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 and Linda Zhao

Sociological Methods & Research 2014 43: 422 originally published online 31 March 2014
DOI: 10.1177/0049124114526375

Regular Article

Sociological Methods & Research
2014, Vol. 43(3) 422-451

DOI: 10.1177/0049124114526375
smr.sagepub.com

Misspecified Mean Function Regression: Making Good Use of Regression Models That Are Wrong

Richard Berk[1,2], Lawrence Brown[1], Andreas Buja[1], Edward George[1], Emil Pitkin[1], Kai Zhang[1] and Linda Zhao[1]

Abstract

There are over three decades of largely unrebutted criticism of regression analysis as practiced in the social sciences. Yet, regression analysis broadly construed remains for many the method of choice for characterizing conditional relationships. One possible explanation is that the existing alternatives sometimes can be seen by researchers as unsatisfying. In this article, we provide a different formulation. We allow the regression model to be incorrect and consider what can be learned nevertheless. To this end, 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. There exist estimators that are asymptotically unbiased and standard errors that are asymptotically correct even when there are important specification errors. Both can be obtained easily from popular statistical packages.

[1] Department of Statistics, University of Pennsylvania, Philadelphia, PA, USA
[2] Department of Criminology, University of Pennsylvania, Philadelphia, PA, USA

Corresponding Author:
Richard Berk, Department of Statistics and Department of Criminology, University of Pennsylvania, Philadelphia, PA, USA.
Email: berkr@wharton.upenn.edu

Keywords

random predictors, linear models, model misspecification, regression models, misspecified mean function regression

Introduction

There is a large literature on the many difficulties with regression modeling in the social sciences (e.g., Angrist and Pischke 2010; Berk 2004; Box 1976; Breiman 2001; Freedman 1987; Holland 1986; Imbens 2009; Leamer 1983; Rubin 1986, 2008). By and large, this literature is unrebutted (Freedman 2005: section 8.9). Yet, even revised editions of popular methods textbooks continue to appear with new material essentially appended to the old (e.g., Aron, Coups, and Aron 2010; Greene 2011; Stock and Watson 2010). Rarely are fundamentals revisited except to tinker around the margins or to attach more elaborate formulations. Research practice proceeds in much the same manner.

Studies on the deterrent effect of capital punishment are an instructive illustration because the literature is an interdisciplinary product of economics, sociology, criminology, political science, and law. In 1977, a National Research Council committee charged with reviewing the relevant 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 et al. 1977:62). But researchers proceeded with regression modeling as usual. Thirty-four years later, another National Research Council committee was given a similar charge. Although there were strong criticisms of the 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 would so many social science researchers maintain their attachment to traditional 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 regression 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 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 parameter 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 summarize 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 formulation has much in common with the "correlation model" favored by Freedman (1981) and with procedures developed by White (1980). Its foundations are similar to those found in computational learning theory from computer science (Bishop 2007: section 7.1.5; Vapnick 1998). Angrist and Pischke (2009: section 3.1.2) provide very accessible motivation for regression as approximation.

The second section reviews briefly some key properties of the linear regression mean function with fixed predictors. This is the conventional formulation and provides an important baseline. The third section considers the regression mean function when predictors are random. We introduce key issues at a broad conceptual level and show how predictors that are random

variables can do more than simply increase conventional uncertainty. When coupled with a misspecified mean function, the joint distribution of the predictors can perversely affect regression estimates. This is an additional source of bias that is irrelevant in fixed-X regression. In the fourth section, 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. The fifth section turns to implementation and practice. In particular, we focus on the meaning of regression coefficients in the setting characterized in the third and fourth sections. In the sixth section, there is an example using real data. The example is kept simple so that instructive visualizations can be provided. The seventh section briefly broadens the discussion to parametric nonlinear regression including the generalized linear model. The earlier discussion by and large carries over. The eighth section offers some broad conclusions. Rather than trying to estimate the parameters of the "correct" model, it can be more honest, more instructive, and even liberating to estimate the parameters of its linear approximation, even though the linear approximation is assumed to be "incorrect." Omitted variables may imply that one's findings are incomplete. But unlike conventional regression, omitted variables do *not* jeopardize the desirable statistical properties of our estimators. These conclusions apply to model endogeneity more generally, at least as commonly characterized by social scientists.

Conventional Linear Regression With Fixed Predictors: The Conventional Baseline

In this section and the next, we raise issues that help motivate the rest of the article. 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 the fourth section.

The standard formulation for linear regression takes the following form:

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

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, independent draw from a distribution of disturbances that has a mean of zero 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 independently 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: $\mu_i|\boldsymbol{X_i}$.[2] These conditional means are found in the real process that nature employs to generate the response. Unbiased estimates of β and σ^2 require that equation (1) be 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] Given a model that is first-order correct, 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.

Regression Mean Functions With Fixed X

For initial 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, which for now is our primary concern, can be more easily represented and studied.

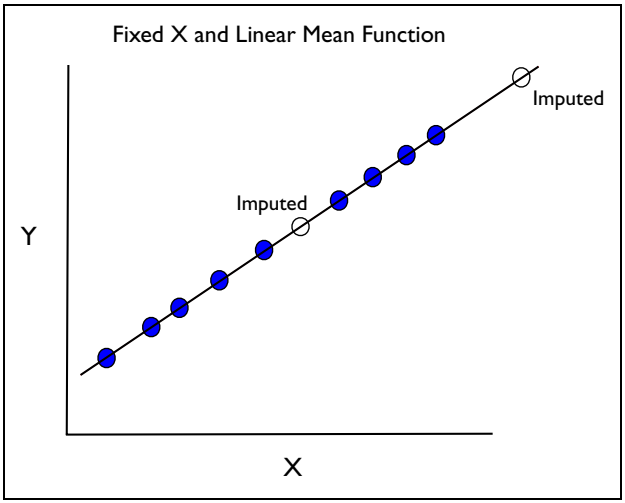

Figure 1. The canonical regression formulation with *Y*, a deterministic linear function of a fixed *X*.

Figure 1 shows the relationship between the response and that single predictor for the conventional linear model. The black line is nature's response function; it represents the true relationship between the predictor and the response. The blue circles are hypothetical observations for the response at some predictor values assuming no disturbances. The observations are hypothetical because we are showing a part of nature's machinery, not a conventional scatter plot constructed from real data.

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 location of the predictor values is unrelated to the estimated slope $\hat{\beta}$, which always corresponds to the true slope β. Whether, for instance, the predictor values are skewed to the left or to the right does not affect the regression coefficient. In practice with real data, one has an unbiased estimate of the same population regression line. Nor does it affect the validity of conventional statistical tests.

It follows that one can condition on the predictor values, the usual practice, and obtain correct regression coefficients from the data. This 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).

Regression Mean Functions With Random *X*

Fixed regressors can be an underappreciated constraint on the conventional regression model. From the fixed-regressor perspective, uncertainty is solely a function of the disturbances. Regression estimates are seen as varying over realizations of the response values with the predictor values unchanged.

Complications 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).[5]

The conventional response is to treat predictor values as fixed once they materialize in the sample and to condition on the observed predictor values. 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.[6] Still, as long as the true 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 again 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.

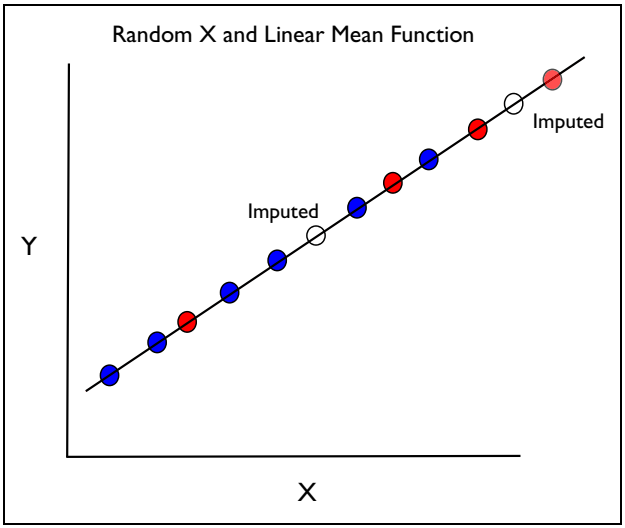

Figure 2. The canonical regression formulation with *Y*, a deterministic linear function of a random *X*.

Figure 3 tells a much darker and more 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.

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. $\hat{\beta}$ now depends on the *nature's predictor distribution*. For example, if that predictor distribution is concentrated at low values, blue circles rather than red circles are more likely to be realized. If that predictor distribution is concentrated at high values, the reverse is true. Hence, the *expected value* of the slope can differ depending on the nature's predictor distribution. Moreover, the usual statistical tests will not perform as they should. These are not issues for the conventional linear regression model because nature's predictor distribution is treated as "ancillary" (Cox and Hinkley 1974).

Here is the reasoning in more detail. For the moment, imagine a large population that could be generated by nature. Imagine being able to compute

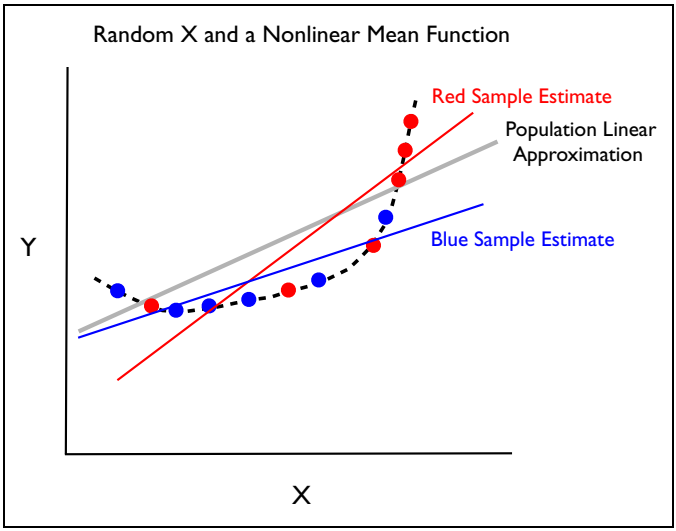

Figure 3. Nonlinear mean function and random *X*.

a bivariate least squares regression in that population. That is, one can treat the population values of the response and the predictor as a data set analyzed by least squares regression. Then, one can quite properly interpret the regression as a feature of that population. In the population, the nonlinear mean function is still the correct functional form. The population regression is a best linear approximation of the true mean function. The approximation is shown as the straight gray line in Figure 3.

Suppose 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 the least squares regression computed from the red data or the blue data failing systematically to get the population linear approximation right.

This disappointing result is shaped on two factors. As already noted, when researchers apply least squares regression, they condition on the predictor values realized in the sample. There is no allowance for any predictor values beyond those actually observed. In addition, any random subset of predictor values provides access to only a random piece of that nonlinear truth. A complete picture of how the response and predictor are related is unavailable; the researcher is necessarily working with fragmentary information.

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 true nonlinear mean

function and random variable predictor, a linear mean function computed from real data will misrepresent the nonlinear population mean function and likely misrepresent a linear population least squares mean function as well. For either estimation target, $\hat{\beta}$ will be biased.

The bias with respect to the population linear approximation can be small, especially in moderate to large samples. We more formally consider this result shortly. We also consider how with random **X**, the combination of a population nonlinear mean function and a population linear least squares approximation leads to a nasty distortion of the disturbances associated with the population linear approximation, regardless of sample size. Conventional standard errors may no longer be valid. Misleading statistical tests and confidence intervals can follow.

Building on random **X**, there is nevertheless a defensible way to proceed. The true mean function is taken to be unknown and not necessarily linear. One accepts that the true mean function cannot be properly estimated from the data. The population linear approximation of the truth has useful substantive information nevertheless even though 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. We will see that it is possible to obtain suitable estimates of the linear approximation and appropriate standard errors, at least in large samples.

We turn now to a more technical discussion to make the rationale and results much more precise. Readers interested primarily in practical implications may wish to skip to the fifth section.

Conceptual Formalities

We expand our discussion to regression with more than one predictor. There is 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 rather like sample means (usually called expectations), variances, and covariances. It can be instructive, therefore, to treat 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 insight 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 $\mathbf{X}$, 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.[7]

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" $\mu(\vec{\boldsymbol{X}})$ in nature's population, that is the *expectation* of the response conditional on given values for the predictors $\vec{\boldsymbol{X}}$. More formally,

$$\mu(\vec{\boldsymbol{X}}) = E[Y|\vec{\boldsymbol{X}}]. \tag{2}$$

A possible population target for estimation is a set of expected values one might conventionally denote by $\mu_i|\boldsymbol{X}_i$. 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 through 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* ordinary least squares linear approximation of the response variable's conditional expectations.

$$\beta^T \vec{\boldsymbol{X}} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, \tag{3}$$

where

$$\beta = \text{argmin}_{\tilde{\beta}} E[(Y - \tilde{\beta}^T \vec{\boldsymbol{X}})^2] = E[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1} E[\mu(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}]. \tag{4}$$

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. One has in the population the best linear approximation of the truth.
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 $\beta_{j\bullet}$:

$$\beta_{j\bullet} = \text{argmin}_{\tilde{\beta}} E[(X_j - \tilde{\beta}^T \vec{\boldsymbol{X}}_{notj})^2] = E[\vec{\boldsymbol{X}}_{notj} \vec{\boldsymbol{X}}_{notj}^T]^{-1} E[\vec{\boldsymbol{X}}_{notj} X_j]. \tag{5}$$

Then the adjusted jth predictor is[8]

$$X_{j\bullet} = X_j - \beta_{j\bullet}^T \vec{X}_{notj}. \tag{6}$$

Each predictor in turn is regressed on all other predictors. Each set of fitted values is then subtracted from its corresponding predictor. Linear dependence between each predictor and all others is removed. Finally, the population regression coefficient for each "residualized" predictor is given by:

$$\beta_j = \frac{E[YX_{j\bullet}]}{E[X_{j\bullet}^2]}, \tag{7}$$

which is just the jth component of equation (4).
6. With respect to the population response surface in equation (2), the population linear approximation in equation (4) has a mean function that is misspecified. We introduce an explicit allowance for the linear

mean function error $\eta(\vec{X})$ that is responsible for disparities between the two:

$$\eta(\vec{X}) = \mu(\vec{X}) - \beta^T\vec{X}. \quad (8)$$

Both terms to the right of the equal sign are random variables because $\vec{X}$ is random. Hence, the difference between the two right-hand side 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

$$\epsilon = Y - \mu(\vec{X}). \quad (9)$$

Even if one knows the true response surface, the population fit of Y will not likely be exact. The best one can possibly 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.[9]

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,

$$\xi = Y - \beta^T\vec{X} = \eta(\vec{X}) + \epsilon. \quad (10)$$

Moreover, ξ is uncorrelated with $\vec{X}$ because the population linear approximation is derived from ordinary least squares.[10] There can be no endogeneity in the population linear approximation.

Figure 4 is a visual aide with one predictor. The difference between a hypothetical value of the response and the population mean function $\eta(\vec{X})$ 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]

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),

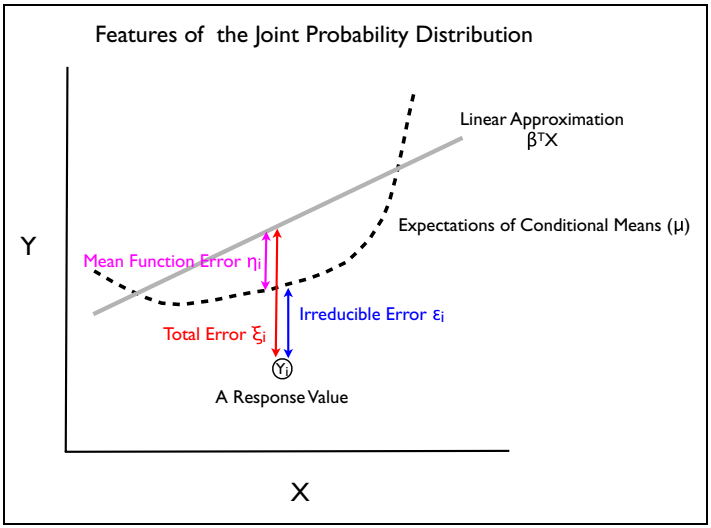

Figure 4. A decomposition of "total error" in the population.

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T. \tag{11}$$

$$\text{argmin}_{\tilde{\beta}}(Y - \mathbf{X}\tilde{\beta})^2. \tag{12}$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y. \tag{13}$$

This is just ordinary least squares. The intent is *not* to estimate nature's true mean function. The intent is to estimate the best population linear *approximation* of $\mu|\boldsymbol{X}$. We have given up on trying to estimate the true mean function, and the model we are applying is explicitly permitted to be wrong. Nevertheless, several of 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:

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

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

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

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 the assumptions underlying conventional least squares regression.

Working With the Linear Approximation

A commitment to estimation of population linear approximations rather than the true response surface leaves unaddressed what subject matter sense one is to make of the regression output from a data set. All of the usual output is still available. Now what?

Interpreting the Regression Coefficients

The estimated regression coefficients are just the usual slopes of a linear least squares fit. But because we are drawing inferences about the best linear least squares fit within a joint probably distribution, the interpretation is a somewhat different. Each slope $\hat{\beta}_j$ is an estimate of 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.

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

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

Equation (17) unpacks the estimation target. The left-hand fraction in brackets is the slope for any hypothetical case shown as 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 right-hand fraction in the brackets, the ratio of the squared deviation score $(X - E(X))^2$ and the expected value of such squared deviation score in the population. The latter is a fixed constant, and the weights sum to 1.0. Slopes of line segments farther from the $E(X)$ are given more weight and have greater influence.

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

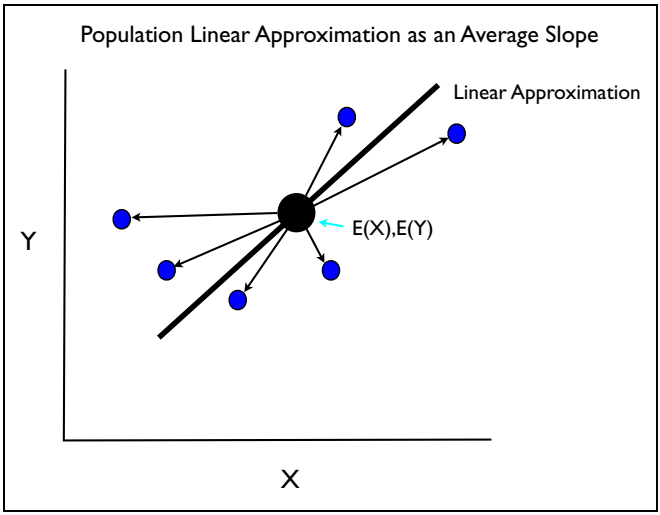

Figure 5. Linear approximation slope as an average of slopes for a single predictor.

to be nonlinear.) In Figure 5, there are six line segments, one for each hypothetical 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. In the joint probability distribution, best linear approximation, shown with a thick black line, has a slope that is the weighted average of the 6. It gets the slope wrong for each hypothetical observation. Nevertheless, it may be an instructive summary of how X and Y are related within their joint probability distribution.[13]

Given our earlier discussion, it is easy to consider how one interprets the slope when there is more than one predictor. The same interpretation of equation (17) 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. Linear dependence among the predictors has been removed.

One must be careful not to ask more of the linear approximation than it can deliver. First, if the true response surface is nonlinear, no linear approximation can properly capture it. The linear approximation in the joint probability distribution is wrong from the start.

Second, the impacts of any omitted variables that are true confounders[14] are absorbed in the regression coefficients for the population linear approximation. The residualization process cannot address that confounding. One's estimation target is the population linear approximation with its omitted variable warts and all. The same reasoning applies to any sources of endogeneity. Although this is no doubt disappointing, it forces researchers to squarely face some real limitations in their regression estimates. It also imposes a direct correspondence between the regression equation applied to data and the regression equation to be estimated. A happy result, discussed shortly, is that estimators with good statistical properties can follow.

Third, 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 requires the same slope for each case.

Because of these three limitations, any step from estimation to causal inference will be challenging. There are the usual conceptual issues such as how to map covariance-adjusted regression coefficients to real-world manipulations of causal variables. But in addition, under what circumstances does it make sense to use linear approximations as the basis for any causal claims?

There may be some important precedents from randomized experiments in which 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 can be seen in the same spirit. But the issues are complicated, and we have yet to consider them in sufficient depth.

Estimation

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

Unfortunately, a difficulty arises that does not exist when $\mathbf{X}$ is considered fixed: The dependence of $\hat{\beta}$ on $\mathbf{X}$ is nonlinear. As a result, $E(\hat{\beta}) \neq \beta$.The nonzero difference is a finite-sample bias that exists whenever $\mu(\boldsymbol{X})$ is not linear in $\mathbf{X}$. It is purely a consequence how the calculations are done coupled with

random $\mathbf{X}$. This bias, however, is of smaller order than the standard error of $\hat{\beta}$ and should not materially matter as long as the number of observations per regression coefficient is not too small.[15]

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

The joint distribution of the estimated regression coefficients is asymptotically normal. The marginal distributions are as well. This means that the stage is nearly set of statistical inference, at least in large samples, a result that has long been exploited in econometrics.

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., homoscedastic). To emphasize this point, the expression for the misspecification disparities in the joint distribution is reproduced subscripted for each potential observation.

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

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. Hypothetical 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 for Y_2. Because the true mean function is nonlinear, its

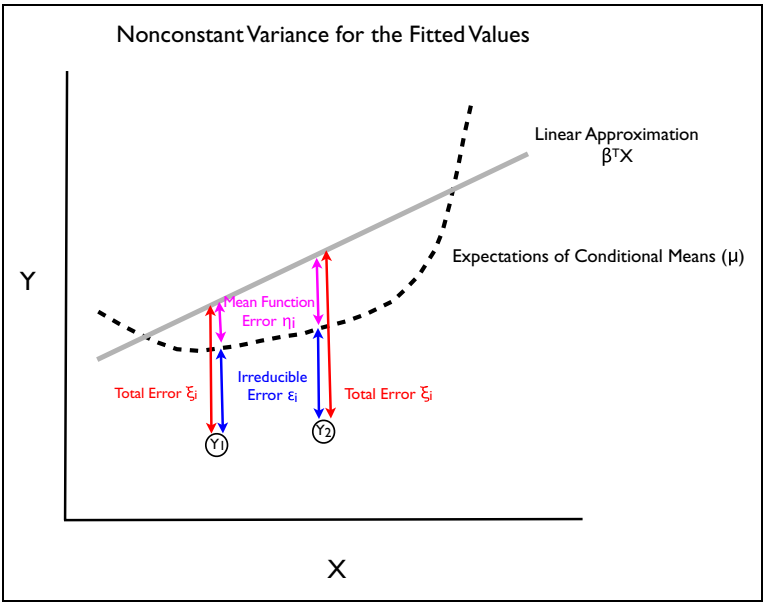

Figure 6. Source of nonconstant variance in linear approximation estimates.

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 in the joint distribution is "noncentered" variance around the linear approximation that as an empirical matter can behave much like nonconstant variance. As a result, 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. Conventional standard errors are still incorrect.

Huber–White standard errors. There are two good ways to obtain *asymptotically* valid standard errors (Berk et al. 2013; Buja et al. 2013). The first uses

Huber–White robust standard errors, sometimes called the "sandwich estimator." Its trick uses squared residuals to estimate the mean squared error between the response and linear approximation to account for both the nonlinearity and the heteroscedastic irreducible errors.[16]

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

$$\mathrm{V}C_{\hat{\beta}} = E[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}E[\delta^2(\vec{\boldsymbol{X}})\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]E[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}, \quad (19)$$

where δ^2 is the conditional mean squared error at $\mathbf{X}$ that includes the variance of the disturbances of the true mean function and the variance resulting from the linear approximation's specification error.

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

Bootstrap standard errors. The bootstrap is essentially a simulation of the frequentist thought experiment. There are two approaches. The "residual bootstrap" 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 are B samples drawn from the data set sometimes called bootstrap samples, $s^{*1}, s^{*2}, \ldots, s^{*B}$. The samples are generated by random sampling of full cases with replacement; both the values for Y and $\mathbf{X}$ are included. More complicated sampling designs are employed 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. If there are N observations in the dataset, there are N observations in each bootstrap sample.
2. There are plug-in estimates one computes for each of the B bootstrap samples $t(Y^{*1}, \mathbf{X}^{*1}), t(Y^{*2}, \mathbf{X}^{*2}), \ldots, t(Y^{*B}, \mathbf{X}^{*B})$. For example, from

each bootstrap sample one might compute regression coefficients. They are called "plug-in" because they are computed in the same manner one would use if there were access to the population from which the original sample was drawn.

3. 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. Some may find the Huber–White standard errors easier to compute, but there is readily available software for both. We are exploring whether the two approaches have different performance characteristics in samples of the size one often sees in the social sciences.

A Simple Example

Consider variables for individuals on probation in a large city. We assume that the individuals represent an independent and identically distributed sample from a population, which allows one to treat the data as realizations from a 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 used in this example:

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 or 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?

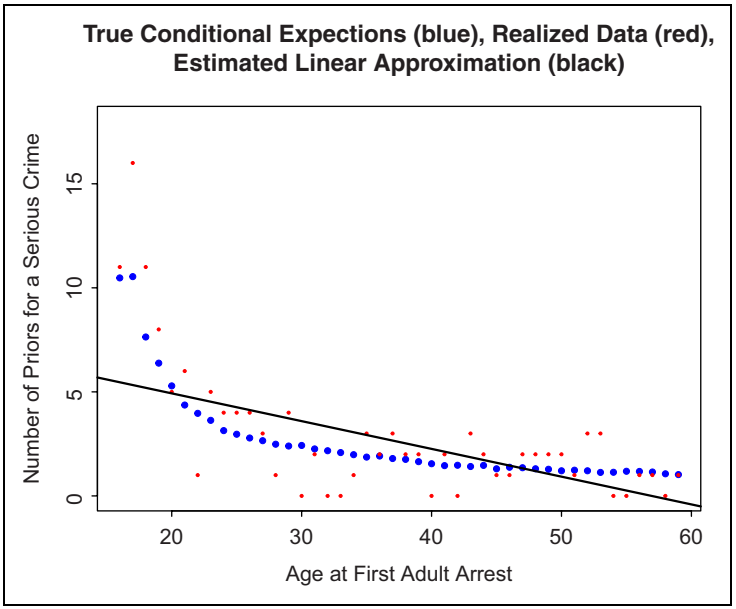

Figure 7. A linear approximation in practice.

Our primary goals are to show visually key features of our approach and to illustrate what sorts of substantive issues that can be usefully addressed. Including more predictors is no doubt a better approximation of usual practice, but introduces many extra details without adding much additional insight.

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 in the available data. Those missing predictors might help explain why the conditional expectations for the number of serious priors do 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 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?

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).

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.

Extensions

All of the discussion to this point applies to any conventional linear regression. But, any sensible functions can be used for the population approximation as long as they are determined before the data analysis begins.[22] For example, one might decide in advance that the approximation in the joint probability distribution should be represented by the log of a linear combination of predictors. As a descriptive matter, interesting features of any associations will perhaps be captured that otherwise would have been missed. Whether the new formulation is a more accurate rendering of the true response surface is unknown. The regression equation one computes with the data is then formulated in the same fashion. One is not limited to linear relationships between the response and the predictors. In short, it is possible to work with linear or nonlinear approximations as long as they are parametric.

Much the same rationale may apply to the entire generalized linear model including logistic regression and Poisson regression (Huber 1967; White 1982). But there are details to be worked out that we are currently exploring. There are also questions about the best way to obtain valid standard errors in studies with small to modest sample sizes.

Conclusion

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 concern 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., <100) or when the sample size is not least several times larger than the number of predictors. 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 linear 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 as noted earlier, if the ATE is instructive for analyses of randomized experiments, perhaps the average slope is instructive for analyses of observational data. Nonlinear approximations can be interpreted in the same spirit, but there will not be a single slope.

If one is only working with approximations, why not proceed with a conventional regression analysis and just *interpret* the regression as a linear approximation? We suspect that this is often de facto practice. Researchers are often unprepared to defend their models as truth.

There are several reasons why reinterpreting conventional regression as a linear approximation 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 best linear approximation in a joint probability distribution with $\mathbf{X}$ random. There are two different answers to the question "estimates of what?" If one's regression estimates are only from a linear approximation, it seems that the estimation 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; the regression estimates used in the tests and confidence intervals will be systematically too large or too small. 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 usual concerns about model misspecification, endogeneity, and other properties of the model's disturbances (and all the diagnostics that can follow) are at least substantially diluted. For example, recall that estimates of the best linear approximation in the joint probability distribution are asymptotically unbiased *even in the presence of unknown 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.

Acknowledgment

Thanks go to John MacDonald, Charles Loeffler, Chris Winship, and three anonymous referees for helpful comments on an earlier draft of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was supported in part by NSF Grant DMS-1007657.

Notes

1. $\boldsymbol{X}$ has N rows and $p + 1$ columns, one for the intercept.
2. That is, $\mu_i|\boldsymbol{X}_i = \boldsymbol{X}_i\beta$.
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. 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**. Predictors and functions of predictors are no longer constants. The predictor matrix is a random matrix, which means that the empirical predictor covariance matrix is also random. As a result, the expectations of the regression estimates depend upon the expectations of the empirical predictor covariance matrix. This is where the estimator nonlinearity enters (Rice 2007: section 14.6). Least squares regression estimators are no longer BLUE because they are no longer linear.
6. This is necessarily true when the predictors are fixed.
7. The researcher really cares about **Z** as measured, commonly recast as Y and **X**. **Z** is not a set of imperfect measures of some other set of random variables **W** that the researcher actually wants to study. Therefore, just as in conventional regression, there is no measurement error. This simplification creates no problems as long as the researcher is comfortable working with **Z** as measured. Allowing for the study of **Z** to be formally a study of **W**, introduces new layers of complexity that are beyond the scope of this article. Fuller's (1987) early treatment remains an excellent reference.
8. The subscript *notj* denotes all predictors but the *j*th predictor. The intercept β_0 is also subject to adjustment, but it is still interpreted as a constant.
9. In general, the "pure noise" ϵ is not stochastically independent of the predictors, though it is uncorrelated with them. Also, its conditional mean is zero.
10. The orthogonality between ξ and $\vec{X}$ is built in, just as it is for ordinary least squares estimates from a sample.
11. In equation form, $Y = \beta^T \vec{X} + \eta(\vec{X}) + \epsilon$.
12. The term "hypothetical" is used because the population is a joint probability distribution. There are no realized observations.
13. It is possible to give the slope of the population linear approximation other interpretations (Yitzhaki 1996), but none translate easily into conventional social science parlance, especially when there is more than a single predictor.
14. They are correlated with the response variable and one or more predictors.
15. With larger samples, 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 estimated linear approximation.
16. There are several proposals that appear to improve the performance 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.
17. In practice, one would have to establish that the composition of the probationer population and the process by which individuals were sentenced to probation did not change in important ways over that five-year period.

18. For this illustration, the blue dots are the means of the response for each arrest age from a real data set with nearly 200,000 observations. With so large a data set, 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.
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 relationship. Current age is not a confounder.
22. If they are determined as part of a data analysis in which different mean functions are examined, there will be model selection bias. Estimates from the model selected will be biased, and statistical tests and confidence intervals will be invalidated. The only question is how serious in practice those problems will be (Berk, Brown, and Zhao 2010).

References

Angrist, J. and S. Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: 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* 24:3-30.

Aron, A., E. Coups, and E. N. Aron. 2010. *Statistics for the Behavioral and Social Sciences*. 5th ed. New York: Pearson.

Berk, R. A. 2004. *Regression Analysis: A Constructive Critique*. Newbury Park, CA: Sage.

Berk, R. A., L. Brown, E. George, M. Traskin, K. Zhang, and L. Zhao. 2013. "What You Can Learn from Wrong Causal Models." Pp. 403-27 in *Handbook of Causal Analysis for Social Research*, edited by S. Morgan. New York: Springer.

Berk, R. A., L. Brown, and L. Zhao. 2010. "Statistical Inference after Model Selection." *Journal of Quantitative Criminology* 26:217-36.

Bishop, C. M. 2007. *Pattern Recognition, and Machine Learning*. New York: Springer.

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

Box, G. E. P. 1976. "Science and Statistics." *Journal of the American Statistical Association* 71:791-99.

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

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

Cox, D. R. and D. V. Hinkley. 1974. *Theoretical Statistics*. New York: Chapman & Hall/CRC.

Freedman, D. A. 1981. "Bootstrapping Regression Models." *Annals of Statistics* 9: 1218-28.

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* 28:267-93.

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

Fuller, W. A. 1987. *Measurement Error Models*. New York: John Wiley.

Greene, W. H. 2011. *Econometric Analysis*. 7th ed. New York: Prentice Hall.

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* 81:945-60.

Huber, P. J. 1967. "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions." Proceedings of the Fifth Berkeley Symposium in Mathmatical Statistics and Probability. University of California Press, Berkeley, CA.

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

Imbens, G. and J. D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62:467-75.

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

Long, J. S. and L. H. Ervin. 2000. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." *American Statistician* 54:217-24.

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, United Kingdom: Cambridge University Press.

Nagin, D. S. and J. V. Pepper. 2012. *Deterrence and the Death Penalty*. Washington, DC: National Research Council.

Rice, J. A. 2007. *Mathematical Statistics and Data Analysis*. 2nd ed. Belmont, CA: Brooks/Cole.

Rosenbaum, P. 2002. *Observational Studies*. 3rd ed. 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* 81:961-62.

Rubin, D. B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *The Annals of Applied Statistics* 2:808-40.

Stock, J. H. and M. W. Watson. 2010. *Introduction to Econometrics*. 2nd ed. New York: Addison-Wesley.

Vapnick, V. N. 1998. *Statistical Learning Theory*. New York: John Wiley.

White, H. 1980. "Using Least Squares to Approximate Unknown Regression Functions." *International Economic Review* 21:149-70.

White, H. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50:1-25.

Yitzhaki, S. 1996. "On Using Linear Regressions in Welfare Economics." *Journal of Business & Economic Statistics* 14:478-86.

Author Biographies

Richard Berk is a professor in the Department of Criminology and the Department of Statistics at the University of Pennsylvania.

Lawrence Brown is a professor in the Department of Statistics at the University of Pennsylvania.

Andreas Buja is a professor in the Department of Statistics at the University of Pennsylvania.

Edward George is a professor in the Department of Statistics at the University of Pennsylvania.

Emil Pitkin is a graduate student in the Department of Statistics at the University of Pennsylvania.

Kai Zang is an assistant professor in the Department of Statistics at the University of North Carolina.

Linda Zhao is a professor in the Department of Statistics at the University of Pennsylvania.