Assumption Lean Regression

Richard Berk, Andreas Buja, Lawrence Brown, Edward George
Arun Kumar Kuchibhotla, Weijie Su, and Linda Zhao
University of Pennsylvania

November 26, 2018

Abstract

It is well known that models used in conventional regression analyses are commonly misspecified. Yet in practice, one tends to proceed with interpretations and inferences that rely on correct specification. Even those who invoke Box's maxim that all models are wrong proceed as if results were generally useful. Misspecification, however, has implications that affect practice. Regression models are approximations to a true response surface and should be treated as such. Accordingly, regression parameters should be interpreted as statistical functionals. Importantly, the regressor distribution affects targets of estimation and regressor randomness affects the sampling variability of estimates. As a consequence, inference should be based on sandwich estimators or the pairs (x-y) bootstrap. Traditional prediction intervals lose their pointwise coverage guarantees, but empirically calibrated intervals can be justified for future populations. We illustrate the key concepts with an empirical application.

1 Introduction

It is old news that models are approximations and that regression analyses of real data commonly employ models that are misspecified in various ways. Conventional approaches are laden with assumptions that are questionable, many of which are effectively untestable (Box, 1976, Leamer, 1878; Rubin, 1986; Cox, 1995; Berk, 2003; Freedman, 2004; 2009). This note discusses some implications of an "assumption lean" reinterpretation of regression. In this reinterpretation, one requires only that the observations are iid, realized at random according to a joint probability distribution of the regressor and response variables. If no model assumptions are made, the parameters of fitted models need to be interpreted as statistical functionals, here called "regression functionals."

For ease and clarity of exposition, we begin with linear regression. Later, we turn to other types of regression and show how the lessons from linear regression carry forward to the generalized linear model and even more broadly. We draw heavily on two papers by Buja et al. (2018a;b), a portion of which draws on early insights of Halbert White (1980).

2 The Parent Joint Probability Distribution

For observational data, suppose there is a set of real-valued random variables that have a joint distribution $\boldsymbol{P}$, also called the "population," that characterizes regressor variables $X_1, \ldots, X_p$ and a response variable Y. The distinction between regressors and the response is determined by the data analyst based on subject matter interest. These designations do not imply any causal mechanisms and or any particular generative models for $\boldsymbol{P}$. Unlike textbook theories of regression, the regressor variables are not interpreted as fixed; they are as random as the response and will be treated as such.

We collect the regressor variables in a $(p{+}1)\times 1$ column random vector $\vec{\boldsymbol{X}} = (1, X_1 \ldots, X_p)'$ with a leading 1 to accommodate an intercept in linear models. We write $\boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ for the joint probability distribution, $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ for the conditional distribution of Y given $\vec{\boldsymbol{X}}$, and $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ for the marginal distribution of $\vec{\boldsymbol{X}}$. The only assumption we make is that the data are realized iid from $\boldsymbol{P}$. The separation of the random variables into regressors and a response implies interest in $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$. Hence, some form of regression analysis is applied. Yet, the regressors being random variables, their marginal distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ cannot be ignored for reasons to be explained below.

3 Estimation Targets

As a feature of $\boldsymbol{P}$ or, more precisely, of $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$, there is a "true response surface" denoted by $\mu(\vec{\boldsymbol{X}})$. Most often, $\mu(\vec{\boldsymbol{X}})$ is the conditional expectation of Y given $\vec{\boldsymbol{X}}$, $\mu(\vec{\boldsymbol{X}}) = \boldsymbol{E}[Y|\vec{\boldsymbol{X}}]$, but there are other possibilities, depending on the context. For example, $\mu(\vec{\boldsymbol{X}})$ might be chosen to be the conditional median or some other conditional quantile of Y given $\vec{\boldsymbol{X}}$. The true response surface is a common estimation target for conventional regression in which a data analyst assumes a specific parametric form. We will *not* proceed in this manner and will not make assumptions about what form $\boldsymbol{P}_{Y|\vec{\boldsymbol{X}}}$ actually takes. Yet, we will make use, for example, of standard ordinary least squares (OLS) fitting of linear equations. We choose OLS for illustrative purposes and for the simplicity of the insights gained, but in later sections, we will consider Poisson regression as an example of GLMs. Using OLS despite a lack of trust in the underlying linear model reflects ambiguities in many data analytic situations; deviations from linearity in $\mu(\vec{\boldsymbol{X}})$ may be difficult to detect with diagnostics, or the linear fit is known to be a deficient approximation of $\mu(\vec{\boldsymbol{X}})$ and yet, OLS is employed because of substantive theories, measurement scales, or considerations of interpretability.

Fitting a linear function $l(\vec{\boldsymbol{X}}) = \boldsymbol{\beta}'\vec{\boldsymbol{X}}$ to Y with OLS can be represented mathematically at the population $\boldsymbol{P}$ without assuming that the response surface $\mu(\vec{\boldsymbol{X}})$ is linear in $\vec{\boldsymbol{X}}$:

$$\boldsymbol{\beta}(\boldsymbol{P}) \;=\; \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\; \boldsymbol{E}\big[\,(Y - \boldsymbol{\beta}'\vec{\boldsymbol{X}})^2\,\big]. \tag{1}$$

The vector $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ is the "population OLS solution" and contains the "population coefficients." Notationally, when we write $\boldsymbol{\beta}$, it is understood to be $\boldsymbol{\beta}(\boldsymbol{P})$. Similar to finite datasets, the OLS solution for the population can be obtained by solving a population version of the normal equations, resulting in

$$\boldsymbol{\beta}(\boldsymbol{P}) \;=\; \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}Y]. \tag{2}$$

Thus, one obtains the best linear approximation to Y as well as to $\mu(\vec{\boldsymbol{X}})$ in the OLS sense. As such, it can be useful without (unrealistically) assuming that $\mu(\vec{\boldsymbol{X}})$ is identical to $\boldsymbol{\beta}'\vec{\boldsymbol{X}}$.

We have worked so far with a distribution/population $\boldsymbol{P}$, not data. We have, therefore, defined a target of estimation: $\boldsymbol{\beta}(\boldsymbol{P})$ obtained from (1) and (2) is the estimand of empirical OLS estimates $\hat{\boldsymbol{\beta}}$ obtained from data. This estimand is well-defined as long as the joint distribution $\boldsymbol{P}$ has second moments and the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ is not perfectly collinear; that is, the second moment matrix $\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']$ is full rank. There is no need to assume linearity of $\mu(\vec{\boldsymbol{X}})$ homoskedasticity or Gaussianity. This constitutes the "assumption lean" or "model robust" framework.

An important question is why one should settle for the best linear *approximation* to the truth? Indeed, those who insist that models must always be "correctly specified" are likely to be unreceptive. They will revise models until diagnostics and goodness of fit tests no longer detect deficiencies so the models can be legitimately treated as correct.

Such thinking warrants careful scrutiny. Data analysis with a fixed sample size requires decisions about how to balance the desire for good models against the costs of data dredging. "Improving" models by searching regressors, trying out transformations of all variables, inventing new regressors from existing ones, using model selection algorithms, performing interactive experiments, applying goodness of fit tests and diagnostic plots can each invalidate subsequent statistical inference. The result often is models that not only fit the data well, but fit them too well (Hong et al. 2017).

Research is underway to provide valid post-selection inference (e.g., Berk et al. 2013, Lee et al. 2016), which is an important special case. The proposed procedures address solely regressor selection, and their initial justifications make strong Gaussian assumptions. Recent developments, however, indicate that extensions of Berk et al. (2013) have asymptotic justifications under misspecification (Bachoc et al. 2016, Kuchibhotla et al. 2018).

Beyond the costs of data dredging, there can be substantive reasons for discouraging "model improvement." Some variables may express phenomena in "natural" or "conventional" units that should not be transformed even if model fit is improved. A substantive theory may require a particular model that does not fit the data well. Identifying important variables may be the primary concern, making quality of the fit less important. Predictors prescribed by subject-matter theory or past research may be unavailable so that the model is the best that can be done. In short, one must consider ways in which valid statistical inference can be undertaken with models acknowledged to be approximations.

We are ***not*** making an argument for discarding model diagnostics. It is always important to learn all that is possible from the data, including model deficiencies. In fact, in

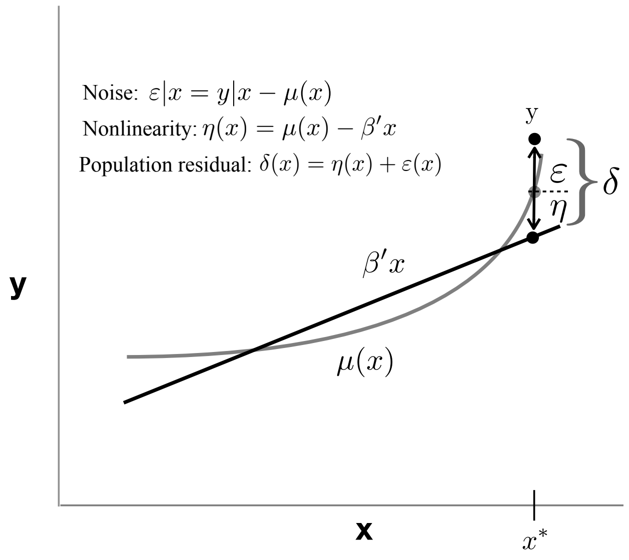
Noise: $\varepsilon|x = y|x - \mu(x)$
Nonlinearity: $\eta(x) = \mu(x) - \beta' x$
Population residual: $\delta(x) = \eta(x) + \varepsilon(x)$
y
ε
δ
η
$\beta' x$
y
$\mu(x)$
x
x^*

Figure 1: A Population Decomposition of $Y|X$ Using the Best Linear Approximation

Buja et al. (2018b) we propose a reweighting diagnostic that are tailored to the regression quantities of interest.

We also are not simply restating Box's maxim that models are always "wrong" in some ways but can be useful despite their deficiencies. Acknowledging models as approximations is one thing. Understanding the consequences is another. What follows, therefore, is a discussion of some of these consequences and an argument in favor of assumption lean inference employing model robust standard errors, such as those obtained from sandwich estimators or the x-y bootstrap.

4 A Population Decomposition of the Conditional Distribution of Y for OLS Fitting

A first step in understanding the statistical properties of the best linear approximation is to consider carefully the potential disparities in the population between $\mu(\vec{\boldsymbol{X}})$ and $\boldsymbol{\beta}'\vec{\boldsymbol{X}}$. Figure 1 provides a visual representation. There is for the moment a response variable Y and a single regressor X. The curved line shows the true response surface $\mu(x)$. The straight line shows the best linear approximation $\beta_0 + \beta_1 x$. Both are features of the joint probability distribution, not a realized dataset.

The figure shows a regressor value x^* drawn from $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ and a response value y drawn from $\boldsymbol{P}_{Y|X=x^*}$. The disparity between y and the fitted value from the best linear approximation is denoted as $\delta = y - (\beta_0 + \beta_1 x^*)$ and will be called the "population residual." The value of δ at x^* can be decomposed into two components:

- The first component results from the disparity between the true response surface, $\mu(x^*)$, and the approximation $\beta_0 + \beta_1 x^*$. We denote this disparity by $\eta = \eta(x^*)$ and call it "the nonlinearity." Because $\beta_0 + \beta_1 x^*$ is an approximation, disparities should be expected. They are the result of mean function misspecification. As a function of the random variable X, the nonlinearity $\eta(X)$ is a random variable as well.
- The second component of δ at x^*, denoted by ε, is random variation around the true conditional mean $\mu(x^*)$. We prefer for such variation the term "noise" over "error." Sometimes it is called "irreducible variation" because it exists even if the true response surface is known.

The components defined here and shown in Figure 1 generalize to regression with arbitrary numbers of regressors, in which case we write $\delta = Y - \boldsymbol{\beta}'\vec{\boldsymbol{X}}$, $\eta = \mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}'\vec{\boldsymbol{X}}$ and $\varepsilon = Y - \mu(\vec{\boldsymbol{X}})$. These random variables should *not be confused with error terms* in the sense of generative models. They share some properties with error terms, but these are *not assumptions*, rather, they are consequences of the definitions that constitute the above OLS-based decompositions. Foremost among properties is that the population residual, the nonlinearity and the noise are all "population-orthogonal" to the regressors:

$$\boldsymbol{E}(X_j\,\delta) = \boldsymbol{E}(X_j\,\eta(\vec{\boldsymbol{X}})) = \boldsymbol{E}(X_j\,\varepsilon) = 0. \tag{3}$$

As was already noted, these properties (3) are ***not*** assumptions. They derive directly from the decomposition described above and the fact that $\boldsymbol{\beta}'\vec{\boldsymbol{X}}$ is the population OLS approximation of Y and also of $\mu(\vec{\boldsymbol{X}})$. This much holds in an assumption lean framework without making any modeling assumptions whatsoever.

Because we assume an intercept to be part of the regressors ($X_0 = 1$), the facts (3) imply that all three terms are marginally population centered:

$$\boldsymbol{E}[\delta] \;=\; \boldsymbol{E}[\eta(\vec{\boldsymbol{X}})] \;=\; \boldsymbol{E}[\varepsilon] = 0. \tag{4}$$

However, δ is ***not*** conditionally centered and ***not*** independent of $\vec{\boldsymbol{X}}$ as would be the case assuming a conventional error term in a linear model. We have instead $\boldsymbol{E}[\delta|\vec{\boldsymbol{X}}] = \eta(\vec{\boldsymbol{X}})$, which, though marginally centered, is a function of $\vec{\boldsymbol{X}}$ and hence, not independent of the regressors (unless it vanishes). By comparison, the noise ε is marginally and conditionally centered, $\boldsymbol{E}[\varepsilon|\vec{\boldsymbol{X}}] = 0$, but not assumed homoskedastic, and hence, not independent of $\vec{\boldsymbol{X}}$.

We emphasize that in contrast to standard practice, the regressor variables have been treated as random and not as fixed. The assumption lean framework has allowed a constructive decomposition that mimics some of the features of a linear model but replaces the usual assumptions made about "error terms" with orthogonality properties associated with the random regressors. These properties are satisfied by the population residuals, the nonlinearity and the noise alike. They are not assumptions. They are consequences of the decomposition.

5 Regressor Distributions Interacting With Misspecification

Because in reality regressors are most often random variables that are as random as the response, it is a peculiarity of common statistical practice that such regressors are treated as fixed (Searle, 1970: Chapter 3). In probabilistic terms, this means that one conditions on the observed regressors. Under the frequentist paradigm, alternative datasets generated from the same model leave regressor values unchanged; only the response values change. Consequently, regression models have nothing to say about the regressor distribution; they only model the conditional distribution of the response given the regressors. This alone might be seen by some as sufficient to justify conditioning on the regressors. There exists, however, a more formal justification. Drawing on principles of mathematical statistics, in any regression model regressors are ancillary for the parameters of the model, and hence, can be conditioned on and treated as fixed. This principle, however, has no validity here because it applies only when the model is correct, which is precisely the assumption discarded by an assumption lean framework. Thus, we are not constrained by statistical principles that apply only in a model trusting framework.

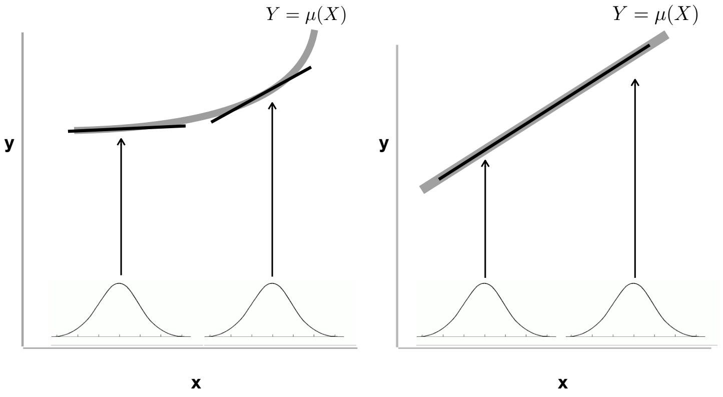

Figure 2: Dependence of the Population Best Linear Approximation on the Marginal Distribution of the Regressors

Ignoring the marginal distributions of the regressor is perilous under misspecification, and Figure 2 shows why. The left and right side pictures both compare the effects of different regressor distributions for a single regressor variable X in two different population settings. The left plot shows misspecification for which the true mean function $\mu(X)$ is nonlinear. Yet a linear function is fitted. The best linear approximation to the nonlinear mean function depends on the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. Therefore, the "true parameters" $\boldsymbol{\beta}$ — the slope and intercept of the best fitting line at the population — will also depend on the regressor distribution. One can see that for the left marginal distribution that the intercept is larger and the slope is smaller than for the right marginal distribution. This implies that under misspecification the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$, thought of as a "non-parametric nuisance parameter," is no longer ancillary.

The right side plot of Figure 2 shows a case of correct specification: The true mean function $\mu(X)$ (gray line) is linear. Consequently, the best linear approximation clearly is the same (black line) for both regressor distributions. In this case, the population marginal distribution of X does not matter for the best linear approximation. There is one value for $\boldsymbol{\beta}$ no matter where the mass of X falls. This makes the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$ ancillary for the parameters of the best linear fit.

The lessons from Figure 2 generalize to multiple linear regression with multivariate $\vec{\boldsymbol{X}}$, but the effects illustrated by the figure are magnified. Although diagnosing misspecification may be easy for a single regressor, it becomes a challenge for progressively larger numbers of regressors, and nearly impossible in "modern" settings for which the number of regressors exceeds the sample size, and data analysts often gamble on sparsity.

In short, it is the combination of a misspecified working model and random regressors that produces the complications; it now matters where the regressor values fall. Three questions arise immediately. First, one may wonder about the meaning of slopes when the model is not assumed to be correct. Second, what is the use of predicted values $\hat{y} = \vec{\boldsymbol{x}}'\boldsymbol{\beta}$? Third, what form should statistical inference take when there is no reliance on the usual assumptions? We will discuss possible answers to these questions in the sections ahead.

6 The Impact of Regressor Distributions Illustrated

The difficulties illustrated by Figure 2 suggests possibilities that may occur in various applications, ranging from modeling of grouped data to meta-analysis. Consider the following hypothetical scenarios that should serve as cautions when interpreting models that are approximations.

Imagine a study of employed females and males in a certain industry, with income as response and a scale measuring educational level as regressor. Consider next the possibility that there is one conditional mean function for income irrespective of gender, but the mean function may be nonlinear in the education scale, as illustrated by the left side picture in Figure 2. A data analyst may fit a linear model, perhaps because of convention, a high level of noise obscuring the nonlinearity, or a lack of graphical data exploration. The analyst may then find that different slopes are required for males and females and may respond by including in the regression an interaction term between gender and education. If, however, the truth is as stipulated, the usual interpretation of interaction effects would be misleading. The driver of the gender difference is not how income responds to education, but the education scale distribution by gender. Put in different language, one may say that the real story is in the consequences of an association between gender and education.

Imagine now meta-analysis of randomized clinical trials (RCTs). RCTs often produce different apparent treatment effects for the same intervention, sometimes called "parameter heterogeneity." Suppose the intervention is a subsidy for higher education, and the response is income at some defined end point. In two different locales, the average education levels may differ. Consequently, in each setting the interventions work off different baselines. There can be an appearance of different treatment effects even though the nonlinear mean returns to education may be the same in both locales. The issue is, once again, that the difference in effects on returns to education may not derive from different conditional mean functions but from differences between regressor distributions.

Apparent parameter heterogeneity also can materialize in the choice of covariates in multiple regression. The coefficient β_1 of the regressor X_1 is not properly interpreted in isolation because β_1 generally depends on which other regressors are included. This is well-known as "confounding." In the simplest case, a regression on X_1 alone, differs from a regression on X_1 and X_2 when the two regressors are correlated. In the extreme, the coefficients β_1 obtained from the two regressions may have different signs, suggesting an instance of Simpson's paradox. (See Berk et al. 2013, Section 2.1, for a more detailed discussion.) For present purposes, exclusion versus inclusion of X_2 can be interpreted as a difference in regressor distributions.

7 Estimation and Standard Errors

Given iid multivariate data $(Y_i, \vec{\boldsymbol{X}}_i) \sim \boldsymbol{P}$ $(i = 1, \ldots, n)$, one can apply OLS and obtain the plug-in estimate $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{P}}_n)$ derived from (1), where $\hat{\boldsymbol{P}}_n$ denotes the empirical distribution of the dataset. By multivariate central limit theorems, $\hat{\boldsymbol{\beta}}$ is asymptotically unbiased and normally distributed, and it is asymptotically efficient in the sense of semi-parametric theory (e.g, Levit 1976, p. 725, ex. 5; Tsiatis, 2006, p. 8 and ch. 4).

7.1 Sandwich Standard Error Estimates

The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ in the assumption lean iid sampling framework deviates from that of linear models theory, which assumes linearity and homoskedasticity. The appropriate expression has a "sandwich" form (White, 1980):

$$\boldsymbol{AV}[\boldsymbol{\beta}, \boldsymbol{P}] = \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1} \, \boldsymbol{E}[\delta^2 \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'] \, \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}. \tag{5}$$

A plug-in estimator is obtained as follows:

$$\widehat{\boldsymbol{AV}} = \boldsymbol{AV}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{P}}_n] = \left(\frac{1}{n}\sum_i \vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i'\right)^{-1} \left(\frac{1}{n}\sum_i r_i^2 \vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i'\right) \left(\frac{1}{n}\sum_i \vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i'\right)^{-1}, \tag{6}$$

where $r_i = Y_i - \vec{\boldsymbol{X}}_i'\hat{\boldsymbol{\beta}}$ are the sample residuals. Equation (6) is the simplest form of a sandwich estimator of asymptotic variance. More refined forms exist but are outside the scope of this article. Standard error estimates for OLS slope estimates $\hat{\beta}_j$ are obtained from (6) using the asymptotic variance estimate in the j'th diagonal element:

$$SE_j = \left(\frac{1}{n} \widehat{\boldsymbol{AV}}_{j,j} \right)^{1/2}.$$

A connection with linear models theory is as follows. If the truth is linear and homoskedastic, and hence, the working model is correctly specified to first and second order, the sandwich formula (5) collapses to the conventional formula for asymptotic variance due to $\boldsymbol{E}[\delta^2 \vec{\boldsymbol{X}}\vec{\boldsymbol{X}}'] = \sigma^2 \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']$, which follows from $\boldsymbol{E}[\delta^2|\vec{\boldsymbol{X}}] = \boldsymbol{E}[\epsilon^2|\vec{\boldsymbol{X}}] = \sigma^2$. The result is $\boldsymbol{AV}[\boldsymbol{\beta}, \boldsymbol{P}] = \sigma^2 \boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}']^{-1}$, the "assumption laden" form of asymptotic variance.

7.2 Bootstrap Standard Error Estimates

Alternative standard error estimates can be obtained from the nonparametric pairwise or x-y bootstrap, which resamples tuples $(Y_i, \vec{\boldsymbol{X}}_i)$. It is assumption lean in that it relies for asymptotic correctness only on iid sampling of the tuples $(Y_i, \vec{\boldsymbol{X}}_i)$ and some moment conditions. The x-y bootstrap, therefore, applies to all manners of regressions, including GLMs.

In contrast, the residual bootstrap is inappropriate because it assumes first order correctness, $\mu(\vec{\boldsymbol{x}}) = \boldsymbol{\beta}'\vec{\boldsymbol{x}}$, as well as exchangeable and hence, homoskedastic population residuals δ. The only step toward assumption leanness is a relaxation of Gaussianity of the noise distribution. Furthermore, it does not apply to other forms of regression such as logistic regression. The residual bootstrap is preferred by those who insist that one should condition on the regressors because they are ancillary. As argued in Section 5, however, the ancillarity argument requires correct specification of the regression model, counter to the idea that models are just approximations.

Sandwich and bootstrap estimators of standard error are identical in the asymptotic limit, and for finite data they tend to be close. Based on either, one may perform conventional statistical tests and form confidence intervals. Although asymptotics are a justification for either, one of the advantages of the bootstrap is that it lends itself to a diagnostic for assessing whether asymptotic normality is a reasonable assumption. One simply creates normal quantile plots of bootstrap estimates obtained in the requisite simulations.

Finally, bootstrap confidence intervals have been addressed in extensive research showing that there are variants that are higher order correct. See for example Hall (1992), Efron and Tibshirani (1994), Davison and Hinkley (1997). An elaborate double-bootstrap procedure for regression is described in McCarthy et al. (2017).

8 Slopes from Best Approximations

When the estimation target is the best linear approximation, one can capitalize on desirable model-robust properties not available from assumption laden linear models theory. The price is that subject-matter interpretations address features of the best linear approximation, not that of a "generative truth;" which, as we have emphasized, is often an unrealistic notion. (Even the assumption of iid sampling adopted here is often unrealistic.)

The most important interpretive issue concerns the regression coefficients of the best linear approximation. The problem is that the standard interpretation of a regression coefficient is not strictly applicable anymore. It no longer holds that

> β_j *is the average difference in* Y *for a unit difference in* X_j *at constant levels of all other regressors* X_k*.*

This statement uses the classical "ceteris paribus" (all things being equal) clause, which only holds when the response function is linear. For proper interpretation that accounts for misspecification, one needs to reformulate the statement in a way that clearly refers to differences in the best approximation $\boldsymbol{\beta}'\vec{\boldsymbol{x}}$, not to differences in the conditional means $\mu(\vec{\boldsymbol{x}})$:

> β_j *is the difference in* ***the best linear approximation to*** Y *for a unit difference in* X_j *at constant levels of all other regressors* X_k*.*

This restatement, unsatisfactory as it may appear at first sight, implies an appropriate admission that there could exist a discrepancy between $\boldsymbol{\beta}'\vec{\boldsymbol{x}}$ and $\mu(\vec{\boldsymbol{x}})$. The main point is that interpretations of regression coefficients should refer not to the response but to the best approximation. This mandate is not particular to OLS linear regression but applies to all types of regressions, as will be rehearsed below for Poisson regressions.

9 Predicted Values $\hat{y}$ from Best Approximations

Also important in regression analysis are the predicted values at specific locations $\vec{\boldsymbol{x}}$ in regressor space, estimated as $\hat{y}_{\vec{\boldsymbol{x}}} = \hat{\boldsymbol{\beta}}'\vec{\boldsymbol{x}}$. In linear models theory, for which the model is assumed correct, there is no bias if it is the response surface that is estimated by predicted values; $\boldsymbol{E}[\hat{y}_{\vec{\boldsymbol{x}}}] = \boldsymbol{\beta}'\vec{\boldsymbol{x}} = \mu(\vec{\boldsymbol{x}})$ because $\boldsymbol{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, where $\boldsymbol{E}[\ldots]$ refers only to the randomness of the response values y_i with the regressor vectors $\vec{\boldsymbol{X}}_i$ treated as fixed.

When the model is mean-misspecified such that $\mu(\vec{\boldsymbol{x}}) \neq \boldsymbol{\beta}'\vec{\boldsymbol{x}}$, then $\hat{y}_{\vec{\boldsymbol{x}}}$ is an estimate of the best linear approximation $\boldsymbol{\beta}'\vec{\boldsymbol{x}}$, not $\mu(\vec{\boldsymbol{x}})$, hence, there exists bias $\mu(\vec{\boldsymbol{x}}) - \boldsymbol{\beta}'\vec{\boldsymbol{x}} = \eta(\vec{\boldsymbol{x}})$ that does not disappear with increasing sample size n. Insisting on consistent prediction with linear equations at a specific location $\vec{\boldsymbol{x}}$ in regressor space is, therefore, impossible.

In order to give meaning to predicted values $\hat{y}_{\vec{\boldsymbol{x}}}$ under misspecification, it is necessary to focus on a population of future observations $(Y_{future}, \vec{\boldsymbol{X}}_{future})$ and to assume that it follows the same joint distribution $\boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}$ as the past training data $(Y_i, \vec{\boldsymbol{X}}_i)$. In particular, the future regressors are not fixed but random according to $\vec{\boldsymbol{X}}_{future} \sim \boldsymbol{P}_{\vec{\boldsymbol{X}}}$. If this is a reasonable assumption, then $\hat{y}_{\vec{\boldsymbol{X}}_{future}}$ is indeed the best linear prediction of $\mu(\vec{\boldsymbol{X}}_{future})$ and Y_{future} for this future population under squared error loss. Averaged over future regressor vectors, there is no systematic bias because $\boldsymbol{E}[\eta(\vec{\boldsymbol{X}}_{future})] = 0$ according to (4) of Section 4.[1] Asymptotically correct prediction intervals for Y_{future} do exist and, in fact, one can use the usual intervals of the form

$$PI_n(\vec{\boldsymbol{x}}; K) \;=\; \left[\hat{y}_{\vec{\boldsymbol{x}}} \pm K \cdot \hat{\sigma} \cdot \left(1 + \vec{\boldsymbol{x}}' \big(\sum \vec{\boldsymbol{X}}_i' \vec{\boldsymbol{X}}_i \big)^{-1} \vec{\boldsymbol{x}} \right) \right]. \tag{7}$$

However, the usual multiplier K is based on linear models theory with fixed regressors, and hence, is not robust to misspecification. There exists a simple alternative for choosing K that has asymptotically correct predictive coverage under misspecification. It can be obtained by calibrating the multiplier K empirically on the training sample such that the desired fraction $1 - \alpha$ of observations $(Y_i, \vec{\boldsymbol{X}}_i)$ falls in their respective intervals. One estimates $\hat{K}$ by satisfying an approximate equality as follows, rounded to $\pm 1/n$:

$$\frac{1}{n} \cdot \# \left\{ i \in \{1, \ldots, n\} : \; Y_i \in PI(\vec{\boldsymbol{X}}_i; \hat{K}) \right\} \;\approx\; 1 - \alpha.$$

Under natural conditions, such multipliers yield asymptotically correct prediction coverage:

$$\boldsymbol{P} \left[Y_{future} \in PI(\vec{\boldsymbol{X}}_{future}; K) \right] \;\to\; 1 - \alpha \quad \text{as} \quad n \to \infty,$$

where $\boldsymbol{P}[\ldots]$ accounts for randomness in the training data as well as in the future data. When the ratio p/n is unfavorable, one may consider a cross-validated version of calibration for $\hat{K}$. Finally we note that empirical calibration of prediction intervals generalizes to arbitrary types of regression with a quantitative response.

10 Causality and Best Approximation

Misspecification creates important challenges for causal inference. Consider first a randomized experiment with potential outcomes Y_1, Y_0 for a binary treatment/intervention $C \in \{0, 1\}$. Because of randomization, the potential outcomes are independent of the intervention: $(Y_1, Y_0) \perp\!\!\!\perp C$. Unbiased estimates of the *Average Treatment Effect* (ATE) follow. Pre-treatment covariates $\vec{\boldsymbol{X}}$ can be used to increase precision (reduce standard errors), similar to control variates in Monte Carlo (MC) experiments. It has been known for some time that the model including the treatment C and the pre-treatment covariates $\vec{\boldsymbol{X}}$ does not need to be correctly specified to provide correct estimation of the ATE and (possibly) an asymptotic reduction of standard errors. That is, the model $Y \sim \tau C + \boldsymbol{\beta}' \vec{\boldsymbol{X}}$ may be arbitrarily misspecified, and yet the ATE agrees with the treatment coefficient τ. (To yield a benefit, however, the covariates $\vec{\boldsymbol{X}}$ must produce a useful increase in R^2 or some other appropriate measure of fit, similar to control variates in MC experiments.)

[1] When regressors are treated as random, there exists a small estimation bias, $\boldsymbol{E}[\hat{\boldsymbol{\beta}}] \neq \boldsymbol{\beta}$ in general, because $\boldsymbol{E}[(\frac{1}{n} \sum \vec{\boldsymbol{X}}_i' \vec{\boldsymbol{X}}_i)^{-1} (\frac{1}{n} \sum \vec{\boldsymbol{X}}_i Y_i)] \neq \boldsymbol{E}[\vec{\boldsymbol{X}}' \vec{\boldsymbol{X}}]^{-1} \boldsymbol{E}[\vec{\boldsymbol{X}} Y]$, causing $\boldsymbol{E}[\hat{y}_{\vec{\boldsymbol{x}}}] \neq \boldsymbol{\beta}' \vec{\boldsymbol{x}}$ for fixed $\vec{\boldsymbol{x}}$. However, this bias is of small order in n and shrinks rapidly with increasing n.

Now consider observational studies. There can be one or more variables that are thought of as causal and which can at least in principle be manipulated independently of the other covariates. If there is just one causal binary variable C, we are returned to a model of the form $Y \sim \tau C + \boldsymbol{\beta}' \vec{\boldsymbol{X}}$, where it would be desirable for τ to be interpretable as an average treatment effect (Angrist and Pischke, 2009, Section 3.2). These are always very strong claims that often call for special scrutiny. It is widely known that causal inference can be properly justified by assuming one of two sufficient conditions, known as "double robustness" (see, e.g, Bang and Robins 2005, Rotnitzki et al. 2012): (1) Either $\mu(\vec{\boldsymbol{x}})$ is correctly specified, which in practice means that there is no "omitted variables" problem for the response and that the fitted functional form for $\mu(\vec{\boldsymbol{x}})$ is correct; or (2) the conditional probability of treatment (called the propensity score) can be correctly modeled, which in practice means that there is no omitted variables problem for treatment probabilities and that the (usually logistic) functional form of the propensity scores is correct. In either case, omitted variable concerns are substantive and can not be satisfactorily addressed by formal statistical methods (Freedman, 2004). There exist diagnostic proposals based on proxies for potentially missing variables or based on instrumental variables, but their assumptions are hardly lean (e.g., Hausman 1978). Misspecification of the functional form in (1) or (2) is probably more amenable to formal diagnostics.

In summary, causal inferences based on observational data are fragile because they depend on one of two kinds of correct specification. Best approximation under misspecification won't do. As a consequence, tremendous importance can fall to misspecification diagnostics. Some useful proposals are given in Buja et al. (2018b).

11 A Generalization: Assumption-Lean Poisson Regression

An obvious generalization of assumption lean modeling is to regressions other than linear OLS, such as generalized linear models. We mention here Poisson regression, to be illustrated with an application in the next section. The response is now a counting variable, which suggests modeling conditional counts with a suitable link function and an objective function other than OLS, namely, the negative log-likelihood of a conditional Poisson model. Interpreting the parameters as functionals allows the conditional distribution of the counting response to be largely arbitrary; the Poisson model does not need to be correct. The working model is a mere heuristic that produces a plausible objective function.

For a counting response $Y \in \{0, 1, 2, ...\}$, one models the log of the conditional expectations of the counts, $\mu(\vec{\boldsymbol{x}}) = \boldsymbol{E}[Y | \vec{\boldsymbol{X}} = \vec{\boldsymbol{x}}]$, with a linear function of the regressors:

$$\log(\mu(\vec{\boldsymbol{x}})) \;\approx\; \boldsymbol{\beta}' \vec{\boldsymbol{x}}.$$

We use "$\approx$" rather than "$=$" to indicate an approximation that allows varying degrees of misspecification. The negative log-likelihood of the model when $n \to \infty$ results in a population objective function whose minimization produces the statistical functional, treated as an estimand or "population parameter:"

$$\boldsymbol{\beta}(\boldsymbol{P}) \;=\; \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\; \boldsymbol{E}\left[\exp\left(\vec{\boldsymbol{X}}'\boldsymbol{\beta}\right) - \left(\vec{\boldsymbol{X}}'\boldsymbol{\beta}\right)Y\right]. \tag{8}$$

The usual estimates $\hat{\boldsymbol{\beta}}$ are obtained by plug-in, replacing the expectation with the mean over the observations and thereby reverting to the negative log-likelihood of the sample.

Interpretations and practice follow much as earlier, with the added complication that the best approximation to $\mu(\vec{\boldsymbol{x}})$ has the form $\exp(\boldsymbol{\beta}'\vec{\boldsymbol{x}})$. The approximation discrepancy $\mu(\vec{\boldsymbol{x}}) - \exp(\boldsymbol{\beta}'\vec{\boldsymbol{x}})$ does not disappear with more data. For statistical tests and confidence intervals, one should use standard error estimates of the appropriate sandwich form or obtained from the nonparametric x-y bootstrap. Finally, under misspecification the regression functional $\boldsymbol{\beta}(\boldsymbol{P})$ will generally, as before, depend on the regressor distribution $\boldsymbol{P}_{\vec{\boldsymbol{X}}}$. The regressors should not be treated as ancillary and not held fixed. The regression functional $\boldsymbol{\beta}(\boldsymbol{P})$ can have different values depending on where in regressor space the data fall.

12 An Empirical Example Using Poisson Regression

We apply Poisson regression to criminology data where the response is the number of charges filed by police after an arrest. One crime event can lead to one charge or many. Each charge for which there is a guilty plea or guilty verdict will have sanctions specified by statute. For example, an aggravated robbery is defined by the use of a deadly weapon, or an object that appears to be a deadly weapon, to take property of value. If that weapon is a firearm, there can then be a charge of aggravated robbery and a second charge of illegal use of a firearm with possible penalties for each. In this illustration, we consider correlates of the number of charges against an offender filed by the police.

The dataset contains 10,000 offenders arrested between 2007 and 2015 in a particular urban jurisdiction. The data are a random sample from over three hundred thousand offenders arrested in the jurisdiction during those years. This pool is sufficiently large to make an assumed infinite population and iid sampling good approximations. During that period, the governing statutes, administrative procedures, and mix of offenders were effectively unchanged; there is a form of criminal justice stationarity. We use as the response variable the number of charges associated with the most recent arrest. The regression exercise is, therefore, not about the number of arrests of a person but about a measure of severity of the alleged crimes that led to the latest arrest. Several regressors are available, all thought to be related to the response. Many other relevant regressors are not available, such as the consequences of the crime for its victims.

We make no claims of correct specification or causal interpretation for the adopted Poisson model. In particular, the binary events constituting the counts do not need to be independent, an assumption that would be unrealistic. For example, if the crime is an armed robbery and the offender struggles with an arresting officer, the charges could be aggravated robbery and resisting arrest. Ordinarily, such dependence would be a concern.

	Coeff	SE	p-value	Boot.SE	Sand.SE	Sand-p
(Intercept)	1.8802	0.0205	0.0000	0.0522	0.0526	0.0000
Age	-0.0147	0.0006	0.0000	0.0016	0.0016	0.0000
Male	0.0823	0.0127	0.0000	0.0284	0.0299	0.0058
Number of Priors	0.0031	0.0002	0.0000	0.0005	0.0005	0.0000
Number of Prior Sentences	0.0002	0.0016	0.8868	0.0040	0.0039	0.9519
Number of Drug Priors	-0.0138	0.0008	0.0000	0.0021	0.0020	0.0000
Age At First Charge	0.0028	0.0009	0.0012	0.0022	0.0021	0.1935

Table 1: Poisson Regression Results for The Number of Crime Charges (n=10,000)

The results of the Poisson regression are shown in Table 1. The columns contain, from left to right, the following quantities:

1. the name of the regressor variable;
2. the usual Poisson regression coefficient;
3. the conventional standard errors;
4. the associated p-values;
5. standard errors computed using a nonparametric *x*-*y* bootstrap;
6. standard errors computed with the sandwich estimator; and
7. the associated sandwich p-values.

Even though the model is likely misspecified by conventional standards for any number of reasons, the coefficient estimates for the population approximation are asymptotically unbiased for the population best approximation. In addition, asymptotic normality holds and can be leveraged to justify approximate confidence intervals and p-values based on sandwich or *x*-*y* bootstrap estimators of standard error. With 10,000 observations, the asymptotic results effectively apply.[2] None of this would be true for inferences based on assumption-laden theories that assume the working model to be correct.

The marginal distribution of the response is skewed upward with the number of charges ranging from 1 to 40. The mean is 4.7 and the standard deviation 5.5. Most offenders have relatively few charges, but a few offenders have many.

Table 1 shows that some of the bootstrap and sandwich standard errors are rather different from the conventional standard errors, indicating indirectly that the conditional Poisson model is misspecified (Buja et al. 2018a). Moreover, there is a reversal of the test's conclusion for "Age at First Charge" (i.e., the earliest arrest that led to a charge as an adult). The null hypothesis is rejected with conventional standard errors but is not rejected with a bootstrap or sandwich standard error. This correction is helpful because past research has often found that the slope of "Age At First Charge" is negative. Typically,

[2] QQ plots of the bootstrap empirical sampling distributions showed close approximations to normality.

individuals who have an arrest and a charge at an early age are more likely to commit crimes later on for which there can be multiple charges.

In the Poisson working model the interpretation of estimated coefficients are about the *estimated best approximation to the conditional response mean.* Attempting to show full awareness of this fact when interpreting coefficients makes for clumsy formulations, hence some simplifications are in order. We will use the shortened expression *response approximation*, which in the current context becomes *charge count approximation.*

The Poisson model implies that the charge count approximation has the form $\hat{y}_{\vec{x}} = \exp(\sum_j \hat{\beta}_j x_j)$. Hence, a unit difference in x_j implies a multiplier of $\exp(\hat{\beta}_j)$ in the charge count approximation, also equivalent to a percentage difference of $(\exp(\hat{\beta}_j) - 1) \cdot 100\%$. Also, for the approximation it is correct to apply the ceteris paribus clause "at fixed levels of all other regressors". It will be implicitly assumed but not explicitly repeated in what follows.

We now interpret each regressor accordingly if the null hypothesis is rejected based on p-values from sandwich/bootstrap standard errors.

- Age: Starting at the top of Table 1, a difference of ten years of age multiplies the charge count approximation by a factor of 0.86. This suggests that older offenders who commit crimes tend to have fewer charges, perhaps because their crimes are different from those of younger offenders.

- Male: According to the approximation, at the same levels of all other covariates, men on the average have an 8% greater number of charges than women.

- Number of Priors: To get the same 8% difference in the charge count approximation from the number of all prior arrests takes an increment of about 25 priors. Such increments are common in the data: about 25% of the cases are first offenders (i.e., no prior arrests), and another 30% have 25 or more prior arrests.

- Number of Drug Priors: According to the approximation, a greater number of prior arrests for drug offenses implies on average fewer charges after controlling for the other covariates. Drug offenders often have a large number of such arrests, so the small coefficient of -0.0138 matters: for 20 additional prior drug arrests, there is a 24% reduction in the charge count approximation. This agrees with the expectation that a long history of drug abuse can be debilitating so that the crimes committed are less likely to involve violence and often entail little more than drug possession.

In summary, the model approximation suggests that offenders who are young males with many prior arrests not for drug possession tend to have substantially more criminal charges. Such offenders perhaps are disproportionately arrested for crimes of violence in which other felonies are committed as well. A larger number of charges would then be expected.

Questions of causality must be left without answers for several reasons. The regressors represent variables that are not subject to intervention. If some of the regressors were causally interpreted, they would affect other regressors downstream. Most importantly, however, the data are missing essential causal factors such as gang membership.

If causal inference is off the table, what have we gained? Though not literally a replication, we reproduced several associations consistent with past research. In an era when the reproducibility of scientific research is being questioned, consistent findings across studies are encouraging. Equally important, the findings can inform the introduction of real interventions that could be beneficial. For example, our replication of the importance of age underscores the relevance of questions about the cost-effectiveness of very long prison sentences and reminds us that the peak crime years are usually during the late teens and early 20s. Priority might be given to beneficial interventions in early childhood, for which there exists strong experimental evidence (e.g., Olds, 2008). On the other hand, if the goal is risk assessment in criminal justice (Berk, 2012), the associations reported here may point to predictors that could help improve existing risk assessment instruments.

There are also statistical lessons. We have seen indirect indications of model misspecification in part because traditional model-trusting standard errors differ from assumption lean sandwich and x-y bootstrap standard errors. As a consequence of model misspecification, it is likely that the parameters of the best fitting model depend on where in regressor space the mass of the regressor distribution falls. This raises concerns about the performance of out-of-sample prediction. If the out-of-sample data are not derived from a source stochastically similar to that of the analyzed sample, such predictions may be wildly inaccurate.

13 Conclusions

Treating models as best approximations should replace treating models as if they were correct. By using best approximations of a fixed model, we explicitly acknowledge approximation discrepancies, sometimes called "model bias," which do not disappear with more data. Contrary to a common misunderstanding, model bias does not create asymptotic bias in parameter estimates of best approximations. Rather, parameters of best approximations are estimated with bias that disappears at the usual rapid rate.

In regression, a fundamental feature of best approximations is that they depend on regressor distributions. Two consequences follow immediately. First, the target of estimation depends on where the regressor distribution falls. Second, one cannot condition on regressors and treat regressors as fixed. Regressor variability must be included in treatments of the sampling variability for any estimates. This can be achieved by using model robust standard error estimates in statistical tests and confidence intervals. Two choices are readily available: sandwich estimators and bootstrap-based estimators of standard errors. In addition, a strong argument can be made in favor of the nonparametric x-y bootstrap over the residual bootstrap, because conditioning on the regressors and treating them as fixed is incorrect when there is model misspecification.

We also described some ways in which the idea of models as approximations requires re-interpretations in practice: (1) model parameters need to be re-interpreted as regression functionals, characterizing best approximations; (2) predictions are for populations rather than at fixed regressor locations and need to be calibrated empirically, not relying on model-based multipliers of pointwise prediction error; and (3) estimation of causal effects from observational data is fragile because it depends critically on correct specification of either response means or treatment probabilities.

In summary, it is easy to agree with G.E.P. Box' famous dictum, but there are real consequences that cannot be ignored or minimized by hand waving. Realizing that models are approximations affects how we interpret estimates and how we obtain valid statistical inferences and predictions.

References

Angrist, J.D. and Pischke. J.-S. (2009) *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Bang, H., and Robins, J.M. (2005) "Doubly Robust Estimation in Missing Data and Causal Inference," *Biometrics* 61, 962–972.

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2017) "Uniformly valid confidence intervals post-model-selection." arXiv:1611.01043.

Berk, R.A., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013) "Valid Post-Selection Inference." *The Annals of Statistics* 41(2): 802–837.

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

Berk, R.A. (2012) *Criminal Justice Forecasts of Risk: A Machine Learning Approach.* New York: Springer

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

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhan, K., and Zhao, L. (2018a) "Models as Approximations — Part I: A Conspiracy of Nonlinearity and Random Regressors in Linear Regression." arXiv:1404.1578

Buja, A., Berk, R., Brown, L., George, E., Arun Kuman Kuchibhotla, and Zhao, L. (2018b) "Models as Approximations — Part II: A General Theory of Model-Robust Regression." arXiv:1612.03257

Davison, A.C., and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*, Cambridge University Press.

Cox, D.R. (1995) "Discussion of Chatfield" (1995). *Journal of the Royal Statistical Society*, Series A 158 (3), 455–456.

Efron, B., and Tibshirani, R.J. (1994) *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press.

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

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.

Hall, P. (1992) *The Bootstrap and Edgeworth Expansion.* (Springer Series in Statistics) New York, NY: Springer Verlag.

Hausman, J. A. (1978) "Specification Tests in Econometrics." *Econometrica* 46 (6): 1251-–1271.

Hong, L., Kuffner, T.A., and Martin R. (2016) "On Overfitting and Post-Selection Uncertainty Assessments." *Biometrika* 103: 1–4.

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

Kuchibhotla, A.K., Brown L.D., Buja, A., George, E., Zhao, L. (2018) "A Model Free Perspective for Linear Regression: Uniform-in-model Bounds for Post Selection Inference." arXiv:1802.05801

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

Levit, B. Y. (1976) "On the efficiency of a class of non-parametric estimates." Theory of Probability & Its Applications, 20(4):723–740.

McCarthy, D., Zhang, K., Berk, R.A., Brown, L., Buja, A., George, E., and Zhao, L.(2017) "Calibrated Percentile Double Bootstrap for Robust Linear Regression Inference." *Statistica Sinica*, forthcoming

Olds, D.L., (2008) "Preventing Child Maltreatment and Crime with Prenatal and Infancy Support of Parents: The Nurse-Family Partnership." *Journal of Scandinavian Studies of Criminology and Crime Prevention*, 9(S1): 2-24.

Rotnitzky, A., Lei, Q., Sued, M. and Robins, J. M. (2012). "Improved Double-Robust Estimation in Missing Data and Causal Inference Models,"*Biometrika* 99, 439—456.

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

Searle, S.R. (1970) *Linear Models.* New York: John Wiley.

Lee, J. D., Sun, D.L., Sun, Y., and Taylor, J.E. (2016) "Exact Post-Selection Inference, with Application to the Lasso." *The Annals of Statistics* 44(3): 907–927.

Tsiatis, A.A. (2006) *Semiparametric Theory and Missing Data.* New York: Springer.

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