plane. This is also apparent from the plot.

In Fig. 19.4, if there is an interaction between median income and proportion residential, it is not readily apparent to the eye. Moreover, when the two predictors were entered separately and the approximation reestimated, the quality of the fit did not degrade.[26]

---

[24] Thin plate splines fit a two-dimensional surface to the data (Hastie et al. 2009: Sect. 5.7).

[25] The software provided a joint test for the null hypothesis that none of the predictors was related to the log of the number of homeless. The null hypothesis was rejected at well below conventional p-values. As already discussed, however, the meaning of such tests is obscure in this context.

[26] When the relationships with a response are linear in both dimensions, and when there are no interactions, the fitted values form a plane. Along either dimension, the slope does not change with the values of the other dimension. Interactions cause the plane to be torqued. The same reasoning applies when either or both of the relationships with

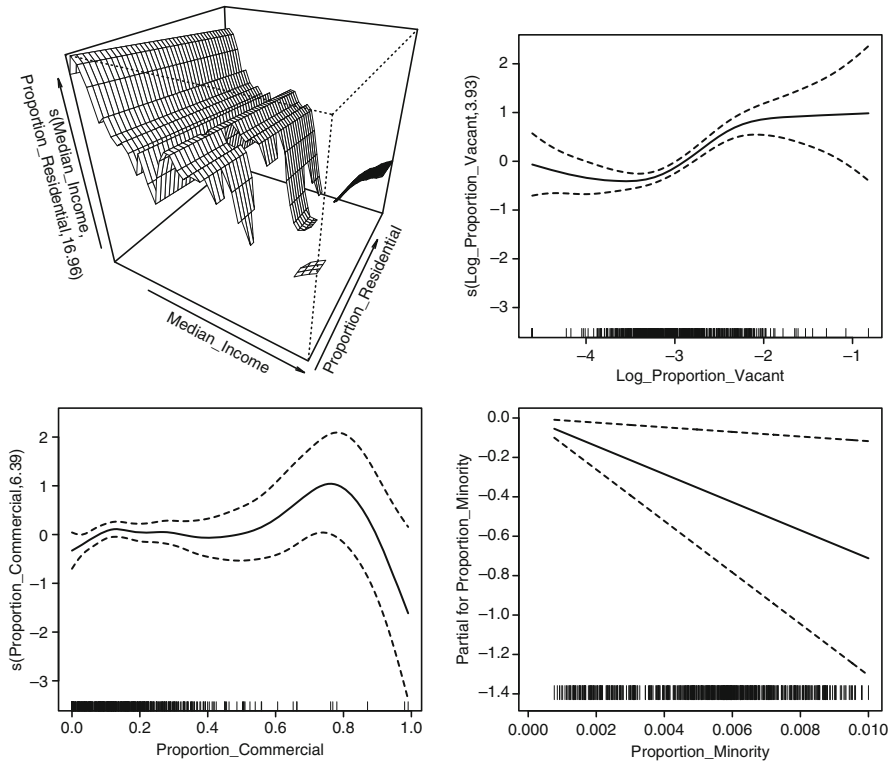

Fig. 19.4 GAM output

Median income has a negative relationship with the response that is stronger when median income is above about $50,000. There is no substantive reason to take the smaller ripples seriously, and they are almost certainly well within any sensible uncertainty bands.[27] Likewise, the upturn in the surface at very high incomes would be difficult to distinguish from noise. The proportion residential also has a negative relationship with the log of the number of homeless, but the relationship is weak.

As a descriptive matter, here is what is going on. For census tracts that are alike with respect to the proportion of dwelling units that are vacant, the proportion of land used for commercial purposes, and the proportion of residents who are minorities, tracts with higher median income have fewer homeless. This relationship is especially strong when median income is more than about $50,000. At the extreme, the difference between a very poor tract and a very rich tract is on the average about 30 homeless individuals when the tracts are otherwise alike. In short, local affluence is inversely related to homelessness, especially for wealthier communities, even if you take into account measures of a tract's racial composition, the condition of its residential housing, and the land use.[28]

The upper right graph is a one-predictor smooth for the logged proportion of dwellings that are vacant. Nearly 4 degrees of freedom are used up indicating that we are again some distance from a

the response are nonlinear (here, especially for median income). When the surface is torqued, the function for one dimension changes with values along the other dimension.

[27] It is not clear how to show uncertainty bands in three dimensions without making a plot unreadable.

[28] The adjustments for related predictors are approximations too. There is no direct correspondence to post-stratification as there is in conventional linear regression.

linear relationship. (A linear relationship would have used up 1 degree of freedom.) The relationship has much the same structure described earlier when no statistical controls were employed.

The lower left graph is a one-predictor smooth for the proportion of land that is used for commercial purposes in a census tract. A little over 6 degrees of freedom are used up indicating that the relationship is substantially nonlinear. One can see that the relationship is largely flat until the proportion tops about 80 %. At that point, the relationship turns sharply negative. There are very few observation on the far right of the graph, but taking the reported error band into account still suggests a substantial negative association after adjustments for the other predictors. Moreover, the highest values represent Los Angeles county's downtown census tracts that are dominated by large, upscale commercial buildings (e.g., for corporate headquarters). The area is very well policed, and there is a large number of private security guards. The homeless are not welcome. So, the relationship revealed makes sense even when adjusting for predictors such as median income.

The lower right graph is a one-predictor plot of the imposed linear relationship between the proportion minority in a census tract and the log of the number of homeless. Taking the log units of the response into account indicates that the slope is not very steep. Indeed, the difference between a tract that is exclusively minority and a tract that is nearly exclusively non-minority is about half a homeless person. If you know a tract's median household income and a key measure of the quality of its housing stock, race by itself is not important. Homeless individuals are not disproportionally found in minority areas, other things roughly equal.

A plot of the residuals against the fitted values showed the model to be inadequate in at least one important way. The homeless distribution is skewed to the right. Seventy-five percent of the tracts have less than about 35 homeless individuals, and 25 % have less than 6. But a few tracts have more than 300, and one tract has over 900. Even using the log of the homeless count, the model grossly underestimates the counts in these census tracts. One reason is that in the tracts with very large numbers of homeless individuals, the homeless live in homeless encampments (e.g., near downtown "skid row"). Encampments have very different dynamics from small and transient concentrations of homeless individuals. Another reason is that social services for the homeless are concentrated in areas with larger numbers of homeless people, which likely makes those areas more hospitable to the homeless. As noted earlier, predictors to capture such phenomena were not available in the data. In short, the descriptive content of the results is at least incomplete.

Broadly speaking, estimation is not a problem for these data. The data are a real random sample. *Researchers* sampled census tracts from a real population of tracts. The fitted values and plots can be taken as approximations of the population's features. They are biased approximations for the reasons discussed earlier, but for researchers interested in the distribution of homeless individuals in Los Angeles, the approximations provide rich information that can support understanding. They can also help inform causal accounts. In short, we have a finite population version of our joint probability distribution model.

More problematic are confidence intervals and statistical tests. Once again, the estimation procedure precludes conventional approaches. But stability intervals can still be helpful. We applied the nonparametric bootstrap as described earlier with the percentile method, and interpretations of the results did not change materially.

Summary and Conclusions

Researchers routinely work with causal regression models that are wrong but proceed as if the models are right. This leads to any number of conceptual confusions beginning with the parameters to be estimated. The statistical inference that follows is then incorrect. In practice, confidence intervals and statistical tests will not perform as researchers assume, and misleading statistical inferences can follow.

Causal inference is also compromised. When pushed, researchers will acknowledge that their models are not literally right but that they are close enough. However, "close enough" is usually undefined, and a factual basis for the claim is obscure. A further retreat concedes that the causal model may be substantially wrong but that it is useful nevertheless. At that point, unfortunately, most anything goes.

Is there a better way? Our approach begins with nature generating the data as a random realization from a joint probability distribution. Researchers designate a response **Y** and predictors **X**. The conditional means of **Y** given **X** are often of substantive interest. The relationship between **Y** and **X**, denoted by $f(\mathbf{X})$, is usually of substantive interest as well.

Parametric, semiparametric, or nonparametric approximations are readily available. Parametric approximations can be easy to estimate and easy to interpret. Semiparametric and nonparametric approximations are richer and can perform well, but they are more complicated. In particular, statistical inference is problematic. Currently, the best one is likely to do is to capture the sampling variance. None of the three approximation flavors provide estimates of causal effects, but the results can inform and be informed by causal reasoning.

The joint probability distribution model is less ambitious than the regression causal model. But it has several important assets. It can be far less vulnerable to untestable assumptions, and it has fewer of those assumptions to begin with. It also has broader applicability not just for parametric, semiparametric, and nonparametric regression but for machine learning and multivariate statistics in general. And perhaps most important, it provides a reasoned framework for what most social science researchers are actually doing. At the same time, a model based on nature's joint probability distribution can be wrong too.

With all of the problems that models cause, are *any* models worth the effort? If the research goal is description, models are no longer relevant. Focus in on the data itself, not how the data came to be. But if a researcher wants to make claims beyond the data on hand, a suitable object for any generalizations must be defined, and a conceptual road map to that object must be provided. And that is precisely what a model does. We suggest that nature's joint probability distribution can be an appropriate and tractable target for a wide variety of data-driven generalizations, and treating the data as a random realization from that distribution supplies the road map.

There is nothing in our formulation that precludes a consideration of causal effects. Causal thinking can help inform how a statistical approximation is specified, and causal thinking can be instrumental when results need to be interpreted. Our approach precludes using regression models with observational data to obtain estimates of how on the average a response variable will change if a given predictor is manipulated independently of all other predictors. If such estimates are desired, the best option is likely to be a randomized experiment or a strong quasi-experiment. If these are not available, there are analysis procedures, not based on conventional regression, that may have more promise (Rosenbaum 2009, 2010).

References

Angrist, J. D., & Pischke, J. (2009). *Most harmless econometrics*. Princeton: Princeton University Press.

Berk, R. A. (2003). *Regression analysis: A constructive critique*. Newberry Park: Sage Publications.

Berk, R. A., Kriegler, B., & Ylvisaker, D. (2008). Counting the homeless in Los Angeles county. In D. Nolan & S. Speed (Eds.), *Probability and statistics: Essays in honor of David A. Freedman* (Monograph series). Beachwood: Institute of Mathematical Statistics.

Berk, R. A., Brown, L., & Zhao, L. (2010). Statistical inference after model selection. *Journal of Quantitative Criminology, 26*, 217–236.

Berk, R. A., Brown, L., Buja, A., George, E., Pitkin, E., Traskin, M., Zhang, K., & Zhao, L. (2011). *Regression with a random design matrix* (Working paper). Pennsylvania: Department of Statistics, University of Pennsylvania.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics*. New York: Academic.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 78*, 829–836.
Cook, D. R., & Weisberg, S. (1999). *Applied regression including computing and graphics*. NewYork: Wiley.
Duncan, O. D. (1975). *Introduction to structural equation models*. New York: Academic.
Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics, 34*, 447–456.
Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 59–82.
Fisher, R. A. (1924). The distribution of the partial correlation coefficient. *Metron, 3,* 329–332.
Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics, 9*(6), 1218–1228.
Freedman, D. A. (2009). Diagnostics cannot have much power against general alternatives. *International Journal of Forecasting, 25*(4), 833–839.
Gelman, A., & Park, D. K. (2008). Splitting a predictor at the upper quarter third and the lower quarter or third. *The American Statistician, 62*(4), 1–8.
Goldberger, A. S., & Duncan, O. D. (1973). *Structural equation modeling in the social sciences*. New York: Seminar Press.
Greene, W. H. (2003). *Econometric analysis* (5th ed.). New York: Prentice Hall.
Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists*. New York: Academic.
Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman & Hall.
Hastie, T. J., Tibshirani, R. J., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.
Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Symposium on Mathematical Statistics and Probability, I*, 221–233.
Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Los Angeles: Sage Publications.
Leeb, H., & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics, 34*(5), 2554–2591.
Leeb, H., & Pötscher, B. M. (2008a). Model selection. In T. G. Anderson, R. A. Davis, J.-P. Kreib, & T. Mikosch (Eds.), *The handbook of financial time series* (pp. 785–821). New York: Springer.
Leeb, H., & Pötscher, B. M. (2008b). Sparse estimators and the oracle property, or the return of Hodges estimator. *Journal of Econometrics, 142*, 201–211.
Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics, 21*(1), 255–285.
Rosenbaum, P. (2009). *Design of observational studies*. New York: Springer.
Rosenbaum, P. (2010). *Observational studies* (2nd ed.). New York: Springer.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B, 58*(1), 267–288.
Thompson, S. (2002). *Sampling* (2nd ed.). New York: Wiley.
White, H. (1980). Using least squares to approximate unknown regression functions. *International Economic Review, 21*(1), 149–170.
Zellner, A. (1984). *Basic issues in econometrics*. Chicago: University of Chicago Press.