Working With Misspecified Regression Models

Richard Berk[a,b], Lawrence Brown[b], Andreas Buja[b],
Edward George[b], Linda Zhao[b]

Department of Criminology[a]
Department of Statistics[b]
University of Pennsylvania

January, 2017

Abstract

Objectives: Conventional statistical modeling in criminology assumes proper model specification. Very strong and unrebutted criticisms have existed for decades. Some respond that although the criticisms are correct, there is for observational data no alternative. In this paper we provide an alternative.

Methods: We draw on work in econometrics and statistics from several decades ago, updated with the most recent thinking to provide a way to properly work with misspecified models.

Results: We show how asymptotically, unbiased regression estimates can be obtained along with valid standard errors. Conventional statistical inference can follow.

Conclusions: If one is prepared to work with explicit approximations of a "true" model, defensible analyses can be obtained. The alternative is working with models about which all of the usual criticisms hold.

1 Introduction

The generalized linear model and its extensions have long been a workhorse for empirical research in criminology. The appeal is clear. The righthand side

is a linear combination of regressors that is easy to interpret. Depending on the disturbance distribution chosen, the response can be numerical or categorical. Conventional statistical tests and confidence intervals can follow, and the regression coefficients can sometimes be given causal interpretation. It is no surprise that two recent issues of *Criminology* (Volume 54, Issues 1 and 2, 2016) have 9 regression applications out of 10 research articles.

But the ease of use is deceptive. Powerful critiques of regression in practice have been widely available since at least the 1970s (e.g., Leamer, 1978, Rubin, 1986; 2008; Freedman, 1987; 2004; Berk, 2003). David Freedman's excellent text on statistical models (2009) can be consulted for an unusually cogent discussion. Moreover, there apparently has never been an effective rebuttal. Freedman (2009: 195) provides an illustrative list of comebacks he received over the years to his criticisms of conventional regression analysis practice.

> *We all know that. Nothing is perfect. Linearity has to be a good first approximation. Log linearity has to be a good first approximation. The assumptions are reasonable. The assumptions don't matter. The assumptions are conservative. You can't prove the assumptions are wrong. The biases will cancel. We can model for the biases. We're only doing what everybody else does. Now we use more sophisticated techniques. If we don't do it, someone else will. What would you do? The decision-maker has to be better off with us than without us. We all have mental models, not using a model is still a model. The models are not totally useless. You have to do the best you can with the data. You have to make assumptions to make progress. You have to give the model the benefit of the doubt. Where's the harm?*

Clearly, Freedman is having some fun while underscoring the lack of real substance from those defending conventional regression practice. But he is also missing an important message: conventional practice can recognize and accept that requisite assumptions are not met and that the empirical results derive from a misspecified model. Criminology craft lore in particular permits working with approximations of the truth.

Yet, justifications of regression by approximation require far more than craft lore. One needs a formal mathematical rationale. Such a rationale was first offered by White (1980a) and Freedman (1981). Accessible summaries

followed (Angrist and Pischke, 2008; Berk et al., 2014b). Buja and his colleagues (2016) recently have developed important extensions. There can be a formal justification for regression by approximation after all.

In this paper, we begin with a review of the criticisms of conventional regression practice. For ease of exposition, we will use the linear regression model, but the problems identified apply, with some modest alterations, to the generalized linear model and multiple equation extensions such as hierarchical linear models and structural equation models. We follow with a discussion of how to justify and make sense of misspecified regression models. The takeaway message is this: there will be many situations in which regression approximations can be appropriate and instructive, but some important revisions of common interpretations are required.

2 Revisiting the Ubiquitous Linear Regression Model

We need to set the stage for regression by approximation with a brief review of the traditional linear regression formulation followed by a short discussion of some of its most telling criticisms. Conventional notation is used.

Y is an $N \times 1$ numerical response variable, sometimes called a dependent variable or an endogenous variable.[1] N is the number of observations. There is an $N \times (p + 1)$ "design matrix" $\mathbf{X}$, where p is the number of predictors, sometimes called regressors, independent variables, or exogenous variables. A leading column of 1s is usually included in $\mathbf{X}$ for the intercept coefficient. Y is a random variable. In this formulation, the p predictors in $\mathbf{X}$ are fixed variables. Whether predictors are fixed or random is not a technical detail, and figures substantially in subsequent material.[2]

[1] This section draws heavily on Berk's textbook on statistical learning (2016: Section 1.3).

[2] In this context, the predictors are treated as fixed variables if in new realizations of the data, their values do not change. This is the approach in conventional regression. It simplifies the mathematics, but at a substantial interpretative price; the regression results can only be generalized to new observations produced by nature in the same fashion with *exactly* the same x-values. In contrast, predictors are treated as random variables if in new realizations of the data, their values changes in an unsystematic manner (e.g., through the equivalent of random sampling). This complicates the mathematics, but one gains the ability to generalize the regression results to new observations produced by nature in the same fashion but with different x-values. To take a cartoon illustration, if a predictor is

The value of Y for the ith case is realized from a linear function that takes the form,

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

where

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

Conventionally, β_0 is the y-intercept associated with the leading column of 1s. There are regression coefficients $\beta_1, \beta_2, \ldots, \beta_p$, and a random perturbation ε_i. One can say that for each case i, nature determines the values of the predictors, multiplies each such value by its corresponding regression coefficient, adds these products, adds the value of the constant, and finally, adds a random perturbation. Each perturbation, ε_i, is a random variable realized as if drawn randomly and independently from a single distribution, often assumed to be normal, with a mean of 0.0. Nature behaves as if she appropriates a linear model, and Equations 1 and 2 are, therefore, a bonafide theory of how some process works. Equations 1 and 2 are not merely a statistical convenience.

The values of Y for each case i can be realized repeatedly because, given X, its values will vary solely because of ε. The predictor values are fixed. For example, one can imagine that a given defendant could have a limitless number of sentence lengths, solely because of the "noise" represented by ε_i. Nothing else in nature's linear combination would change: the defendant's prior record, conviction offense, age, martial status, and so on. This is more than a statistical formality. It is an essential part of the theory for how sentences are determined.[3]

age, and the values in the dataset are ages 24, 25, 30, 31,32 and 35, these are the only ages to which generalizations are permitted even if the true relationship is really linear. Should one want to apply results to, say, a 26 year old, one has to alter the mathematics to allow for realizations of ages that were not in the data. In other words, one has to allow for the x-values to have been different. This introduces a new source of uncertainty not addressed in the usual, fixed-x regression formulation. If one's regression model is correctly specified, the impact of the additional uncertainty can be in practice small. But as we shall see, it matters a great deal if one wants to allow properly for model misspecification (Freedman, 1981).

[3]If on substantive grounds one allows for nature to set more than one value for any given predictor and defendant, a temporal process may be implied. Then, there is systematic temporal variation to build into the regression equation. This can be done, but the formulation is more complicated, requires that nature be still more cooperative, and for the points to be made here, adds unnecessary complexity.

It is important to distinguish between the mean function and the disturbances (also called the residual error). The mean function is the expectation of Equation 1.[4] A conventional linear regression model is "first order correct" when Equation 1 is *literally* what nature used to generate the means of Y for different values of the predictors. To proceed in this manner the data analyst (1) must know the predictors nature is using, (2) must know what transformations, if any, nature applies to those predictors, (3) must know that the predictors are linearly combined, and (4) has those predictors in the dataset to be analyzed. In short, for the first order condition to be met, the mean function specified in Equation 1 must be the mean function nature used to generate Y. The only unknowns are the values of the y-intercept and the regression coefficients.

Equation 2 is the disturbance function. A conventional linear regression model will be "second order correct" when the first order conditions are met and when the "errors" behave *exactly* as Equation 2 specifies. That is, the data analyst knows that each perturbation is realized independently of all other perturbations and that each is realized from a single distribution that has an expectation of 0.0. Because there is a single disturbance distribution, the variance of that distribution is said to be "constant." These are the usual second order conditions. Sometimes the disturbance is also assumed to be normal with variance σ^2. When N is much larger than p, the normality assumption is unnecessary.

Suppose that the first order conditions are met, and ordinary least squares is applied to the data. Estimates of the slopes and y-intercept are then unbiased estimates of the corresponding values that nature uses. When the first order conditions and the second order conditions are met, the disturbance variance can be estimated in an unbiased fashion using the residuals from the realized data. Conventional confidence intervals and statistical tests properly follow, and by the Gauss-Markov theorem, each estimated β has the smallest possible sampling variation of any other linear estimator of nature's regression parameters. A similar discussion applies to the entire generalized linear model and its multi-equation extensions, although that reasoning depends

[4]The expectation is essentially the mean Equation 1 over a limitless number of independent realizations of the data conditional on the x-values in the dataset. In the expectation, the values of regression coefficients are their means, and the value of the disturbance term is 0.0. The left hand side is then the means of Y for different values of predictors in the original dataset.

on asymptotics.[5]

There is nothing in the first or second order conditions about causal inference because causal inference is an interpretative overlay. It is not a formal feature of the regression model and depends conceptually on a potential outcomes perspective first proposed by Neyman (1927) and extended by Rubin (Rubin and Imbens, 2015). As Cook and Weisberg (1999:27) explain, the goal of a regression analysis is to understand "as far as possible with the available data how the conditional distribution of some response **y** varies across subpopulations determined by the possible values of the predictor or predictors." Cause is nowhere to be found. For example, one might compare for descriptive purposes the length of sentence given to 25 year old males, convicted of aggravated assault, with two prior felony convictions to 25 year old females, convicted of aggravated assault, with two prior felony convictions. Perhaps the males' distribution in the data has a larger mean and a longer tail to the right. There is no need for a causal interpretation and in any case, with observational data, causal inference can be very controversial (Friedman, 1987; 2004). In short, a regression model does not have to be a causal model.

3 Problems in Practice for Conventional Regression

In order to obtain unbiased estimates of the linear regression parameters, the first order conditions must be met; the mean function specified is the mean function used by nature. If these conditions are not met, any formal justification for estimation, confidence intervals, and statistical tests evaporates. In order to obtain valid statistical tests and confidence intervals, the first order conditions and the second order conditions must be met; the disturbances must be generated by nature as independent draws from a single distribution with a mean of 0.0.

[5]The term "asymptotics" in this context refers to the performance of regression estimates (e.g., the regression coefficients) when the number of observations increases without limit. Often this mathematical exercise shows that estimation biases decrease with larger sample sizes, and disappear with a limitless number of observations. Good asymptotic performance can be a fallback position for statistical procedures whose estimates are otherwise biased. Then, if the number of observations is far larger than the number of predictors, estimation biases are likely to be small.

One properly can proceed if the first order conditions are met even if the second order condition of constant disturbance variance is violated. Halbert White (1980b) provides valid, asymptotic standard errors when the disturbance variances are not constant (i.e., heteroscedasticity-consistent standard errors). Valid confidence intervals and statistical tests can follow.[6] However, sometimes these standard errors are characterized as "robust" which perhaps has led criminologists to use them when they do not apply. For example, they do not adjust properly for dependence between disturbances and most assuredly do not correct for mean function misspecification.

These sorts of details matter because it is usually impossible to know whether the regression model specified by the analyst is the means by which the data were generated. A common fallback, therefore, is to claim that the model specified is "close enough." But there is no way to know what "close enough" means. One requires the truth to quantify a model's disparities from the truth, and were the truth known, there would be no need to analyze any data.

Nevertheless, three strategies often are used to address the "close enough" requirement. First, sometimes researchers try to cover their bets by offering a suite of possible models. But, it is not clear what to make for this exercise. Perhaps most important, even if a single model is designated at the best, one cannot claim that the model chosen is properly specified. It may be just the best of a bad lot. Moreover, there are difficult conceptual and mathematical problems inherent in the concept of "best." For example, it does not follow that a better fitting model is closer to the correct model. One might be improving the fit by including predictors that are correlated with the response variable, but not actually a feature of the true model. It is also challenging to properly compare the different models in part because any statistical tests or confidence intervals are only correct for the single correct model, which is unknown. Even if that model happens to be among those examined, there is no way to determine which one it is.

Second, there are a large number of regression diagnostics taking a variety of forms including graphical procedures, statistical tests, and the comparative

[6]Other work by White (1980a) and others, to be addressed shortly, allows for asymptotically valid tests when the mean function is misspecified. But that work does not apply to the conventional linear regression model. By "valid" one means that the probabilities computed for statistical tests and confidence intervals, have the properties they are supposed to have. For example, the 95% confidence interval really does cover the value of the population parameter in 95% of possible realized datasets.

performance of alternative model specifications (Weisberg, 2005: Chapters 9-10). These tools can sometimes identify problems with the linear model. Most are designed to detect single difficulties in isolation when in practice, there can be many difficulties at once. For example, is evidence of non-constant variance a product of mean function misspecification, disturbances generated from different distributions, or both? In addition, diagnostic tools using statistical tests typically have weak statistical power (Freedman, 2009b: 193).

Compounding matters, when for model misspecification tests a null hypothesis is not rejected, analysts commonly "accept" the null hypothesis as if the model were correct (Goodman, 2016). In fact, there are effectively a limitless number of other null hypotheses that would also not be rejected. This is sometimes called "the fallacy of accepting the null" (Rozeboom, 1960).[7]

Finally, even if some model misspecification is accurately identified, there may be little guidance on how to fix it, especially within the limitation of the data available, and trying to re-specify the model can introduce new sources of bias. It is now well known that model selection and model estimation undertaken on the same data (e.g., statistical tests for a set of nested models) lead to biased estimates and/or incorrect statistical inference even if by some good fortune the correct model is found (Leeb and Pötscher, 2005; 2006; 2008; Berk et al., 2010; 2014).[8]

Third, when regression results make sense and are consistent with – or at least not contradicted by – existing theory and past research, some argue that the regression model must be reasonably close to right. Some go so far as to claim that earlier findings have been replicated, and that the model under consideration has been validated.

As a logical matter, these arguments about replicability do not parse. An obvious complication is that the study protocols must be comparable. If hot spots policing and community policing are both associated with crime

[7]For example, if the null hypothesis for a given regression coefficient is 0.0, there will almost always be many reasonable null values close to 0.0 that would also not be rejected. And even a coefficient value close to 0.0 can meaningfully change the model specification and the estimated values of the other regression coefficients. A predictor with a small regression coefficient may be strongly correlated with other predictors so that their estimated regression coefficients will vary substantially depending on whether that variable is included in the regression.

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

reductions, one would be hard pressed to claim reproducibility (Ioannidis, 2014; Harris, 2012; Open Science Collaboration, 2015). The same reasoning applies to studies using different regression models. And should the study protocols be comparable, one may well be reproducing results that are incorrect. Indeed, there is ample room within a claim of reproducibility to replicate nonsense. The model under scrutiny and the previous models to which comparisons are made may all be substantially wrong. Even many wrongs don't make a right.[9]

In summary, a close look at the requirements of conventional regression reveal a standard that is extremely difficult to meet. All statistical models are wrong (Box, 1976), not just because statistical models are by design simplifications, but because the formal requirements can be too strict for real world practice. So what is a researcher to do? In the pages ahead, we provide a more permissive formulation that comports better with how quantitative research on criminology is actually done.

4 A Statistical Formulation for Misspecified Regression Models

The conventional linear regression model requires that the data are realized *exactly* as described in Equations 1 and 2. A more permissive formulation allows each case to be realized independently from some joint probability distribution and does not require the first order and second order conditions essential for conventional linear regression.

4.1 A Finite Population Approach

One can get a grounded sense of what this means by thinking about a two-dimensional histogram. A more technical and complete discussion follows. As shown in Figure 1, there are two simulated variables X and Y that define

[9]These problems and more carry over to formal meta-analyses (Berk, 2007). For example, the set of studies being summarize are not a probability sample of anything and are not realized in an independent fashion. Indeed, one of the key features of the scientific enterprise is that later studies build on early studies. As a result, all statistical tests and confidence intervals are likely to be bogus. The one exception is when all of the studies are randomized experiments, but then the inferential formulation is somewhat different. Within that framework, one can have valid statistical inference.

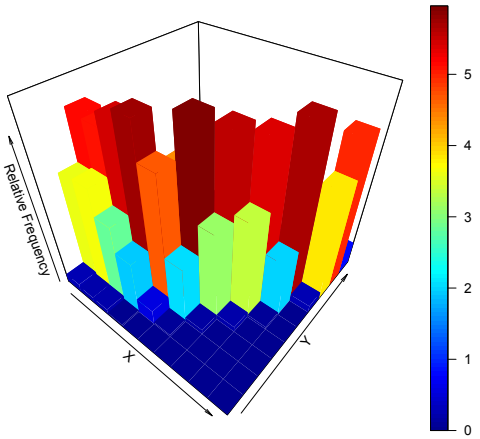

Figure 1: A Three-Dimensional Histogram of a Finite Population With Y And X As The Variables

a plane. Sitting on that plane are bars representing the relative frequencies of observations. The location of each bar is determined by a value of X and a value of Y (in this case, binned for visualization purposes). The proportion of cases contained within each bar can be approximately ascertained using the color legend on the far right.

Figure 1 is a visual summary of a joint distribution for the two variables X and Y. One can think of the data shown in Figure 1 as a finite population, as one might within a traditional random sampling framework. The population shown in Figure 1 has means for Y and X, variances for Y and X and a covariance between Y and X. These are the relevant moments of the joint population distribution. The variables Y and X are fixed; they do not change.

Suppose one had access to all of the population data shown in the histogram. Figure 2, is a birds eye view of Figure 1 and actually a scatter plot. Y is by construction a cubic function of X, although there are population residuals around the cubic function. The cubic mean function characterizes the population conditional means of Y for different values of X and consti-

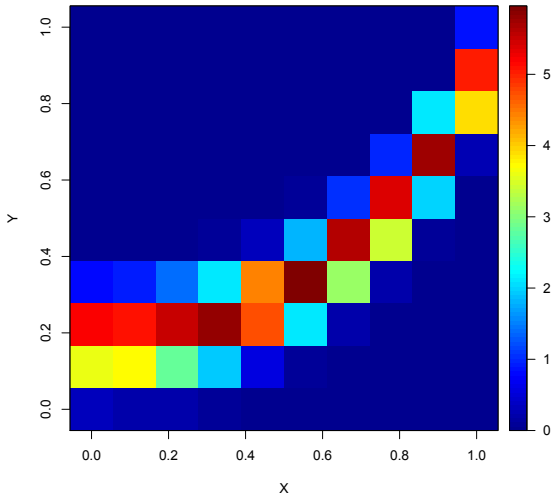

Figure 2: A Two-Dimensional Histogram Looking Down on Figure 1

tutes the "true response surface."[10]

Looking at Figure 2, a linear fit would less than ideal. Nevertheless, suppose a population linear regression of Y on X was computed by ordinary least squares. Clearly, the linear mean function is misspecified. What useful information might it convey?

Figure 3 provides some answers in a conventional scatter plot format. The blue circles are observations. (There is a lot of overprinting.) The green line is the true population response surface composed of the true conditional means. The red line is the best population linear approximation of those true conditional means. In this simple example, the linear function captures the positive monotonic association between X and Y. The slope represents the average change in Y for a unit change in X over the range of x-values in the population. Moreover, because the linear function is computed using ordinary least squares, one properly can claim that the population linear mean function is the best linear approximation of the true response surface.

Now imagine drawing a simple random sample from the population; the

[10]There is nothing special about the cubic function except its relative simplicity. We could have used here virtually any nonlinear function, but the price would have been a more difficult exposition.

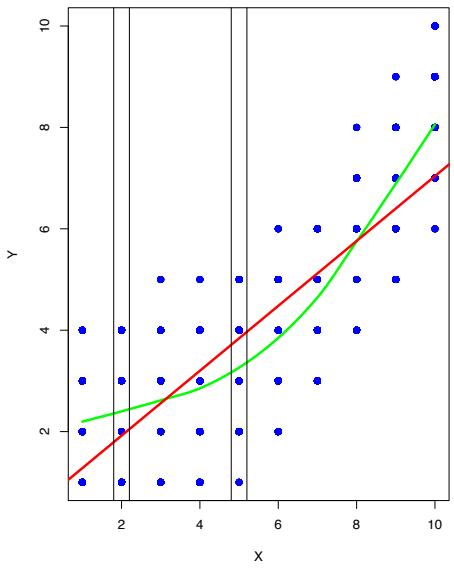

Figure 3: A Population Scatter Plot with True Response Surface In Green and Best Linear Approximation in Red

data are generated by real random sampling. Even though Y and X are fixed variables in the population, they are now random variables in the sample. Were a new random sample drawn, the sample values of both variables would differ by chance. The data does not result from nature appropriating the linear model. Allowing predictors to be random variables requires a fundamental *reformulation* of our estimation procedures.

We begin by abandoning the true response surface as the target of estimation. The true response function is assumed to be unknown and certainly not limited to a linear function. Adopting a prudent strategy, the data analyst wishes to *estimate* with ordinary least squares the population best linear approximation.[11] That is, the data analyst seeks an estimate of the red line in Figure 3. What are the properties of such an estimate, given random X and unknown true response surface that could well be nonlinear? To answer that question, we must leave behind the finite population and begin a more technical and abstract discussion.

4.2 Treating a Joint Probability Distribution as the Population

We begin with a statistical abstraction from a joint empirical distribution. There is now a population composed of variables $\mathbf{Z}$ in which the number of observations is limitless. The population is described by a joint probability distribution having usual sorts of parameters such as the mean and variance for each variable.[12] Because the number of observations is limitless, these parameters are expectations. For example, the mean for a particular Z is the expected value of that Z.

Within the joint probability distribution, there is no distinction between predictors and responses. For the population variables $\mathbf{Z}$, a *researcher* distinguishes between predictors $\mathbf{X}$ and responses $\mathbf{Y}$. Some of the variables in $\mathbf{Z}$ may be discarded because they are not relevant for the substantive or policy issues at hand. These practitioner decisions have nothing to do with how the data were generated.

[11] The framework to follow applies to any parametric approximation of the true response surface, not just a linear approximation. But working with a linear function makes the exposition much easier.

[12] There is a subtlety here. The variables in the joint probability distribution may well be correlated. But those correlations have no role in how the data are generated.

With the predictors and response determined, there is for these variables a true response surface that is a feature of the joint probability distribution. The true response surface is the set of conditional expectations of Y for the predictors X and can be highly nonlinear. No particular functional form is assumed and in practice, the functional form is unknown. Another feature of the joint probability distribution is a best linear approximation of that true response surface that is a least squares multiple regression of $Y|X$ in high dimensions of X.

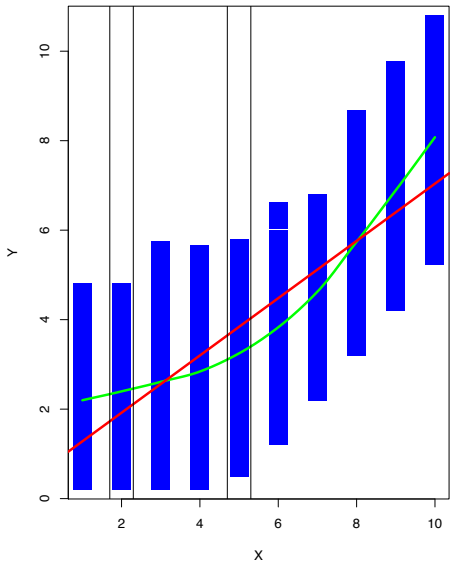

Figure 4: Nonconstant Variability Caused By Working Model Misspecification, A Nonlinear True Response Surface, and X Realized at Random

Figure 4 is much like Figure 3, but is meant to represent the joint probability distribution for Y and a one-dimensional X. The conditional distributions $Y|X$ are shown with solid blue bars. As before, the green line is the true response surface, and the red line is the best linear approximation.

Even if the variability around each true conditional expectation happens to be the same, the variability around the conditional expectations of the best linear approximation will likely differ. For the fifth vertical slice (boxed), the best linear approximation falls above the true response. Therefore, space between the two lines represents specification error. Because X as realized is

a random variable, the specification error is also random and gets folded into the disturbance variability. For the second vertical slice (boxed), the linear approximation falls below the true response surface. This is a specification error in the other direction but because X as realized is a random variable, it too is folded into the disturbance variability.

Across the entire range of X, specification error becomes part of the disturbance variability except when the true response surface and the best linear approximation have the same conditional expectation. Because the size of the specification error varies, so does the resulting variability around the best linear approximation. In short, the combination of a nonlinear true response surface and a best linear approximation, coupled with a randomly realized X, produces heteroscdasticity. This has estimation implications to be addressed shortly.

In practice, some finite number of observations are independently realized as the data to be analyzed. That is, the data are produced by a natural process equivalent to random sampling. Suppose now that a researcher analyzing such data takes as a working model a conventional linear regression. According to the working model, the conditional means over cases, $\boldsymbol{\mu}$, is assumed to be related to $\mathbf{X}$ by $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. $\mathbf{Y}$ is then $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Because the form of true response surface is unknown, there is no justification for treating the working model as correctly specified. But the researcher can treat the working model as a vehicle with which to estimate the best linear approximation of that unknown, true response surface. The researcher forgoes trying to estimate the true conditional means and settles for trying to estimate the best linear approximation of that truth. Very little is given up because as noted earlier, any working model will likely be misspecified if the true response surface is the estimation target.

Immediately there are important benefits. There is no longer any model misspecification because there is no such thing as omitted variables or incorrect functional forms. The estimation target is the best linear approximation specified by the researcher's working model, whatever that happens to be. Working models can be more or less informative, but they cannot be more of less incorrect. For a model to be more or less correct, a comparison to the true model is required. In addition, because there is no longer such a thing as model misspecification, there is no longer a need to examine regression diagnostics with the hope of patching up the mean function. Regression diagnostics can play a role, but only to improve estimates of the best linear approximation or perhaps suggest a different parametric approximation.

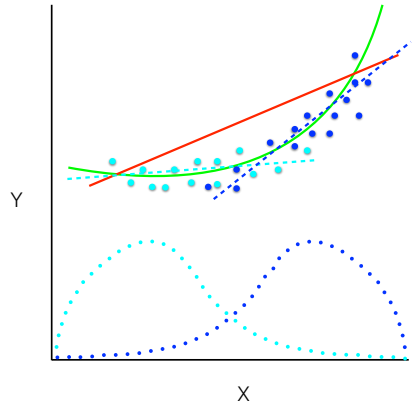

Figure 5: Estimation Complications Because of Random X And A Nonlinear True Response Surface

Unfortunately, estimation comes with complications. If as conventionally done, X is treated as fixed, an estimate of *the best linear approximation* in the population will be biased. Figure 5 shows why. The solid green line shows the true response surface. The solid red line is the best linear approximation in the population.

Suppose that in the sample, the distribution of X is skewed to the right. This is illustrated by the cyan distribution at the bottom of figure. It follows that low values of X will dominate the x-values in the sample. These are shown by the cyan-filled circles. The estimate of the population best linear approximation is shown by the cyan dashed line. Clearly, the slope is too small.

Suppose that in the sample, the distribution of X is skewed to the left. This is illustrated by the blue distribution at the bottom of the figure. It follows that high values of X will dominate the x-values in the sample. These are shown by the blue-filed circles. The estimate of the population best linear approximation is shown by the blue dashed line. Clearly, the slope is too large.

The technical point is that when, as conventionally done, X is treated as fixed, and there is mean function misspecification, the distribution of X in the sample matters even when the best linear approximation is the estimation target.[13] The practical point is that when, as conventionally done, X is

[13]Skewness is not essential. All one requires is that potential distributions of X have

treated as fixed, all of the usual estimation problems remain.

However, under our joint probability distribution formulation with observations independently realized, X is not fixed; X is a random variable. This means that over realizations of the dataset, one will get to see x-values from the full population distribution of X. Sometimes an estimated slope will be too flat, and sometimes an estimated slope will be too steep, just a shown in the figure. But over realizations of the dataset, the different slopes will be averaged and asymptotically, an estimate of the best linear approximation will be unbiased. In finite samples of even modest size, the bias will be small.[14]

In summary, the slope obtained from a given dataset can be interpreted as an asymptotically unbiased estimate of the the average slope over the full range of the unknown true response function. For any particular sample, the average may be too flat or too steep, and there is no way to know which or by how much. Nevertheless, in Figure 5 the estimate of the best linear approximation accurately conveys that by and large the true relationship is positive and monotonic.

The same reasoning can be applied when there is more than one predictor. The main difference is that each regression coefficient is, as usual, adjusted for its correlations with all other predictors; one has "partial" regression coefficients.

In addition, the best approximation can be the best nonlinear, parametric approximation. This allows for convenient mean functions such as polynomials. For example, in Figure 5, the approximation could be parabolic.

One might think that if the predictors are all categorical, there can be no nonlinear true response surface, and the problems addressed would disappear. This view is correct if for the true response surface, all of the categorical predictors are included additively. But if there are interaction effects as products of any predictors, and if those interaction effects are not included in the working mean function, one again has a nonlinear true response surface and a best linear approximation.

Finally, we come to proper estimates of the standard errors. The heteroscedasticy described earlier means that conventional standard errors for the estimated regression coefficients are not valid. It follows that the cor-

different expected values.

[14]The reliance on asymptotics is widespread in statistical and econometric applications. For example, even if the mean function for a logistic regression is correct, estimates of the regression coefficients are only unbiased asymptotically.

responding statistical tests and confidence intervals are also not valid. But there are two readily available solutions. First, one can apply a nonparametric bootstrap in which rows of the dataset are sampled at random with replacement. This produces asymptotically valid standard error leading to asymptotically valid statistical tests and confidence intervals (Freedman, 1981; McCarthy et al., 2016). Second, one can employ White's "sandwich estimator" to the same end (White, 1980a, Freedman, 1981). Both solutions can be accessed in popular statistical packages or easily coded in programming languages such as R.

4.3 Causal Inference

Working with misspecified regression models and observational data, even within our framework, presents significant challenges for causal inference. Perhaps most fundamentally, the working regression will probably not correspond to the real world setting to which causal inferences are to be drawn. Causal inferences are conventionally made from estimates of the true response surface, not from an explicit approximation whose correspondence to the truth is unknown.[15]

Still, with observational data, an estimate of the best linear approximation will usually be all one has to work from. Perhaps one can capitalize on the common practice in randomized experiments of estimating an average treatment effect (ATE).[16] After all, the best linear approximation is an average slope.

In randomized experiments, interventions are assigned, and the usual potential outcome framework is easily applied (Rubin and Imbens, 2015). For example, the standard model of treatment effects in randomized experiments allows each study unit to have its own pre-existing value for the response. Then for each unit, the response value is shifted up or down additively by some constant amount attributed to the intervention (Rosenbaum, 2002: Section 2.5.3). An ATE averages over these pre-existing differences to arrive at the additive constant.

[15]The same difficulties arise if regression is replaced by matching.

[16]For groups, an ATE formally is the difference between their two response variable means. Whether that difference can be *interpreted* as a causal effect depends on the the research design and in particular whether there is an intervention subject to manipulation. This requirements is met in randomized experiments and strong quasi-experiments. It can be very problematic in observational studies.

The best linear approximation is a different kind of average. The best linear approximation averages over slopes between pairs of observations that will vary in their distance from one another. Therefore, each pair of observations is being subjected to different interventions – the "change in X" will likely vary. And because the true response surface can be nonlinear, where the x-values are located matters too – the "change in Y" can differ for observations that are the same distance in X from one another. If one wishes to treat the slopes between pairs of observations as causal effects, the slope of the best linear approximation is a weighted average of causal effects.[17] For the conventional ATE, there is but one causal effect somewhat obscured by pre-existing differences between study units.

In short, the best linear approximation is not a conventional tool for causal inference. If conventional causal inference is a key study feature, one must try to estimate the true response surface. Randomized experiments or strong quasi-experiments are needed.

5 An Example

We now turn to an illustration of how the ideas discussed above can play out in a real application. The application is relatively simple. A richer application would require a relatively lengthy digression into substantive issues, which for this paper would be a diversion.

Variation in prison sentences has long been studied and can be a controversial policy issue. For example, the U.S. Sentencing Commission regularly publishes reports of federal sentencing outcomes by features of offenders, crimes and jurisdictions (http://www.ussc.gov). For illustrative purposes, we consider the sentences of 500 inmates incarcerated in a state prison system. They are a convenience sample from a recent year. The response variable is the nominal length of the prison sentence given by a sentencing judge.[18] We will address empirically the possible role of gender and other offender features in the lengths of sentences imposed (Steffensmeier et al., 1993; Ulmer and Bradley, 2006; Starr, 2015)

There are two ways to think about the population to which inferences are

[17] A thorough discussion the weights are can found in Buja and his colleagues (2016: Section 10). Perhaps the most important conclusion is that although the weights are formally required, they further complicate how an average causal effect is interpreted.

[18] Time actually served can differ, sometimes dramatically.

to be made. There is the set of all inmates in that prison system for several years around the time the data were collected. Although the inmates in the study are not a random sample, one can view them as random realizations from the social processes associated with prison sentences. There is no reasonable evidence that over that interval, there were important changes in the mix of inmates, relevant statutes, or courts' administrative practices. These inmates, therefore, would constitute a finite population of several hundred thousand that could be described by a joint empirical distribution. It is a relatively short conceptual step to imagine a limitless population of inmates that could have been realized from the same social processes over the time period of interest, described by a joint probability distribution. Because in this example, the finite population is so large relative to the sample size, either conception would in practice suffice.

We emphasize that such reasoning depends on subject-matter knowledge, and how well the reality corresponds to the formal statistical requirements will be a matter of degree. However, sometimes data from the population help. For example, if the prison system were able to provide for all current inmates key summary statistics (e.g., the current distribution of prison sentence lengths), comparisons could be made to the sample. Should such information be available over several years, more convincing comparisons could be made. In this instance, we actually have many summary statistics for the relevant population, and they correspond well to the summary statistics in the sample of 500.

Perhaps more demanding is the requirement that the 500 observations are realized independently. That too will be a matter of degree and would depend on such factors as whether earlier sentences given to convicted offenders significantly shape the sentences given to later convicted offenders in a state that has advisory sentencing guidelines. That is, given the guideline sentences, are the sentences realized independently?

In short, whether a dataset can be properly seen as a set of independently realized observations from a joint probably distribution needs to be justified on substantive grounds and will typically be a matter of degree. If the case cannot be made, statistical inference is off the table, and the analysis is limited to description of the data on hand.

Table 1 shows the regression results. For each predictor, the first three columns contain the usual ordinary least squares results. The last two columns show the "sandwich" standard errors and the proper t-values. Asterisks next to a t-value indicate that the p-value is less than .05 for a two

Predictor	Coefficient	Std. Err	t-Value	Sandwich	Proper t-Value
Intercept	0.45	1.99	0.22	1.84	0.22
Violent Record	4.64	0.51	9.13*	0.55	8.43*
Sex Offender	-1.15	1.17	-0.98	1.73	-0.66
Number of Prior Charges	0.01	0.02	0.67	0.02	0.67
First Arrest Age	-0.20	0.05	-4.23*	0.06	-3.33*
Number of Prior Arrests	-0.27	0.09	-3.02*	0.09	-3.00*
Gender	2.63	0.75	3.52*	0.50	5.26*
IQ	-0.01	0.02	-0.61	0.02	-0.61
Age	0.26	0.03	7.99*	0.05	5.20*

Table 1: Regression Results for Nominal Prison Sentence with Proper "Sandwich" Standard Errors (N=500)

tailed test.

We have specified on purpose a model whose mean function is clearly incorrect. For example, we do not include the crimes for which the offender was convicted despite requirements of the sentencing guidelines. The variable "Violent Record" only indicates whether the conviction offense and/or other prior convictions were for violent crimes. There are also reasons to believe that some nonlinear relationships have been overlooked. For example, age likely has a nonlinear relationship with sentence length. The working regression provides estimates of the population best linear approximation of the true response surface.

Consider first how one should interpret the results for a conventional linear regression. One literally has in the regression coefficients estimates of the constants nature used when constructing the linear combination of predictors responsible for average sentence length. One can generalize the results to all offenders and settings in which nature proceeded in very same way.

For five of the eight predictors, the conventional least squares regression leads to a rejection of the usual null hypothesis of no linear association. If one takes the tests at face value, one still has the linear machinery nature used, but with some predictors that nature did not in fact employ. There might be a very strong temptation to re-estimate of regression coefficients for a specification that did not include the predictors whose null hypothesis of 0.0 could not be rejected. But if the same data were used, the new coefficients will be estimated in a biased manner, and all subsequent statistical tests

will be invalidated. This point was made earlier when "model selection" was briefly addressed.

Of particular interest is that holding all else in the linear regression equation constant, being male is associated with an average increase of 2.63 years in sentence length. In conventional terms, this is taken to be an unbiased estimate of the true relationship between gender and sentence length, holding *all possible* confounders constant. Moreover, the increment of 2.63 years can be an estimate of the average treatment effect (ATE) of gender. If one changed a convicted offender's gender from female to male, the sentence given would be on the average 2.63 years longer.[19] The signs and magnitudes of the other "significant" coefficients are consistent with expectations except for the number of prior arrests, which has a negative association with sentence length. Interpretations the other regression coefficients would take much the same form as the interpretation for gender.

Consider now the results from the perspective of a best linear approximation. Some of the sandwich standard errors differ substantially from the conventional standard errors. In particular, the sandwich standard error for gender is 0.50, and the conventional standard error is 0.75. Because the valid standard error is about a third smaller, the 95% confidence interval around the gender regression coefficient is about a third smaller as well. The gender t-value using the sandwich standard error is nearly 50% larger, but in either case, the null hypothesis is easily rejected at the .05 level.

Getting the proper standard errors is largely a technical matter. More challenging is how to interpret properly the regression coefficients from the linear approximation. The regression coefficient for gender is again a good illustration. The longer average sentence for men of 2.63 years represents an association from a linear approximation of the unknown, true relationship. It is not an unbiased estimate of the true relationship, but an asymptotically unbiased estimate from a linear approximation of the true relationship. For the true relationship, there might be no association between sentence length and gender, or it might be that women on the average receive longer sentences. Moreover, one has only an association, not an estimated causal effect.[20]

[19]There are well-known interpretative problems treating gender as a cause because it is not manipulable, but that is often overlooked when causal interpretations are provided for regression results. Causal interpretations for race have the same problem (Berk, 2003: Chapter 5).

[20]Language matters too. One must be careful about using verbs like "affect," "impact,"

Nevertheless, if one were concerned about gender bias in sentencing, there is evidence that holding constant the number of prior arrests, the age at first arrest, the number of prior charges, and several other predictors thought to be related to sentence length, men on the average receive substantially longer sentences. One has results consistent with gender discrimination even if evidence is a this point not very compelling. The weak results apply to all offenders whose sentences are subject to the very same criminal justice processes.

Likewise, offenders with a violent criminal history have sentences that are on the average 4.64 year longer. For each additional year of age at which an offender's first arrest as an adult occurred, sentence length is on the average about .20 years shorter. A first arrest at 15 compared to a first arrest at 20 is associated with an average sentence that is about a year shorter. Such associations are being estimated in a nearly unbiased manner for a sample of 500, but they are estimates for the linear approximation, not the true response surface.

In short, the regression coefficients have much in common with partial correlation coefficients.[21] Each is a measure of association adjusted for correlations with the other predictors included in the working regression model. Because the original units of the response and the predictors are retained, the size of the association can be given a grounded interpretation.

A few regression diagnostics were examined. Perhaps most important were the variance inflation factors associated with each predictor. One might wonder if dependence between predictors was diluting estimation precision. The variance inflation factors were all relatively small. Most of the variances for the estimated regression coefficients were less than twice the size they would have been had all of the predictors been uncorrelated with one another. For these kinds of data, that is a good result.

Also examined were simple transformations of several predictors to consider nonlinear relationships. For example, the variable "Age" was replaced the the square of "Age." None of these transformations changed the results in important ways. To confirm these conclusions, the working model was re-estimated within a generalized additive formulation. All numerical pre-

or "influence," which can be read as implying causality.

[21]The partial correlation is not used much any more despite have an impressive pedigree (Fisher, 1924). It is just the usual Pearson correlation, but between two variables from which any linear dependence with other specified variables has been removed, much as in multiple regression.

dictors were smoothed. The quality of the fit improved a bit, but the overall results were much the same. There were for these predictors apparently no strong nonlinear relationships overlooked.[22]

6 Discussion

For the practitioner, the computational changes associated with our best approximation approach are easily made. One can run the usual regression software and then compute sandwich or nonparametric bootstrap standard errors. Proper statistical tests and confidence intervals can then follow as usual.

A far more challenging matter is how to think about the underlying assumptions. There are no longer the first or second order conditions required by conventional regression. According to those rules, the working regression model is misspecified. The only requirement is that the data are generated as independent realizations from a substantively appropriate population. The real world must have provided the data by the equivalent of random sampling. Such a claim will rest on substantive considerations and will typically be a matter of degree.

There is still a model of the data generation process, but one that we have called "assumption lean" (Buja et al., 2016). Conventional regression requires a very similar conception for the regression disturbances but *in addition*, requires that the mean function specified is the mean function used by nature. We have called the conventional regression formulation "assumption-laden" (Buja et al., 2016).

Some readers may long for a regression approach that is totally model free. If one is satisfied using regression solely to describe interesting features of the data on hand, there is no need for a generative model accounting for how the data can to be. And rich description is surely a worthy scientific and policy goal. But if one wishes to draw inferences beyond the data on hand, there must be a good answer to the question: inferences to what? Without a credible answer, estimates from the data are a statistical bridge to nowhere.

[22]One might wonder why the generalized additive model was not used instead of linear regression. The generalized additive model is an inductive procedure that adapts empirically to the data through a tuning parameter. This constitutes model selection that introduces significant complications for all statistical inference (Berk 2016, Chapter 2). A discussion of these issues is well beyond the scope of this paper.

Moreover, there must be a good answer to a second question: how close to probability sampling are the means by which the data were generated? Unless a credible case can be made that the correspondence is reasonably close, there is no way to build a statistical bridge to begin with.

In practice, answers to both questions are necessarily derived from subject-matter knowledge and will be matters of degree. There is no room for "assume and proceed" statistics nor for "point-and-click" statistical analyses. Technical expertise must be combined with substantive judgement.

7 Implications for Practice

The "wrong model" perspective has important implications for practice. These have been introduced earlier paper. We now provide them summary form.

1. We have given criminologists a formal rationale for the common practice of not taking specified models literally. It can be far more sensible to explicitly and correctly make use of misspecified regression models than to proceed as if misspecified models can be properly interpreted as if they were specified correctly. Our approach is internally consistent and honest. The conventional approach is neither.

2. Under conventional regression formulations, data are generated by nature using a linear expression with several additional assumptions. Under the wrong model perspective, the data are generated independently and randomly from a joint probability distribution. Neither formulation is required if the goal is description of the data on hand. But if inferences are to be drawn beyond the data, those inferences have to be drawn to something. In sample surveys, inferences are typically made to a well defined, finite population. We provide a mathematical abstraction of that basic idea. Our approach is "assumption lean." The conventional approach is "assumption laden."

3. For conventional regression, the estimation target is the function by which nature actually generated the data – the "true model." For our approach, the estimation target is an acknowledge parametric approximation of the true model. The approximation is "best" when it is the product of ordinary least squares.

4. With a wrong regression model specified, one can proceed as usual with one's software of choice to obtain estimates of regression coefficients. Regression coefficients retain their usual descriptive interpretation: how much the mean of the response differs depending on a one unit of variation in a given predictor with the linear dependence between that predictor on all other predictors removed (i.e., with all other predictors "held constant"). When the estimation target is the truth, regression coefficient estimates will almost certainly be biased, even asymptotically. When the estimation target is the approximation, regression coefficient estimates will be asymptotically unbiased. If the number of observations is substantially larger than the number of predictors, the biases in a given sample will be small.

5. If the estimation target is the true model, estimated standard errors will be biased, even asymptotically. It follows that statistical tests and confidence intervals will not perform as they should and any inferential conclusions could be seriously in error. One may be rejecting a null hypothesis when one should not, or one may be failing to reject a null hypothesis when one should. Confidence intervals will not have their advertised coverage. If the estimation target is the approximation, one cannot use the usual standard error estimates routinely provided by popular software. One needs to employ either the nonparametric bootstrap or the "sandwich" estimator. Both are readily available in standard regression packages. Then, standard errors, statistical tests, and confidence intervals will be asymptotically correct.

6. With conventional regression, causal inference is often a central goal. Causal inferences can be very misleading with a misspecified regression model. Under the wrong model approach, causal inference is not an option. Causal *interpretations* may be useful, but one does not have *estimates* of causal effects. For example, one can choose to *interpret* an offender's prior record as a cause of sentence length, but not take the value of the associated regression coefficients as an estimate of its causal effect. One might say that a regression coefficient in the expected direction is consistent with a causal impact, but not say how much the expected sentence length would change if the number of prior convictions was altered to be one more or one less. One is working with a regression summary statistic much like a partial correlation coefficient,

but not in standardized units.

7. Under either the right model or wrong model formulation, casual inference is almost certainly problematic. A far better approach, when practical, is to implement a randomized experiment or a strong quasi-experiment. Sometimes instructive natural experiments are available.[23]

8. Working within the wrong model perspective means that model misspecification is not longer relevant. Some working models will be more instructive, complete or interesting than others, but they are all treated as wrong. Regression diagnostics can help researchers find better misspecified models, but not a model that is demonstrably correctly specified.[24]

These implications for practice go more to how one thinks about regression analysis than to its mechanics. Required is a foundational attitude adjustment. We are not advocating another technical elaboration on top of usual practice.

8 Summary and Conclusions

Telling criticisms of linear regression are old news, and there has yet to be an effective rebuttal. At least implicitly, many researchers seem to understand the situation. They will readily acknowledge that their working models are only approximations of the true relationships. However, they still proceed with all of the formal trappings of conventional regression that by and large no longer apply. This can lead to all manner of unnecessary labor, incorrect statistical inference, and misleading interpretations of results.

In this paper, we provide a more permissive approach allowing one properly to work with misspecified regression models. But the newfound freedom comes at a price. One must acknowledge that the estimation target is an approximation of the truth from which causal inference are very difficult to justify. Causal interpretations of the estimated associations can be in play,

[23] In a natural experiment, nature provides a good approximation of a randomized experiment or quasi-experiment.

[24] There are a number of subtle issues when using of regression diagnostics with explicitly misspecified models that are beyond the scope of this paper. But generally, visual and graphical tools can be properly employed. Formal tests are likely to be problematic.

but the estimates are *not* conventional ATEs. The coefficients do not convey what will happen if a given predictor is manipulated.

Some readers will argue that the price is too high. But in fact, there is rarely any price to be paid. It is very difficult to find regression models in criminology, or in the social sciences more generally, for which a strong case for proper specification can be made (Berk, 2003; Angrist and Pischke, 2008; Freedman, 2009). Misspecified models are ubiquitous. If credible estimates of causal effects are an essential feature of an analysis, the best option is to undertake a randomized experiment or a very strong quasi-experiment.

References

Angrist, J., and S. Pischke (2008) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Berk, R.A. (2003) *Regression Analysis: A Constructive Critique*. Newbury Park, CA.: Sage.

Berk, R.A., (2007) "Meta-Analysis and Statistical Inference" (with commentary), *Journal of Experimental Criminology*, 3(3): 247- 297,

Berk, R.A., (2009) "The Role of Race in Forecasts of Violent Crime." *Race and Social Problems* 1: 231–242.

Berk, R.A. (2016) *Statistical Learning from a Regression Perspective*, second edition. New York: Springer.

Berk, R.A., Baek, A., Ladd, H., and H. Graziano (2002) "A Randomized Experiment Testing Inmate Classification System." *Criminology & Public Policy* 2 (1): 239–256.

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

Berk., R.A., Brown, L., Buja, A., Zhang, K., and L. Zhao (2014a) "Valid Post-Selection Inference." *Annals of Statistics* 41(2).

Berk., R.A., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., and L. Zhao (2014b) "Misspecified Mean Function Regression: Making Good Use of Regression Models that are Wrong." *Sociological Methods and Research* 43: 422-451, 2014.

Box, G.E.P. (1976) "Science and Statistics." *Journal of the American Statistical Association* 71(356): 791–799.

Buja, A., Berk, R.A., Brown, L., George, E., Pitkin, E., Traskin, M., Zhao, L., and K. Zhang (2016) "Models as Approximations — A Conspiracy of Random Regressors and Model Violations Against Classical Inference in Regression." $imsart-sts\ ver.2015/07/30 : Buja\_et\_al\_Conspiracy-v2.texdate : July\ 23, 2015.$

Bushway, S., and A. Morrison Piehl. (2001) "Judging Judicial Discretion: Legal Factors and Racial Discrimination in Sentencing." *Law and Society Review* 35(4): 733–764.

Cook, D.R. and S. Weisberg (1999) *Applied Regression Including Computing and Graphics*. New York: John Wiley and Sons.

Fisher, R.A. (1924) "The Distribution of the Partial correlation Coefficient." *Metron* 3: 329-332.

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

Freedman, D.A. (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 for Causation and the Identification Problem." *Evaluation Review* 28: 267–293.

Freedman, D.A. (2009) *Statistical Models* Cambridge, UK: Cambridge University Press.

Goodman,, S.N. (2016) "Aligning Statistical and Scientific Reasoning," *Science* 352(6290): 1180–1181.

Harris, C.R. (2012) "Is The Replicability Crisis Overblown? Three Arguments Examined." *Perspectives on Psychological Science* 7(6): 531–536.

Ioannidis, J.P.A. (2012) "Why Science Is Not necessarily Self-Correcting." *Perspectives on Psychological Science* 7(6): 645–654.

Leamer, E.E. (1978) *Specification Searches: Ad Hoc Inference with Non-Experimental Data*. New York, John Wiley.

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

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

Leeb, H., B.M. Pötscher (2008) "Model Selection," in T.G. Anderson, R.A. Davis, J.-P. Kreib, and T. Mikosch (eds.), *The Handbook of Financial Time Series*, New York, Springer: 785–821.

McCarthy, D., Zhang, K., Berk, R.A., Brown, L.,Buja, A., George, E., and L. Zhao (2016) "Calibrated Percentile Double Bootstrap for Robust Linear Regression Inference." Working Paper. Department of Statistics, University of Pennsylvania.

Neyman, J. (1923) "On the Application of Probability Theory to Agricultural Experiments: Essays on Principles. Section 9. *Roczniki Nauk Rolniczch Tom X* [in Polish]; translated in *Statistical Science* 5: 588-606, 1990.

Open Science Collaboration (2015) "Estimating The Reproducibility of Psychological Science." *Science* 346(6251): 943.

Rosenbaum, P.R. (2002) *Observational Studies*. New York: Springer.

Rozeboom, W.W. (1960) "The Fallacy of the Null Hypothesis Significance Test. *Psychological Bulletin* 57: 416–428.

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." *Annals of Applied Statistics* 2(3): 808–840.

Rubin, D.B., and G.W. Imbens (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge University Press.

Starr, S.B., (2015) "Estimating Gender Disparities in Federal Criminal Cases." *American Law & Economics Review* 17(1): 127–159.

Steffensmeier, D., Kramer, J.,and C. Streifel (1993) "Gender and Imprisonment Decisions." *Criminology* 31:411-46.

Ulmer, J,T, and M.S. Bradley (2006) "Variation in Trial Penalties Among Serious Violent Offenders." *Criminology* 44: 631–670.

Weisberg, S. (2013) *Applied Linear Regression*, fourth edition. New York: Wiley.

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

White, H. (1980b) "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4): 817–838.