Misspecified Mean Function Regression: Making Good Use of Regression Models that are Wrong[*]

Richard Berk, Lawrence Brown, Andreas Buja,
Edward George, Emil Pitkin, Kai Zhang, Linda Zhao
Department of Statistics
Department of Criminology
University of Pennsylvania

April 10, 2013

Abstract

There are over three decades of largely unrebutted criticism of regression analysis as practiced in the social sciences. All of the criticisms apply to criminology as well. Yet, regression analysis broadly construed remains for many the method of choice for characterizing conditional relationships, sometimes interpreted in causal terms. One possible explanation is that the existing alternatives are seen by too many researchers as unsatisfying. In this paper, we provide a different approach. We proceed assuming the regression model is wrong and consider what can be learned nevertheless. The search for a "correct" model is abandoned. We offer instead a rigorous way to learn from regression approximations. These approximations, not "the truth," are the estimation targets. For these, we provide estimators that are asymptotically unbiased and standard errors that are asymptotically correct. Both can be obtained easily from popular statistical packages.

1 Introduction

There is a large literature on the many difficulties with regression modeling in the social sciences (e.g., Box, 1976; Leamer, 1983; Holland, 1986; Rubin, 1986; 2008; Freedman, 1987; Breiman, 2001; Berk, 2004; Imbens, 2009; Angrist and Pischke, 2010). By and large, this literature is unrebutted (Freedman, 2005: Sction 8.9), and regression modeling in criminology is subject to the same concerns (Blumstein et al., 1978; Welford et al., 2005; Nagin and Pepper, 2012).

Research on the deterrent effect of capital punishment is an instructive illustration. In 1978, an NRC Committee charged with reviewing the relevant

[*] Thanks go to John MacDonald and Charles Loeffler for helpful comments on an earlier draft of this paper.

research was “skeptical that the death penalty [as practiced in the United States] can ever be subjected to the kind of statistical analysis that would validly establish the presence or absence of a deterrent effect” (Blumstein, 1978: 62). But researchers proceeded with regression modeling as usual. Thirty-four years later, another NRC committee was given a similar charge. Although there were strong criticisms of theoretical foundations on which the research rested, regression modeling was again indicted.

> “The standard procedure in capital punishment research has been to impose sufficiently strong assumptions to yield definitive findings on deterrence. ... The use of strong assumptions hides the problem that the study of deterrence is plagued by model uncertainty and that many of the assumptions used in the research lack credibility” (Nagin and Pepper, 2012: 7).

Further,

> “The committee concludes that research to date on the effect of capital punishment on homicide is not informative about whether capital punishment decreases, increases, or has no effect on homicide rates. ... Consequently, claims that research demonstrates that capital punishment decreases or increases the homicide rate by a specific amount or has no effect on the homicide rate should not influence policy judgments about capital punishment” (Nagin and Pepper, 2012: 2).

Why in the face of powerful critiques and a disappointing track record do so many criminal justice researchers maintain their attachment to regression modeling? One reason may be that the existing analysis alternatives for observational data can sometimes be unattractive. For example, multiple equation and hierarchical models layer on additional complexity without really addressing the causal modeling critique (Freedman, 2005: Chapter 8). Matching methods are more robust (Rosenbaum, 2002; 2010), but borrow heavily from the experimental paradigm, which some find limiting (Heckman and Smith, 1995). Although combining the formal logic of causal inference with acyclic graphs (Morgan and Winship 2007) has considerable appeal, the empirical leverage provided by graphical models of causation can be substantially overstated (Freedman, 2004).

There is another way. Rather than trying to find acceptable alternatives to regression modeling, researchers can perhaps learn to make better use of the regression tools they already have. A key may be to make research aspirations more consistent with what can actually be accomplished with observational data. From this point of view, Manski (2003) places bounds around causal effect estimates to capture the impact of identification weaknesses in certain estimation procedures. Imbens and Angrist (1994) provide “local” estimates for subpopulations within which the modeling assumptions may be more credible.

We offer another approach that depends on reduced aspirations. In contrast to conventional regression practice, we explicitly discard the goal of getting a model "right." We consider what can be learned from empirical results that are manifestly approximations of unknown relationships in a target population. Our approach has much in common the "correlation model" proposed by Freedman (1981), and its foundation is closely related to procedures formulated by White (1980). Angrist and Pischke (2009; Section 3.1.2) provide very accessible motivation for regression as approximation.

Section 2 reviews some key properties of the linear regression mean function with fixed predictors. Section 3 considers the mean function when predictors are random. The intent is to introduce key issues at a broad conceptual level. In Section 4, we provide the technical background and justifications for working with linear approximations. Some readers may choose to skip this section if they are prepared to accept our main arguments at face value. Section 5 turns to implementation and practice. In section 6, there is a simple example using real data. Section 7 briefly broadens the discussion to include parametric nonlinear regression and smoothers. Section 8 offers some broad conclusions.

2 Conventional Linear Regression with Fixed Predictors: Once Over Lightly

In this section and the next, we raise issues that help motivate the rest of the paper. The intent is to highlight problems in a relatively nontechnical manner, which we later intend to solve. Readers seeking a more formal treatment will have to wait until Section 4.

The standard formulation for linear regression takes the following form:

$$Y = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_N, \sigma^2 \boldsymbol{I}_{N\times N}), \tag{1}$$

where Y is the response variable, N is the number of observations, and $\boldsymbol{X}$ has p predictors with an additional column of 1s for the intercept.[1]

The usual interpretation attached to Equation 1 is that one has a true account of how the N values of the response variable Y are produced by "nature." For each case i, one might say that nature first determines the values of the p predictors in $\mathbf{X}$, then combines them and the leading constant in a linear fashion using the corresponding regression coefficients, and adds a random draw from a distribution of disturbances that has a mean of 0 and a single variance applicable to each case. That distribution is taken to be normal, although in our context, normality is not an important assumption. Nature is able to repeat this process a limitless number of times for each case using the given, *fixed* values of the predictors. The disturbances are the only source of randomness in Y. Over realizations for a given case, the response values can change, but the predictor values cannot.

Equation 1 is "first order correct" if the mean function corresponds to nature's "true" conditional means: $\boldsymbol{\mu}_i | \boldsymbol{X}_i$.[2] These conditional means are found in

[1] $\boldsymbol{X}$ has N rows and $p+1$ columns.

[2] That is, $\boldsymbol{\mu}_i | \boldsymbol{X}_i = \boldsymbol{X}_i \boldsymbol{\beta}$.

the real process that nature employs to generate the response. Unbiased estimates of $\boldsymbol{\beta}$ and σ^2 require that Equation 1 is first order correct. Equation 1 is "second order correct" if the disturbances have the properties specified in Equation 1, although formally, normality is really not a second order condition, but a convenient assumption about a distributional form.[3] Second order correctness is necessary for statistical tests and confidence intervals to perform as they should.

When researchers consider whether a regression model is second order correct, they usually assume that the model is already first order correct. Otherwise it is very difficult to empirically distinguish between first order errors and second order errors. For example, if the mean function is incorrect, there will likely be the appearance of nonconstant variance even if σ_i^2 is the same for each case. Such confounding can undermine a range of diagnostic tools.

Regression models and their close cousins have been quite properly criticized because there is usually no definitive way to know if either the first order or the second order conditions are met.[4] The result too often is science by hand waving. There is a large, accessible literature on such matters that can be consulted. However, to help motivate our alternative perspective, we have to briefly consider a particular set of difficulties. We focus on the regression mean function, which is the statistical bedrock for conventional regression analysis.

2.1 Regression Mean Functions with Fixed X

For expositional purposes, suppose for the moment that the response is a linear, *deterministic* function of a single, fixed predictor — there are no disturbances. When there are no disturbances, the regression mean function is much easier to visualize.

Figure 1 shows the relationship between the response and that single predictor for the conventional linear model. The black line is nature's mean function. The blue circles are hypothetical observations for the response at some predictor values assuming no disturbances.[5] Because one has the correct linear function, one can determine the value of the response for any value of the predictor, even when the value of the predictor is not observed. In effect, one can impute the value of the response using the correct linear function. This means that the regression results can be properly generalized beyond the data's predictor values. Put another way, no matter what fixed values X one has, the conditional means of the response map out the correct linear function.

Under these circumstances, the distribution of the predictor is unrelated to $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. It follows that one can condition on the predictor, the usual practice, and obtain valid estimates of the regression parameters from the data. This

[3] With a sufficiently large sample size, the normality assumption can be safely ignored.

[4] Close cousins include the generalized linear model and extensions to models with more than one response variable. The defining feature is a focus on the conditional distribution of one or more responses that depend on one or more predictors.

[5] We will later explain why the observations can be more properly seen as nature's conditional expectations for the response variable.

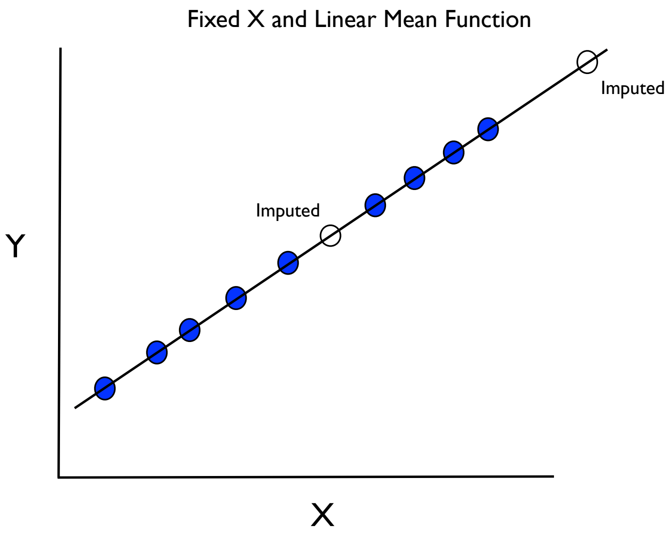

Figure 1: The Canonical Regression Formulation with Y a Deterministic Linear Function of a Fixed X

is the usual backstory in a wide variety of applications for which the mean function is really linear and the researcher knows it. One can think of this as the conventional hubris (Freedman, 2005).

3 Regression Mean Functions with Random X

Fixed regressors can be an under-appreciated constraint on the conventional regression model. Uncertainty is solely a function of the disturbances. Regression estimates are seen as varying over realizations of the data with the predictor values constant.

Three complications can follow. First, if the predictors are actually random variables, an additional source of uncertainty is neglected. For example, survey data constructed by random sampling necessarily makes all predictors random variables. Predictors generated in other ways can be random as well. With predictors as random variables, some important properties of least squares regression no longer hold (Freedman, 2005: section 4.11).[6] A popular response is to treat the predictor values as fixed once they materialize in the sample. But that leads to a second complication: generalizations beyond the data on hand can be jeopardized. Formally, the regression results apply only to the particular predictor values appearing in the sample.[7] Still, as long as mean function is linear and the researcher knows it, generalization beyond the predictor values in the data can be justified.

Figure 2 illustrates why. The predictor values in the data are random realizations from the predictor's underlying distribution. Any time that Y and X are observed, the values of both random variables could have been different, not just for Y. In Figure 2, the black line is nature's mean function. The blue circles are observations for one random data realization. The red circles are observations from another random data realization. The implications for a correct mean function are much the same as before with one important addition: we are able to map the correct, linear form for the conditional means of the response no matter which predictor values *happen* to appear in the data. Conditioning on the predictor values once again permits valid estimates. This is another backstory, perhaps less common, but like the first requires that the mean function is linear and the researcher knows it. Generalizations to nonlinear relationships can sometimes be justified by a similar account, but there can also be complications we will address later.

Figure 3 tells a much darker and complicated tale. Just as in Figure 2, both the response and the predictor are random variables. Nature's mean function, shown by the broken line in black, is now *nonlinear*. Clearly, any linear function will fail to reproduce nature's true mean function. But there is much more to the story.

[6] With fixed **X**, when the regression model is first order and second order correct, regression coefficient estimates from least squares regression are the "best *linear* unbiased estimates" (BLUE) available. But with random **X**, least squares estimates are *nonlinear* in **X**.

[7] This is necessarily true when the predictors are fixed.

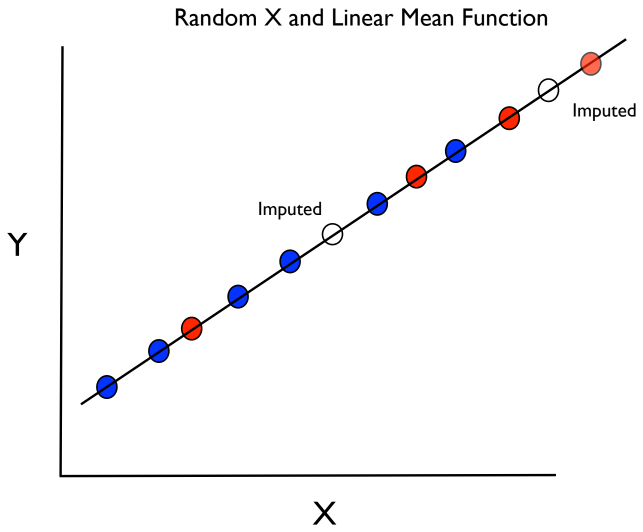

Figure 2: The Canonical Regression Formulation with Y a Deterministic Linear Function of a Random X

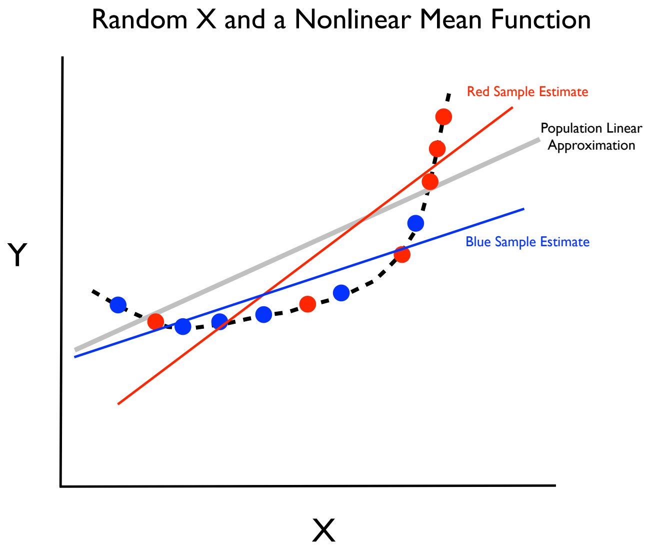

Figure 3: Nonlinear Mean Function and Random X

Because of the nonlinear mean function, the predictor values in one's sample matter in new and important ways. If a researcher happens to get the data shown with the red circles, the conditional means from a linear least squares regression result in a substantially steeper slope than if the researcher happens to get the data shown with the blue circles. The population predictor distribution, therefore, is related to $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. For example, if the predictor distribution is concentrated at low values, blue circles rather than red circles are more likely to be realized. If the predictor distribution is concentrated as high values, the reverse is true. Regression estimates no longer have their desirable properties.

But it's worse. For the moment, imagine a population that could be generated by nature. Imagine being able to compute a bivariate least squares regression in that population. One can quite properly interpret that regression as a feature of the population. In the population, the nonlinear mean function is still the correct functional form. The population regression is a linear approximation of the true mean function. The approximation is shown as the straight gray line in Figure 3.

Suppose the data on hand can be seen as random sample from such a population. How well do linear regressions computed from different samples estimate the population linear approximation? Figure 3 shows that the least squares regressions computed from the red data or the blue data fail to get the population linear approximation right. In a least squares regression, researchers condition on the predictor values realized in the sample. In actual samples, the least squares regression will almost surely differ systematically from a population linear approximation. From any given sample, a researcher only gets to see a random piece of that nonlinear truth. As shown in Figure 2, this is not a problem when the true relationship is actually linear. From Figure 3, one learns that with a nonlinear mean function and random variable predictor, a linear mean function computed from real data will misrepresent the population nonlinear mean function and likely misrepresent a linear, population least squares mean function as well. Any given sample will not have all of the information required.[8]

Building on random $\mathbf{X}$, there is nevertheless a defensible way to proceed. The true mean function is taken to be unknown and with no requirement of linearity. One accepts that the true mean function cannot be properly estimated from the data. But perhaps a population linear approximation of the truth has useful substantive information even through it is almost surely incorrect. *It is this population linear approximation that one seeks to estimate.* The same approach can apply when there are many predictors. The estimation target is then a population hyperplane.

Despite the disappointing conclusions from Figure 3, it is possible to obtain suitable estimates of the linear approximation and appropriate standard errors, at least in large samples. We turn to a more technical discussion to make

[8] We have been assuming that the regression predictor is a random variable because we will make good use of random predictors shortly. But the problems just described have some similar counterparts with fixed predictors, although a somewhat different formation from the ground up is required to characterize them.

the rationale and results much more precise. Readers interested primarily in practical implications may wish to skip to Section 5.

4 Conceptual Formalities

We begin with a set of q *random* variables $Z_1, Z_2, \ldots, Z_q$ characterized by a joint probability distribution. Because $Z_1, Z_2, \ldots, Z_q$ are random variables, they have mathematically defined properties such as means (usually called expectations), variances, and covariances, much like a real data set. It can be instructive, therefore, to refer to the joint probability distribution as a "population." We will on occasion refer to this population as a feature of "nature."

It is important to stress that in contrast to populations associated with conventional regression, in this population all variables are random variables. We assume each random variable has second moments that exist and that the covariance matrix of the random variables is full rank (i.e., no subset of variables is an exact linear function of another subset of variables). These requirements for the random variables are not important constraints in practice. No particular distributional form is imposed (e.g., multivariate normality).

A researcher designates one of the random variables as a response variable, denoted by Y. The researcher also designates p other random variables as predictors, denoted by $X_1, X_2, \ldots, X_p$. All predictors are collected in a matrix $\boldsymbol{X}$ with $p + 1$ columns that includes a leading column of 1s. The distinction between a response and its predictors is *not* inherent in nature's population. It derives from subject-matter knowledge and interests that a researcher imposes on the random variables.

Data on hand are treated as random realizations from nature's joint probability distribution. Each observation i is one such realization, and all of the observations are realized independently. One usefully can think of each observation as a random, independent draw from nature's population. Even though a researcher has made a distinction between Y and $\boldsymbol{X}$, the data are *not* in general a realization from the regression formulation shown in Equation 1. This is a fundamental difference between conventional regression modeling and the formulation to follow.

4.1 Some Features of the Population

For this formulation to play through, it is essential to be far more precise about the population and its properties. For a more detailed discussion see Buja et al. (2013).

1. As a notational convenience, we write the set of *random* variables $\vec{\boldsymbol{X}} = (1, X_1, \ldots, X_p)^T$ as a column vector that includes a 1 for the intercept. The variables may be quantitative or categorical. This notation will make some of the expressions to follow seem unfamiliar.

2. There is a “true response surface” $\mu(\overrightarrow{\boldsymbol{X}})$ in nature’s population, that is the *expectation* of the response conditional on given values for the predictors $\overrightarrow{\boldsymbol{X}}$. More formally,
$$\mu(\overrightarrow{\boldsymbol{X}}) = E[Y|\overrightarrow{\boldsymbol{X}}]. \tag{2}$$
   A possible population target for estimation is a set of expected values one might conventionally denote by $\boldsymbol{\mu}_i|\boldsymbol{X}_i$. But the conventional notation does not indicate that the population “means” are actually expected values, so the notation in Equation 2 is preferred. This is a key difference from the usual inferential approach and is a game-changer. There is randomness in the population that cascades though any regression analysis.

3. There is no assumption of linearity for the true response surface. Indeed, the working assumption is that it is nonlinear. It might be truly nonlinear or be nonlinear because of omitted variables or other factors. At this point, no distinctions are made between different reasons for the nonlinearity.

4. There is a *population* linear least squares approximation of the response variable’s conditional expectations.
$$\boldsymbol{\beta}^T\overrightarrow{\boldsymbol{X}} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, \tag{3}$$
   where
$$\boldsymbol{\beta} = \text{argmin}_{\tilde{\beta}} E[(Y - \tilde{\boldsymbol{\beta}}^T\overrightarrow{\boldsymbol{X}})^2] = E[\overrightarrow{\boldsymbol{X}}\overrightarrow{\boldsymbol{X}}^T]^{-1}E[\mu(\overrightarrow{\boldsymbol{X}})\overrightarrow{\boldsymbol{X}}]. \tag{4}$$
   The regression coefficients are a function of expectations that depend on the predictors. In effect, one is working with expectations of cross-product matrices rather than realized cross-product matrices.

5. The usual covariance adjustments are in play, but are now a function of the random predictors. Consider the p-vector of regression coefficients denoted by $\boldsymbol{\beta}_{j\bullet}$:
$$\boldsymbol{\beta}_{j\bullet} = \text{argmin}_{\tilde{\beta}} E[(X_j - \tilde{\boldsymbol{\beta}}^T\overrightarrow{\boldsymbol{X}}_{notj})^2] = E[\overrightarrow{\boldsymbol{X}}_{notj}\overrightarrow{\boldsymbol{X}}^T_{notj}]^{-1}E[\overrightarrow{\boldsymbol{X}}_{notj}X_j]. \tag{5}$$
   Then the adjusted jth predictor is[9]
$$X_{j\bullet} = X_j - \boldsymbol{\beta}^T_{j\bullet}\overrightarrow{\boldsymbol{X}}_{notj} \tag{6}$$
   Finally, the population regression coefficient for that predictor is
$$\beta_j = \frac{E[YX_{j\bullet]}}{E[X^2_{j\bullet}]}, \tag{7}$$
   which is just the jth component of Equation 4.

[9] The subscript *notj* is all predictors but the jth predictor. The intercept β_0 is also subject to adjustment, but it is still interpreted as a constant.

6. Because the linear approximation from Equation 4 is just an approximation of the true response surface from Equation 2, there must be explicit allowance for mean function error $\eta(\overrightarrow{\boldsymbol{X}})$ that is responsible for disparities between the two:

   $$\eta(\overrightarrow{\boldsymbol{X}}) = \boldsymbol{\beta}_T\overrightarrow{\boldsymbol{X}} - \mu(\overrightarrow{\boldsymbol{X}}). \tag{8}$$

   Both terms to the right of the equal sign are random variables because $\overrightarrow{\boldsymbol{X}}$ is random. Hence, the difference between the two terms is a random variable as well. In effect, there is a new kind of disturbance term. This will have important implications for how uncertainty in statistics from samples is addressed.

7. There is, in addition, "pure noise" (also called "irreducible error") ϵ defined as

   $$\epsilon = Y - \mu(\overrightarrow{\boldsymbol{X}}). \tag{9}$$

   Even if one knows the true response surface, the population fit of Y will not likely be exact. The best one can usually do is the true conditional means, and there will be a distribution of response values around each. This is a consequence of the joint probability distribution formulation. The variance of ϵ can vary over predictor values. There is no requirement of homoscedasticity. Even more, the conditional distribution itself of ϵ can differ over different locations in the predictor space.[10]

8. It follows that in the population the total disparity between any hypothetical value of the response and the population linear approximation can be written as,

   $$\xi = Y - \boldsymbol{\beta}^T\overrightarrow{\boldsymbol{X}} = \eta(\overrightarrow{\boldsymbol{X}}) + \epsilon. \tag{10}$$

Figure 4 is a visual aide. The difference between a hypothetical value of the response and the population mean function $\eta(\overrightarrow{\boldsymbol{X}})$ we call "total error." It can be decomposed into mean function error and irreducible error, both random quantities. The decomposition will figure significantly in later material.[11]

4.2 Sample Properties

Suppose least squares regression computations are applied to the realized data in the usual way. In conventional matrix notation (because we are working with a sample),

$$\begin{aligned} \hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T &= \qquad (11) \\ \text{argmin}_{\tilde{\beta}}(Y - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^2 &= \qquad (12) \\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^TY. & \qquad (13) \end{aligned}$$

[10] In general, the "pure noise" ϵ is not stochastically independent of the predictors, though it is uncorrelated with them.

[11] In equation form, $Y = \boldsymbol{\beta}_T\overrightarrow{\boldsymbol{X}} + \eta(\overrightarrow{\boldsymbol{X}}) + \epsilon$.

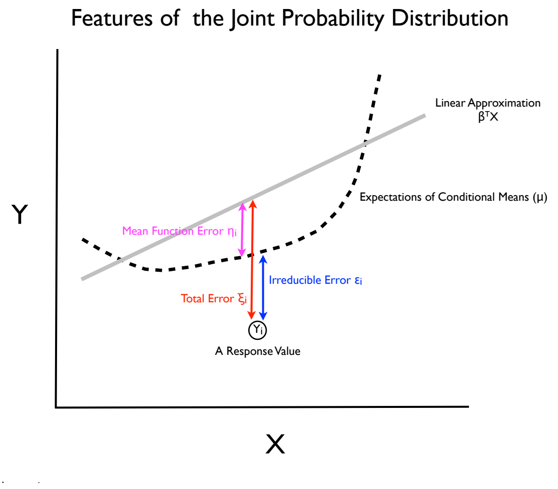

Figure 4: A Decomposition of "Total Error" in the Population

The intent is *not* to estimate nature's true mean function. The intent is to estimate the population linear *approximation* of $\boldsymbol{\mu}|\boldsymbol{X}$. We have given up on trying to estimate the "truth," and the model we are applying is explicitly permitted to be wrong. Nevertheless, the usual expressions follow. For the hat or projection matrix:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T. \tag{14}$$

For the fitted values:

$$\hat{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}Y. \tag{15}$$

For sample residuals, which are not the population "residuals" ξ,

$$\boldsymbol{r} = Y - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{H})Y. \tag{16}$$

In summary, the estimation target is not nature's response surface. The estimation target is nature's linear approximation of that surface. How closely the two correspond is unknown. Researchers are to make the best they can of the linear approximation that with respect to the "truth" is explicitly allowed to be wrong. We have even more thoroughly parted company with conventional least squares regression.

5 Working with the Linear Approximation

5.1 Interpreting the Regression Coefficients

What is the nature of the slopes for the population linear approximation? The regression coefficients are just the usual slopes of a linear least squares fit. Each slope represents the difference in the expectation of the response for a unit difference in the predictor, after adjusting for the predictor's linear association with all other predictors. What the population slopes convey with respect to the true conditional means is rather different.

For ease of exposition, consider the slope when there is a single predictor.

$$\beta = E\left[\frac{\frac{Y-E(Y)}{X-E(X)}(X-E(X))^2}{E[(X-E(X))^2]}\right], \tag{17}$$

From the fraction in the numerator, the slope for any hypothetical case is based on the slope of a line segment from the center point $(E(X), E(Y))$ to a hypothetical data point (X, Y).[12] Each such slope is weighted by the ratio of the squared deviation score $(X - E(X))^2$ and the expected value of such squared deviation score in the population. Slopes of line segments farther from the $E(X)$ are given more weight because they have greater influence on the population slope.

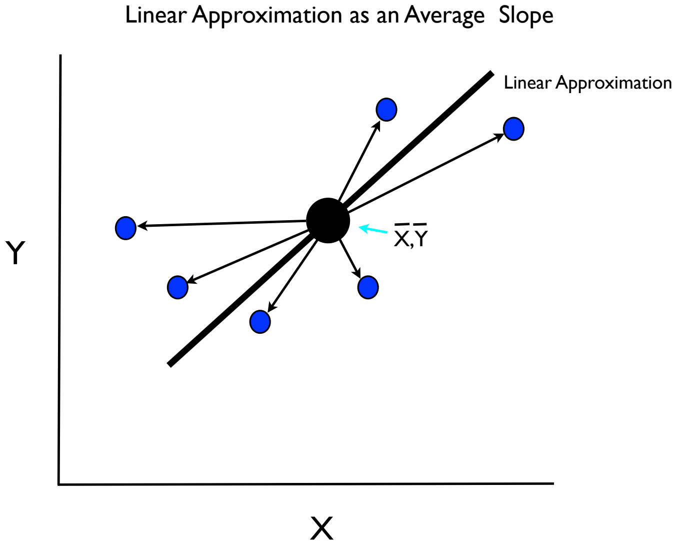
Linear Approximation as an Average Slope
Linear Approximation
$\overline{X},\overline{Y}$
Y
X

Figure 5: Linear Approximation Slope as an Average of Slopes for a Single Predictor

One can interpret β as an average slope. Figure 5 provides a visual rendering of what is being estimated. (As before, one allows the population true mean

[12] The term "hypothetical" is used because the population is a joint probability distribution. There are no realized observations.

function to be nonlinear.) In Figure 5, there are six line segments, one for each observation shown. These are slopes. Each slope goes through the expectations of the response and the predictor, shown by the large black circle, as it should. The population least squares linear approximation, shown with a thick black line, has a slope that is the weighted average of the six. It gets the slope wrong for each observation. Nevertheless, it may be an instructive summary of how X and Y are related within the population. The population linear approximation is properly estimated by the sample least squares line.

When there is more than one predictor, the same interpretation applies with the qualification that each slope is a "partial" slope subject to the usual covariance adjustments. In effect, each predictor is residualized by removing any linear associations it has with all other predictors. Each slope represents the average difference in the expectation of the response for a one unit difference in the residualized predictor.

One must be careful not to ask more of the population linear approximation than it can deliver. The impact of any omitted variables that are true confounders[13] are absorbed in the regression coefficients used in the linear approximation. The residualization process cannot address that confounding. Moreover, the slope is an average. Consequently, it will likely overstate or understate the slope at any particular observation. For example, the slope for one more year of education beyond 9th grade could be very different from the slope for one more year of education beyond 11th grade. Yet, the linear approximation imposes the same slope. One risks both misleading description and misleading causal inferences. If it makes sense to talk about causal effects, they may actually differ by predictor values, but under the linear approximation all causal effects are of the same direction and size.[14]

Still, working with an average slope may not be as limiting as it first appears. In randomized experiments, for example, it is common to seek estimates of the average treatment effect (ATE). This is an average over study subjects for which heterogeneity in potential responses under both the experimental or control condition is assumed (Holland, 1986). The slope of our linear approximation is in the same spirit, even when given no causal interpretation.

5.2 Estimation

We have abandoned trying to estimate nature's true conditional expectations $\boldsymbol{\mu}|\overrightarrow{\boldsymbol{X}}$, and are prepared to settle for a linear approximation $\boldsymbol{\beta}^T\overrightarrow{\boldsymbol{X}}$. We seek to estimate the linear conditional mean function within nature's joint probability distribution. We do this with the available data.

But, as already noted, the estimated least squares regression coefficients depend on which predictor values happen to appear in the data. Moreover, the

[13] They are correlated with the response variable and one or more predictors.

[14] The issues can be subtle. In the presence of confounding, one is at best getting an estimate of the average causal effect when both the designated predictor and the omitted predictors with which it is confounded are manipulated. What are those omitted predictors and are they actually manipulable?

estimates from any given sample will be derived from incomplete information because response values can only be observed for a random subset of predictor values. Empirical realizations for the full response surface not available. The result is that any given sample will provide incorrect estimates of the population linear approximation, and over realizations of the data, those mistakes do not cancel out. There is bias.

With larger samples, however, the regions of the true response surface that are not observed will be fewer. One can imagine that as the sample size grows without limit, the entire response surface will be observed. There is, then, no bias in the linear approximation. Stated more conventionally, the population linear approximation can be estimated with conventional least squares so that the bias disappears asymptotically. This means that in practice, the bias can be small in large samples.

The joint distribution of the regression coefficients is asymptotically normal. The marginal distributions are as well. This means that the stage is nearly set for conventional statistical inference, at least in large samples.

It may be important to underscore that the regression coefficients can be estimated in an asymptotically unbiased manner even in the presence of confounding. Omitted variables can raise interpretative problems to be sure, but in contrast to conventional regression, do not preclude valid statistical inference.

5.3 Standard Errors

If uncertainty in estimates of the linear approximation is to be properly addressed, appropriate standard errors can be essential. One might think that because the linear approximation is just least squares regression, the usual regression standard errors would suffice. They don't.

One problem is that by working with random rather than fixed predictors, there is an additional source of uncertainty. Estimates are not limited to the predictor values in the data. Another problem is that because the disparities between the expectations of the fitted values from the linear approximation and the expectations of nature's conditional mean are not constant, neither is the variance around the linear approximation. There can be nonconstant variance in the overall error ξ even if the variance in the irreducible error ϵ is constant (i.e., homoscedastisic).To emphasize this point, the expression for the misspecification disparities is reproduced subscripted for each possible observation.

$$\eta(\overrightarrow{\boldsymbol{X}_i}) = \boldsymbol{\beta}^T \overrightarrow{\boldsymbol{X}_i} - \mu(\overrightarrow{\boldsymbol{X}_i}). \tag{18}$$

Because $\eta(\overrightarrow{\boldsymbol{X}_i})$ is a function of the random predictors, it is also a random quantity. And as such, it contributes to the random variation around the expectations of the linear approximation's fitted values and causes the variances to differ.

Figure 6 illustrates how. Observations Y_1 and Y_2 happen to have the same sized irreducible errors shown in blue. Yet, the total error, shown in red, is larger for Y_2. The reason is that the mean function error, shown in magenta, is larger

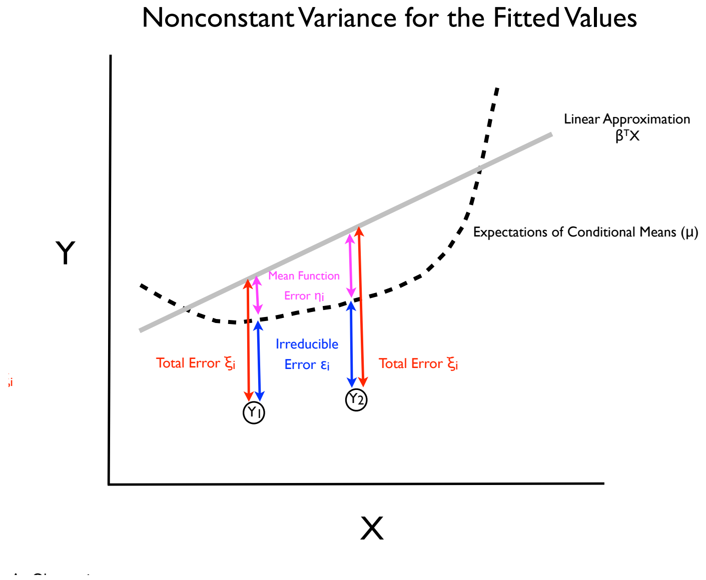

Figure 6: Source of Nonconstant Variance in Linear Approximation Estimates

for Y_2. Because the true mean function is nonlinear, its distance from the linear approximation will vary, and that varying distance is built into the variance of disparities between observations and the linear approximation. The result is nonconstant variance around the fitted values from the linear approximation. Conventional least squares regression standard errors estimated from the data are incorrect and potentially misleading. In general, they will be too small. There is false power.

One might think that the problems with the conventional standard errors are less serious in larger samples. Actually, the problems remain. As the sample size increases, the mean squared error of the estimated approximation decreases. However, it still will have two components. The first is the irreducible error resulting from variation around nature's true conditional means. The second results from the misspecification inherent in the linear approximation coupled with random predictors. Both decline in larger samples, but at the same rate; the consequences of model misspecification do not decline relative to the irreducible error. No matter what the sample size, therefore, one is still faced with nonconstant variability in the estimated linear approximation.

5.3.1 Huber-White Standard Errors

There are two good ways to obtain *asymptotically* valid standard errors. The first uses Huber-White robust standard errors, sometime called the "sandwich estimator." It's trick is to allow the square of each case's residual, which will vary in response to nonconstant variance, to directly affect the calculations one by one.[15]

Within our joint probability distribution framework, the Huber-White variance-covariance matrix for the linear approximation's regression coefficients can be written as,

$$\mathrm{VC}_{\hat{\beta}} = E[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1} E[\sigma^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] E[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}. \tag{19}$$

The square root of the main diagonal elements are the standard errors. In practice, estimates will depend on the usual predictor matrix $\boldsymbol{X}$.[16] $E[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]$ is estimated $(\boldsymbol{X}^T\boldsymbol{X})/N$, and the estimated mean squared error $\hat{\sigma}^2$ is obtained from the standard regression output.

5.3.2 Bootstrap Standard Errors

The bootstrap is essentially a simulation of the frequentist thought experiment. There are two approaches. The "parametric" method takes the regression model as at least first order correct. It follows that the simulation addresses uncertainty in Y caused by the disturbances only. Because the predictors are taken to be fixed, they cannot be a source of uncertainty in Y. The "nonparametric" method, consistent with the perspective taken here, treats uncertainty in Y as a result of the disturbances and the predictors, which are random variables. The nonparametric approach proceeds in the following manner.

1. There is a joint probability distribution F, as described earlier, that is the source of the data.

2. The data are a random realization of size n.

3. There are in the data observed units $u_1, u_2, \ldots, u_n$ that are of the same type as found in the population. Here, the units might be individual offenders.

4. For each realized unit, there are observed measurements for the response variable and the predictor variables that are the same types as in the joint probability distribution. For example, there might be measures of criminal activity and the usual background variables. These measures have an *empirical* joint distribution $\hat{F}$. $\hat{F}$ is an estimate of F.

[15] There are several proposals that appear to improve the perfomance of Huber-White standard errors in remarkably small samples (Long and Ervin, 2000). The formal rationale, however, is incomplete and may not be appropriate in our setting.

[16] There is more going on than might initially be apparent. The notational use of σ^2 does not convey that the variance of the disturbances depends on the values of the predictors. Other notation, such as $\sigma^2(\vec{\boldsymbol{X}})$, could have been used, but would probably be unfamiliar. In addition, each squared residual is an inconsistent estimate of the corresponding disturbance variance. The bias does not go to 0 as the sample size increases without limit. It is the expectation that is a consistent estimate. Nevertheless, Equation 19 "works" asymptotically.

5. There are “plug-in” estimates for the parameter(s) $\theta = t(F)$ that take the form $\hat{\theta} = t(\hat{F})$. One can use the same function $t(.)$ with the data that one would use in the population if one could get to it. For example, θ can be the set of regression coefficients from the population linear approximation and $\hat{\theta}$ could be the regression coefficient estimates from the data.

6. There are B samples drawn from the dataset sometimes called bootstrap samples, $\mathbf{s}^{*1}, \mathbf{s}^{*2}, \ldots, \mathbf{s}^{*B}$. The samples at generated by random sampling with replacement, although more complicated sampling designs are used if they were used to generate the actual data originally. In practice, B can be as small as 30 or larger than 1,000, depending on the data and the purpose of the bootstrap.

7. There are plug-in estimates one computes for each of the B bootstrap samples is $t(Y^{*1}, \mathbf{X}^{*1}), t(Y^{*2}, \mathbf{X}^{*2}), \ldots, t(Y^{*B}, \mathbf{X}^{*B})$. For example, from each bootstrap sample one might compute regression coefficients.

8. The set of plug-in estimates can be used to construct an *empirical* sampling distribution $\hat{\theta}$. The standard deviation of the empirical sampling distribution for each plug-in estimate is an estimate of the standard error. For example, the standard deviation for each regression coefficient over bootstrap samples is an estimate of each regression coefficient’s standard error. It is also possible to undertake statistical tests and/or confidence intervals directly from the empirical sampling distribution of the plug-in estimates.

Both the Huber-White standard errors and the bootstrap standard errors are only justified asymptotically and are asymptotically comparable. Which one uses seems at this point to be a matter of convenience. We are exploring whether the two approaches have different performance characteristics in samples of the size one often sees in the social sciences.

6 A Simple Example

Consider a joint probability distribution of random variables for individuals on probation in a large city. There is a dataset that can be sensibly seen as a random realization from that joint distribution. For example, the joint distribution characterizes all individuals on probation in that city for a five year period, whereas the data are for all individuals from an arbitrary four-month interval.[17]

Although the joint probability distribution is composed of many random variables, only two are available:

[17] In practice, one would have to establish that the composition of the probationer population and the process by which individuals were sentenced to parole did not change in important ways over that five-year period.

1. The number of *prior* charges for a serious crime at the time the individual was sentenced to probation; and
2. The age at which an offender had his/her first arrest leading to a court appearance charged an adult.

The researcher treats the first as the response and treats the second as a predictor. A “serious” prior charge includes murder, attempted murder, robbery, aggravated assault, and rape. At the time when an individual begins probation supervision, what is the relationship between the age at which a first arrest occurs and the number of prior charges for serious crimes?

The blue dots in Figure 7 represent the conditional expectations in the joint probability distribution, which can be seen as the population. These conditional expectations constitute the unknowable true response surface. Although for visualization purposes they are plotted against the single predictor, they would in practice be related to other predictors not included on the available data. Those missing predictors might help explain why the conditional expectations for the number of serious priors does not decline for the two youngest age groups in contrast to the smooth, nearly monotonic decline thereafter. For example, some of the charges for those under 18 may be treated as juvenile offenses and not become part of the adult record. The number of prior charges is too small. The conditional expectation for the youngest age group could be 13, not 10.2.[18]

The smaller red dots are a random realizations from the joint probability distribution. The red dots are what the researcher gets to see. In this example, the sample size is small to make the plot more visually accessible.[19] The solid black line is the *estimated* linear approximation. Its intercept is 6.2, and its slope is -.13. As usual, the intercept is required to vertically locate the estimated linear approximation, but in this instance has no substantive interpretation.[20] The estimated slope indicates that on the average, the estimated mean number of serious priors declines by .13 for every additional year of age at first arrest. For some age intervals, however, the true slope is more steep. For other age intervals, the true slope is less steep. For the two youngest ages, the true slope is actually positive. Clearly, the linear approximation is missing important features of the true relationship between conditional expectations of the response and the predictor.

At the same time, both the estimated intercept and slope are asymptotically unbiased estimates of the intercept and slope of the linear approximation within nature’s joint probability distribution. In practice, a researcher would need to decide whether the estimated linear approximation is substantively instructive. Is it instructive to know the average slope between the age at first arrest and the number of priors for serious priors?

[18] For this illustration, the blue dots are the means of the response for each arrest age from a real dataset with nearly 200,000 observations. With so large a dataset, conditional means can be treated for this illustration as if they were the population conditional expectations. Our theoretical work, however, is based on a joint probability distribution, not a finite empirical population.

[19] The realizations were randomly drawn for the empirical, finite population.

[20] It is the estimated mean number of serious priors at birth.

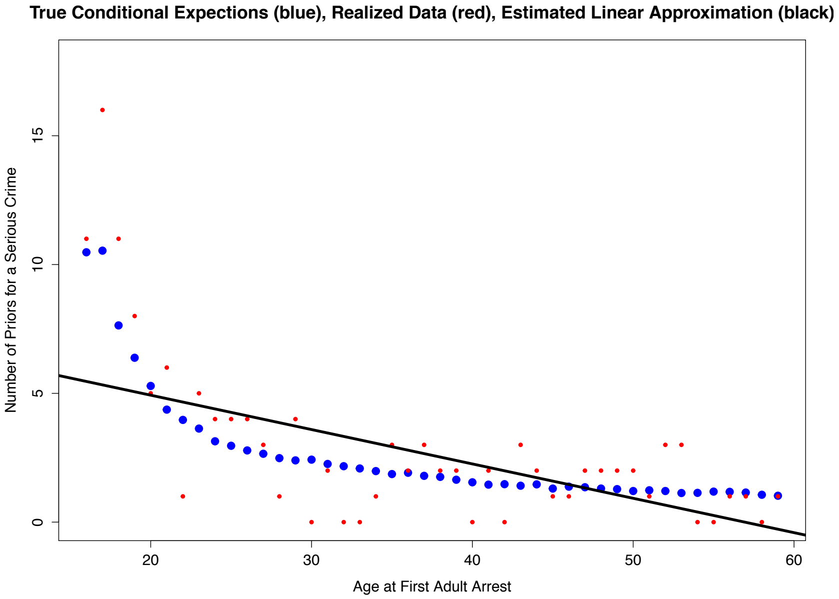

Figure 7: A Linear Approximation in Practice

Some might argue that the relationship is uninteresting because criminal activity that starts at an earlier age simply provides more time to acquire priors. However, individuals who start committing serious crimes at an early age probably spend more time incarcerated. Despite being incapacitated for significant intervals, criminals who start early still manage to accumulate a greater number of serious priors. One policy implication may be that the incarcerations do not overcome a proclivity of early offenders to commit serious crimes. Another policy implication may be that the criminogenic impact of prison coupled with its impact of subsequent employment dilute incapacitation and potential deterrence.[21]

The conventional standard error for the slope is .021. The Huber-White standard error .032. Even with larger Huber-White standard error, one would reject the null hypothesis that the slope of the population linear approximation was equal to zero. However, one would have more confidence in the test's validity with a somewhat larger sample. Because the Huber-White estimate is substantially different from the conventional standard error, there is evidence that there is misspecification of first order (nonlinearity) and/or second order (heteroscedasticity) (Buja et al., 2013).

[21] An offender's current age would seem to be an obvious confounder. Older individuals have more time to accumulate priors. However, whether current age is also related to the age at which a first arrest occurs is an empirical question. And in these data, there is effectively no relatonship. Current age is not a confounder.

In most applications, there would likely be additional predictors included in analysis. Then the estimated slopes would be adjusted in the usual manner. One would be estimating the population hyperplane in an asymptotically unbiased fashion, and each regression coefficient would be an average slope with all other predictors "held constant." The same asymptotically justified tests could follow.

7 Extensions

All of the discussion to this point applies to any conventional parametric regression. Any sensible basis functions can be used as long as they are determined before the data analysis begins.[22] For example, one might decide in advance to fit cubic functions of certain predictors and indicator variables for others. One is not limited to linear relationships between the response and the predictors. In short, one should be able to work with linear or nonlinear approximations as long as they are parametric.

More challenging is nonparametric regression in which tuning parameters are determined by the data. Consider, for example, smoothing splines.[23] There can be one smoothed function with the response for each predictor, and each smoothed function can have its own tuning parameter determining the degree of smoothing.

A key complication is that the tuning parameters can legitimately vary with sample size. With a larger sample, one can fit a more complex function to reduce the bias with no necessary increase in the variance. Smaller values for the tuning parameters follow. Under these circumstances, it is difficult to think about population smoothers, especially if that population has a limitless number of observations. What does a population of limitless size mean for the values of a tuning parameter? And how can any population smoother be a legitimate estimation target for a smoother from a finite sample of a particular size? The sample function and the population function will not be the same.

A major challenge, therefore, is how to arrive at sensible values for the tuning parameters. In the case of a single predictor with a single tuning parameter, Chaudhuri and Marron (2000) propose using a number of different tuning parameter values chosen to represent a range of sensible possibilities. Any given

[22] If they are determined as part of a data analysis in which different mean functions are exained, there will be model selection bias. Estimates from the model selected will be biased, and statistical tests and confidence intervals will be invaldiated. The only question is how serious in practice those problems will be (Berk et al., 2010).

[23] For a single predictor,

$$\text{PSS}(\hat{f}, \lambda) = \sum_{i=1}^{N} [Y_i - \hat{f}(X_i)]^2 + \lambda \int [\hat{f}''(t)]^2 dt. \tag{20}$$

PSS denotes penalized sum of squares to be minimized subject to a penalty tuning parameter λ, and $\hat{f}(X_i)$ is the unknown function to be determined. The integral of the second derivatives over X defines the complexity penalty. The expression can be generalized so that for p predictors there are p such expressions combined additively to form the generalized additive model (Hastie et al., 2009: section 9.1) turning parameter would define a population smoother. That tuning parameter would also be applied to the sample. Then, fitted values from the sample would be an asymptotically unbiased estimate of the fitted, conditional expectations in the population for that turning parameter. The same approach would follow for each tuning parameter value. One would not have single set of fitted values, but a suite of fitted values. The estimation target would not be a approximation line, but an approximation band.

Although this idea has real merit, there are to date serious practical limitations. For example, it is not clear how one would work with more than one predictor if for no other reason than computational demands. One would require a multidimensional grid of tuning parameter values. A second difficulty is constructing proper confidence intervals and statistical tests. Although there has been some progress (Chaudhuri and Marron, 1999), the existing procedures are not designed for conventional confidence intervals or tests and are based on incomplete formal justification. In short, estimation for population smoothers as approximations is an ongoing research topic.[24]

8 Conclusions

The good news is that by and large, one can work with linear and nonlinear parametric approximations using conventional regression software. The bad news is that nevertheless, interpretation of the results requires substantial care.

A major conceptual complication is getting the right standard errors. One can properly interpret the estimated regression coefficients obtained from the usual regression output, but the usual standard errors will be wrong. Fortunately, many statistical packages provide access to Huber-White and/or nonparametric bootstrap standard errors for regression coefficients. As long as one keeps in mind that the estimated approximation and standard errors are only justified asymptotically, proper use can follow. In practice, this means that all bets are off in small samples (e.g., < 50). Then the only legitimate regression enterprise is description of relationships in the data on hand. And that can be very useful. It is still possible to learn lots of interesting things.

Proper interpretation of the results is more challenging. One has an estimate of the approximation only. One does not have an estimate of the true mean function. All substantive conclusions must rest on how instructive the approximation is for the questions being addressed. Recall that in the parametric case, each slope estimates a weighted average slope over the range of a given predictor once that predictor is residualized for all other predictors. But if the average treatment effect is instructive for analyses of randomized experiments, perhaps the average slope is instructive for analyses of observational data.

If one is only working with approximations, why not proceed with a conventional regression analysis and just interpret the regression as an approximation?

[24] Using smoothing results from a sample to estimate nature's true response surface, raises another set of very difficult problems (Chaudhuri and Marron, 1999, section 2; Berk et al., 2013).

We suspect that this is often *de facto* practice. Researchers not really prepared to defend their models as "truth."

There are several reasons why this is a bad idea. First, in conventional regression the estimation target is the true conditional means of Y with $\mathbf{X}$ fixed. Within our formulation, the estimation target is the population linear approximation with $\mathbf{X}$ random. There two different answers to the question "estimates of what?" If one's regression model is only an approximation, the estimator target should be no different.

Second, it follows that in conventional regression, the estimates are likely to be biased in finite samples and asymptotically as well. The bias is undesirable in its own right, and undermines statistical tests and confidence intervals. In our approach, the estimates are unbiased asymptotically. Surely this is preferable.

Third, conventional estimates of the standard errors are likely to be biased in finite samples and asymptotically. All statistical tests and confidence intervals can be very misleading. Huber-White standard error estimates can provide asymptotically unbiased standard error estimates for the regression coefficients, but one is still undercut by the biased estimates of regression coefficients and fitted values. Statistical tests and confidence intervals will not perform as intended. In our approach, the estimated standard errors are asymptotically unbiased and at least in reasonably large samples, statistical tests and confidence intervals will behave as they should. This too should be preferable.

Fourth, in conventional regression, there can be strong incentives to treat regression coefficients as estimates of causal effects even though it is very unusual for a social science causal model to meet the requisite assumptions when the data are observational. Our approximation approach is explicitly agnostic with respect to cause and effect, and there are no claims that one is getting causal effect estimates. In that sense, our approach is conservative.

Finally, the conventional concerns about model misspecification and the properties of the model's disturbances (and all the diagnostics that can follow) are at least substantially diluted. Recall that estimates of the population linear approximation are asymptotically unbiased *even in the presence of confounders.* Valid statistical inference can follow.

In short, to the degree that the many critiques of conventional regression analysis have merit, we offer a constructive option. But one must be prepared to abandon a framework in which with observational data one proceeds as if valid estimates of nature's true conditional means can be routinely obtained. In practice, however, not much is being given up. Such estimates are rarely available anyway.

References

Angrist, J., and S. Pischke (2009) *Mostly Harmless Econometrics.* Princeton: Princeton University Press.

Angrist, J., and S. Pischke (2010) "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24(2).

Berk, R.A. (2004) *Regression Analysis: A Constructive Critique.* Newbury Park: Sage Publications.

Berk, R.A. (2008) *Statistical Learning from a Regression Perspective.* New York: Springer.

Berk, R.A., Brown, L., and L. Zhao (2010) "Statistical Inference After Model Selection." *Journal of Quantitative Criminology* 26(2): 217–236

Berk, R.A., Brown, L., George, E., Traskin, M., Zhang, K., and L. Zhao. (2013) "What You Can Learn From Wrong Causal Models." In S. Morgan (ed.) *Handbook of Causal Analysis for Social Research.* New York: Springer.

Blumstein, A., Cohen, J., and D. Nagin (1977) *Deterrence and Incapacitation: Estimating the Effect of Criminal Sactions on Crime Rates.* Washington, D.C. National Research Council.

Box, G,E.P. (1976) "Science and Statistics." *Journal of the American Statistical Association* 71: 791–799.

Breiman, L. (2001) "Statistical Modeling: Two Cultures," (with discussion). *Statistical Science* 16: 199–231.

Buja, A., Berk, R., Brown, L., George, E., Traskin, M., Zhang, K., Zhao. (2013) "A Conspiracy of Random X and Model Violation against Classical Inference in Linear Regression." Working Paper, Department of Statistics, University of Pennsylvania.

Chaudhuri, P., and J.S. Marron (1999) "SiZer for Exploration of Structures in Curves." *Journal of the American Statistical Association* 94(477): 807–823.

Chaudhuri, P., and J.S. Marron (2000) "Scale Space View of Curve Estimation." *Annals of Statistics* 28(2): 408–428.

Freedman, D.A. (1981) "Bootstrapping Regression Models." *Annals of Statistics* 9(6): 1218–1228.

Freedman, D.A. (1987) "As Others See Us: A Case Study in Path Analysis," (with discussion). *Journal of Educational Statistics* 12: 101–223.

Freedman, D.A. (2004) "Graphical Models of Causation, and the Identification Problem." *Evaluation Review* 28(4): 267-293.

Freedman, D.A. (2005) *Statistical Models: Theory and Practice.* Cambridge: Cambridge University Press.

Hastie, T., Tibshirani, R., and J. Friedman (2009) *The Elements of Statistical Learning*, Second Edition. New York: Springer.

Heckman, J.J., and J.A. Smith (1995) "Assessing the Case for Randomized Social Experiments. "*Journal of Economic Perspectives* 9: 85-110.

Holland, P.W. (1986) "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945–960.

Imbens, G., and J.D. Angrist (1994) "Identification and Estimation of Local Average Treatment Effects" *Econometrica* 62(2): 467–475.

Imbens, G. (2009) "Better Late Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48 (June): 399-423.

Leamer, E.E. (1983) "Let's Take the Con of Econometrics." *American Economics Review* 73: 31–43.

Leeb, H. and B.M. Pötscher (2005) "Model Selection and Inference: Facts and Fiction." *Econometric Theory* 21: 21–59.

Leeb, H., B.M. Pötscher (2006) "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *The Annals of Statistics* 34(5): 2554–2591.

Long, J.S., and Ervin, L.H. (2000) "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." *American Statistician* 54(3): 217-224.

Manski, C.F., (2003) *Partial Identification and Probability Distributions.* New York: Springer.

Morgan, S.L, and C. Winship (2007) *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge: Cambridge University Press.

Nagin, D.S. and J.V. Pepper (2012) *Deterrence and the Death Penalty.* Washington,D.C. National Research Council.

Rosenbaum, P. (2002) *Observational Studies*, Second Edition. New York: Springer-Verlag.

Rosenbaum, P. (2010) *The Design of Observational Studies.* New York: Springer-Verlag.

Rubin, D.B. (1986) “Which Ifs Have Causal Answers?” *Journal of the American Statistical Association* 81: 961–962.

Rubin, D.B. (2008) “For Objective Causal Inference, Design Trumps Analysis.” *The Annals of Applied Statistics* 2(3): 808–840.

Welford, C.F. Pepper, J.V. and C.V. Petrie (2005) *Firearms and Violence: A Critical Review.* Washington, DC: National Resaerch Council.

White, H. (1980) “Using Least Squares to Approximate Unknown Regression Functions.” *International Economic Review* 21(1): 149–170.