Misspecification in Econometrics: A Selective Review

Timothy B. Armstrong*
University of Southern California

July 1, 2025

Abstract

This article is a selective review of literatures in statistics and econometrics that have attempted to formalize the idea that statistical procedures based on misspecified models can be used to reach valid conclusions. In addition to papers that explicitly treat misspecification, we discuss approaches from the literatures on nonparametric statistics and set identified models. We note that a large part of all of these literatures take the same conceptual approach to misspecification: expand the misspecified model to form a new model that one can defend as correctly specified. Applying standard statistical concepts to the expanded model then yields procedures that are deemed robust to misspecification. We discuss some examples and review how some familiar estimators from the literatures on robust estimation and nonparametric statistics arise from this approach. We also discuss issues related to model validation and specification testing.

1 Introduction

Aside from purely descriptive work, nearly all empirical research in economics relies on models in some way. A statistical model gives a probabilistic description of how the data at hand was generated. In addition, statistical models often go beyond the data at hand to describe a scientific theory of how some aspect of the world works. This may be as simple as positing that the data generating process will continue when new observations are drawn, or it may involve a theory (sometimes called "structural" or "causal") of some policy change

*email: timothy.armstrong@usc.edu

or intervention. For the purposes of this article, we will say that the model is *misspecified* if either or both of these two aspects of the modeling process do not provide an accurate description of reality.

Empirical studies that use models typically do so in order to draw conclusions about the real world. What can we say about these conclusions in the presence of model misspecification? In many cases, authors will acknowledge the possibility of misspecification while arguing that their conclusions are nonetheless valid or at least useful in some way. However, these arguments are often vague. The lack of precise arguments about the validity of conclusions drawn from misspecified models has drawn criticism. In a discussion of modeling assumptions used in the difference-in-differences literature, Manski and Pepper (2018, p. 237) state: "Empirical researchers often say that such assumptions are approximations, but they do not formalize what this means." Andrews et al. (2017, Section II) note that discussions in empirical papers about modeling assumptions often make claims about the relative importance of these assumptions that "have no obvious formal meaning."

Is there any real meaning to the idea that misspecified models can be taken to data to draw valid or useful conclusions? If so, how can we make this idea precise? Do we need to adjust our estimators, CIs and other statistical procedures to take misspecification into account? If so, how? The present article is a selective review of literatures in statistics and econometrics that have offered answers to these questions. In addition to papers that explicitly frame their contribution as one of "misspecification robustness," we discuss approaches from the literatures on nonparametric estimation and set identified models. We focus on a particular approach to misspecification that encompasses a large part of all three literatures: given a misspecified econometric model, expand the model to a larger model that one can defend as being correctly specified. This approach gives precise answers to the questions posed above: estimators and other procedures should be judged by how they perform in the expanded model. In particular, one does not need new statistical concepts to formalize the notion of "robustness to misspecification" or "validity under misspecification:" one simply applies standard statistical concepts to the expanded model.

While the literatures on misspecification robustness, nonparametric statistics and set identified models all treat misspecification by expanding the original model, these literatures tend to do so in different ways. Nonetheless, many of the problems considered in these literatures are amenable to the same basic idea: compute estimators that optimally trade off worst-case bias and variance in the expanded model and use these estimators to form CIs and other procedures. This approach leads to procedures that have minimax optimality

properties in many settings. Applying this approach after expanding the original model in different ways leads to different "robust" procedures, some of which may be familiar to readers. In particular, the Huber estimator from the robust estimation literature and the local polynomial estimator from the nonparametric statistics literature arise from applications of this approach. We also discuss how similar ideas involving bias-variance tradeoffs play a role in analyzing tests and estimators and forming optimal procedures in set identified econometric models.

The fact that different ways of expanding the model to allow for misspecification lead to different statistical procedures is a manifestation of the fact that how one expands the model to account for missepcification will, in general, affect the conclusions one draws. To make precise the statement that an estimator or other procedure is "robust to misspecification," one must be explicit about how one is expanding the model to allow for misspecification. Thus, the question of how to expand a potentially misspecified model to be robust to misspecification is, itself, an important research topic. In addition to discussing methods for estimation and inference once the model has been expanded to allow for misspecification, we discuss various proposals for expanding the parameter space to account for misspecification in different settings, as well as the role of specification testing and model validation in this endeavor.

The remainder of this paper is organized as follows. Section 2 introduces the basic setup and introduces examples from the literature. Section 3 uses a simple example to show how misspecification can severely affect the conclusions of an analysis. Section 4 reviews concepts from statistical decision theory and statistical practice that can be used to analyze estimators and other procedures once the model has been expanded to allow for misspecification. Section 5 presents a general approach to misspecification robust inference based on bias-variance tradeoffs and explains how it has been used in our running examples. Section 6 discusses other approaches to data analysis in misspecified models. Section 7 discusses the choices involved in expanding the parameter space to allow for misspecification and the role of validation and specification tests. Section 8 discusses some of the historical connections between the literatures on misspecification, nonparametrics and set identified models. Section 9 concludes.

2 Setup

2.1 Basic setup and notation

We consider a general setting where a researcher observes data Y. The researcher specifies a *model*, which posits that the data Y follows a distribution indexed by an unknown parameter θ in a parameter space Θ. We use subscripts E_θ and P_θ to be explicit about expectations and probability statements depending on θ when such explicit notation is needed. Here, θ may include nuisance parameters such as distributions of error terms.

The researcher is interested in some transformation $T(\theta)$. In addition to specifying the data generating process P_θ, the model specifies an interpretation of $T(\theta)$ as an object that is policy relevant or of scientific interest in some way. We will say that the model is *misspecified* if either (1) the data Y does not follow the distribution P_θ for any $\theta \in \Theta$ or (2) the modeling assumptions that lead us to be interested in $T(\theta)$ as a policy relevant parameter fail.

2.2 Misspecification robustness as expanded parameter space

In this review article, we focus on a general approach that formally allows for misspecification by expanding the parameter space. In particular, one posits a new (larger) parameter space Θ and set of distributions P_θ for this parameter space. An important part of this approach is that the transformation $T(\theta)$ must be defined on this larger parameter space so that $T(\theta)$ retains its policy relevant or scientific interpretation.

Definition. We refer to the approach just described as the *expanded parameter space* approach to misspecification. We use the term *baseline model* or *original model* for the original, possibly misspecified model. We use the term *expanded model* to refer to the larger model.

Ideally, the expanded model is defined in such a way that the researcher can argue convincingly that the expanded model is an adequate description of reality and therefore does not itself suffer from misspecification. From a Bayesian perspective, one can think of the expanded model as including all parameters allowed by a reasonable prior (see Remark 4.1 below). While debates about the meaning of mappings between models and reality are not the main focus of this article, we note that a statistical model as defined in this article falls into the general decision theoretic setup of Wald (1950). Therefore, debates about the interpretation of decision theory going back to Savage (1954) are relevant to this question. Recent discussions of this issue in econometrics include Stoye (2012), Manski (2021) and

Hansen and Sargent (2024). See Armstrong et al. (2025) for a recent review of some of these debates as they relate to empirical practice in economics.

Remark 2.1. In all of our examples, the baseline model can be parameterized as a subset of the expanded model. For example, we may write the parameter in the expanded model as $\theta = (\beta, \gamma)$ with parameter space $B \times \Gamma$, where Γ is a subset of a vector space, with the baseline model taking the same form with parameter space $B \times \{0\}$. However, to avoid notational clutter, we will not introduce any general notation to separately describe the baseline model and expanded model, nor will we introduce general notation to parameterize the former as a subset of the latter.

Remark 2.2. Strictly speaking, the parameter θ includes all nuisance parameters such as distributions of error terms. However, often it is possible to treat some of these parameters as fixed and known when asymptotic approximations such as the central limit theorem are used. For example, in the linear regression model (Example 2 below), one can proceed as if the error term is normally distributed with known variance conditional on the covariates and then plug in an initial estimate of the conditional variance to obtain a feasible procedure. The idea of treating some parameters as known for the purpose of asymptotic analysis (or using asymptotics to make other simplifications) can be formalized using asymptotic efficiency theory. See Hirano and Porter (2020) for a recent review and Grama and Nussbaum (2002), Armstrong and Kolesár (2018) and Armstrong and Kolesár (2021b) for some results that are relevant to the examples we consider below. In some parts of this article, we will make use of such approximations and use θ and Θ to denote the parameter and parameter space after an asymptotic approximation of this form where some unknown quantities are fixed.

2.3 Examples

We explain how several examples of approaches to misspecification from the literature fit into our setup.

Example 1. Huber (1964) considered estimation of a location parameter μ in a parametric family. For concreteness, let us consider the normal distribution with variance 1, which was the main case Huber (1964) considered. The baseline model then specifies that the data $Y = (Y_1, \ldots, Y_n)$ are iid with each following the $N(\mu, 1)$ distribution. This fits into our framework with μ playing the role of θ.

The assumption that the sampling distribution is exactly described by a normal distribution is quite strong. Huber (1964) proposed to deal with misspecification by instead assuming

that, given the location parameter μ, $Y_i - \mu$ follows the distribution with cumulative distribution function (cdf) $F = (1-M)\Phi + MH$ where Φ denotes the $N(0,1)$ distribution and H denotes an arbitrary distribution. Here M determines the magnitude of misspecification from the original model that the researcher allows, and is taken to be known.

In our terminology, Huber's setting gives an expanded model with parameter $\theta = (\mu, F)$ and parameter space $\Theta = \Theta(M) = \mathbb{R} \times \mathcal{F}$ where $\mathcal{F}(M)$ is the set of all probability distributions on the real line that take the form $(1-M)\Phi + MH$ for some cdf H. The parameter of interest is still $T(\mu, F) = \mu$. This expanded model is sometimes called the *gross error* model due to the following interpretation: for each i, we draw from the correctly specified $N(\mu, 1)$ model with probability $1 - M$, but with probability $1 - M$ our sampling process makes an error and draws from some distribution that can be arbitrarily different from the $N(\mu, 1)$ distribution.

The gross error model expands the original $N(\mu, 1)$ model by associating the parameter θ not only with the $N(\mu, 1)$ distribution but also with any distribution F that satisfies $d_{\text{g.e.}}(F; N(\mu, 1)) \leq M$, where $d_{\text{g.e.}}(P; Q)$ is the smallest value of $\tilde{M}$ such that P can be written as $P = (1 - \tilde{M})Q + MH$ for some distribution H. The function $d_{\text{g.e.}}(P; Q)$ is one way of quantifying the distance between distributions P and Q. More generally, one can consider other notions of $d(P; Q)$ of distance between distributions, thereby leading to other expanded models. One can also consider baseline parametric models other then the normal location model, or even baseline models that are not fully parametric in the sense that they do not parametrically specify all error distributions. A large literature in statistics and econometrics has considered expanded models that fall into this class. References include Huber (2004); Donoho and Liu (1988); Kitamura et al. (2013); Andrews et al. (2020); Bonhomme and Weidner (2022); Christensen and Connault (2023).

Example 2. Consider a linear regression model: $\{(X_i, Y_i)\}_{i=1}^n$ are iid and we assume

$$Y_i = \psi(X_i)'\beta + U_i, \quad E[U_i|X_i] = 0 \tag{1}$$

for some known $p \times 1$ vector of functions $\psi(x)$. For example, suppose X_i is univariate and we take $\psi(x) = (1, x, \ldots, x^{p-1})$, in which case we are positing a $(p-1)$th order polynomial for the regression function. We take this parametric regression model to be the baseline model, with β as the parameter.

Sacks and Ylvisaker (1978) considered the possibility of misspecification in this baseline

parametric model. In particular, they considered the expanded model

$$Y_i = \psi(X_i)'\beta + r(X_i) + U_i, \quad E[U_i|X_i] = 0, \quad |r(x)| \leq M(x)$$

where $r(\cdot)$ is specification error and $M(x)$ is a bound on specification error posited by the researcher. They referred to this as an *approximately linear* regression model (i.e. approximately linear in $\psi(x)$). In this expanded model, the parameter θ is $(\beta, r(\cdot))$ and the parameter space $\Theta = \mathbb{R}^p \times \{r(\cdot) : |r(x)| \leq M(x) \text{ all } x\}$.[1]

One of the main cases considered by Sacks and Ylvisaker (1978) is when the regression function is p times differentiable near a point x_0, with a bound M on the pth derivative. It then follows from taking a Taylor approximation at $x = x_0$ that the expanded model holds with $\psi(x) = (1, x - x_0, \ldots, (x - x_0)^{p-1})$ and $M(x) = (M/p!) \cdot |x - x_0|^p$. The transformation $T(\beta, r(\cdot)) = \beta_j$ then corresponds to the $(j+1)$th derivative of the regression function at x_0, with the intercept β_1 corresponding to the regression function at x_0. As we will see, the theory of estimation in this setting is tied closely to classical nonparametric estimation theory.

Example 3. The econometrics literature on set identification has considered numerous models that can be interpreted as expanded models for a misspecified baseline model. One strand of this literature, going back to Manski (1989, 1990), has focused on relaxing assumptions of random sample selection and random selection into treatment groups. To give a simple example, consider a binary variable of interest Y_i^* that is observed only for part of the population. We are interested in the distribution of Y_i^*, which can be summarized by $p = P(Y_i^* = 1)$. Letting W_i be an indicator variable for Y_i^* being observed in the sample, we observe $\{(Y_i, W_i)_{i=1}^n\}$ iid where $Y_i = Y_i^* W_i$.

Under the assumption of random selection into the sample, Y_i^* is independent of W_i. Letting $q = P(W_i = 1)$, the joint distribution of an observation Y_i, W_i is then determined by the parameter $\theta = (p, q)$, with $T(p, q) = p$ being the object of interest. Manski (1989, 1990) considered relaxing the assumption of independence between Y_i^* and W_i in various ways. If we make no assumptions on the joint distribution of Y_i^* and W_i, we obtain an expanded model where the distribution of Y_i, W_i can be characterized by the parameter vector $\theta = (p, q_0, q_1)$

[1]Formally, the parameter θ also includes the unknown conditional distribution $G_U(\cdot|x)$ of U_i and the distribution G_X of X_i, so that we can write $\theta = (\beta, G_U(\cdot|x), G_X)$. However, it is often possible to treat the error distribution as known for the purpose of asymptotic analysis as discussed in Remark 2.2. Thus, we can take the parameter to be $\theta = (\beta, G_X)$ for the purpose of asymptotic analysis and, in some cases such as the problem of estimating the regression function at a point considered below, we can treat G_X as fixed and take $\theta = (\beta, r(x))$ to be the unknown parameter in the expanded model and $\theta = \beta$ in the baseline model.

where $p = P(Y_i^* = 1)$, $q_1 = P(W_i = 1|Y_i^* = 1)$ and $q_0 = P(W_i = 0|Y_i^* = 0)$. The object of interest is still $T(p, q_0, q_1) = p$.

Manski (1989, 1990) also considered relaxing the random selection assumption while maintaining various other assumptions on the joint distribution of W_i, Y_i^*, as well as related settings involving covariates, selection into treatment and exclusion restrictions. See Manski (2003) for a review of this and some of the subsequent work by Manski and coauthors on the topic. The literature on partial identification has also considered relaxing various modeling assumptions in other settings, such as equilibrium assumptions in games (e.g. Ciliberto and Tamer, 2009). Further references to this literature can be found in review articles such as Tamer (2010) and Canay et al. (2023).

Example 4. In the linear instrumental variables (IV) model, we observe $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ where

$$Y_i = X_i'\beta + U_i, \quad E[U_i Z_i] = 0.$$

Here, Y_i is a scalar outcome variable, X_i is a $p \times 1$ vector and Z_i is a $k \times 1$ vector with $k \geq p$. The unknown parameter θ includes β as well as the unknown distribution $G_{X,Z,U}$ of X_i, Z_i, U_i and the parameter space Θ consists of $\beta \in \mathbb{R}^p$ and $G_{X,Z,U}$ such that $E[U_i Z_i] = 0$ under the distribution $G_{X,Z,U}$.

A large literature has examined settings where the exogeneity assumption $E[U_i Z_i] = 0$ holds only approximately. A common approach in this literature (e.g. Conley et al., 2010; Masten and Poirier, 2021) is to allow Z_i to enter directly into the outcome equation:

$$Y_i = X_i'\beta + Z_i'\gamma + U_i, \quad E[U_i Z_i] = 0.$$

This leads to an expanded model with parameters β, γ (as well as the nuisance parameter $G_{X,Z,U}$). The parameter of interest is still $T(\theta) = T(\beta, \gamma, G_{X,Z,U}) = \beta$ (or perhaps a single element of β).

Without further restrictions, the expanded model is clearly unidentified: if (β, γ) range over all of $\mathbb{R}^{p+k}$, then the identified set for any element of β will be the entire real line. One approach used in the literature to obtain informative bounds (e.g. Armstrong and Kolesár, 2021b; Conley et al., 2010; Masten and Poirier, 2021) is to place a bound on the magnitude of γ. Formally, this can be done by assuming that $\|\gamma\| \leq M$ for some norm $\|\cdot\|$ and constant M. One can also place other assumptions on γ, such as sign assumptions. Letting Γ denote the set of values of γ that satisfy the researcher's assumptions, the parameter space for

$(\beta, \gamma, G_{X,Z,U})$ is $\mathbb{R}^k \times \Gamma \times \mathcal{G}$ where $\mathcal{G}$ is a set of distributions $G_{X,Z,U}$ for which $E[U_i Z_i] = 0$ for all distributions in $\mathcal{G}$.

More generally, one can consider the nonlinear generalized method of moments (GMM) model, which imposes $Eg(W_i, \beta) = 0$ for data W_i and a function $g()$ specified by the researcher. The expanded model introduces an additional parameter c and specifies $Eg(W_i, \beta) = c$, with some bound on the magnitude of c. The misspecified IV model falls into this framework with $g(W_i, \beta) = (Y_i - X_i'\beta)Z_i$ and $c = EZ_i'\gamma$. This setting has been considered by Andrews et al. (2017, 2020); Armstrong and Kolesár (2021b).

2.4 Defining $T(\theta)$ in the expanded model

As noted above, one part of forming the expanded model is defining the object of interest $T(\theta)$ on the expanded parameter space. This should be done in a way so that $T(\theta)$ retains its interpretation as an object that is policy relevant or of scientific interest. How one defines $T(\theta)$ will sometimes depend on the source of misspecification and how one interprets it, as the following example illustrates.

Example 1 (continued)**.** Recall that we defined the object of interest to be $T(\mu, F) = \mu$ in the expanded model where we observe data Y_i where $Y_i - \mu$ is distributed according to $F = (1-M)\Phi + MH$ with H an unknown cdf. This can be given the following interpretation: we sample from a population described by the $N(\mu, 1)$ distribution but, with probability M, our data is contaminated so that we draw from an arbitrary unknown cdf instead. In a survey setting, this may occur because a proportion M of our respondent's do not take the survey seriously and give arbitrary answers (see the discussion at the beginning of Section 7 in Huber, 1964). In this *contaminated normal model*, the object of interest is still the original $N(\mu, 1)$ distribution, leading to defining $T(\mu, F) = \mu$ (or some other transformation of μ characterizing the $N(\mu, 1)$ distribution) in the expanded model.

Another setting described by the same expanded model is where one is not concerned with data contamination, but where one views the $N(\mu, 1)$ assumption as holding only approximately. That is, we sample from a population $P_{\mu,F}$ with cdf $y \mapsto F(y - \mu)$ where we know that $F = (1 - M)\Phi + MH$ and we do not face any issues with data contamination. In this case, we would define $T(\mu, F)$ to summarize the distribution $P_{\mu,F}$ of the observed data (e.g. the median or mean of this distribution).

2.5 Comparison to the pseudo-parameter approach

We focus in this review on approaches to misspecification that take the object of interest $T(\theta)$ as given. Another approach to misspecification is to take an estimator $\hat{T}$ of $T(\theta)$ that is viewed as reasonable in the baseline model and consider its behavior when this model is misspecified in essentially arbitrary ways (i.e. when we allow the distribution P of the data to vary in essentially arbitrary ways). We then look for an estimand $T(P)$ such that $\hat{T}$ can be viewed as a reasonable estimator for $T(P)$ even when the model is misspecified (i.e when we do not have $P = P_\theta$ for any θ). The estimand $T(P)$ is often referred to as a *pseudo-parameter*. A classic example of this approach is the best linear predictor interpretation of the ordinary least squares (OLS) estimator.

Example 2 (continued)**.** The ordinary least squares (OLS) estimator for β in the linear regression model (1) is given by $\hat{\beta}_{\text{OLS}} = \arg\min_b \sum_{i=1}^n (Y_i - \psi(X_i)'b)^2$. The *best linear predictor* is defined by $T_{\text{BLP}}(P) = \arg\min_b E_P(Y_i - X_i'b)^2$. Even if the linear model (1) doesn't hold, $T_{\text{BLP}}(P)$ is defined and the OLS estimator is consistent for $T_{\text{BLP}}(P)$ so long as X_i and Y_i have finite second moments. For this reason, $T_{\text{BLP}}(P)$ is commonly used as a pseudo-parameter in linear regression settings under misspecification. Tests and CIs based on the OLS estimator are valid for $T_{\text{BLP}}(P)$ even if the original linear model (1) is misspecified so long as one uses robust standard errors as in White (1980a).

The pseudo-parameter approach contrasts with the approach that is the focus of the present review paper, in which one fixes the parameter of interest $T(\theta)$ along with an expanded model and seeks estimators $\hat{T}$ that perform well in this expanded model. While these approaches are conceptually distinct, they are often combined or used in tandem to motivate a particular statistical procedure. For example, Huber used both approaches to motivate a certain class of estimators that trim outliers in the setting of Example 1.

Approaches based on pseudo-parameters are perhaps most useful when the pseudo-parameter can be related to an object of interest in the original model. We now discuss some results of this form that have been obtained in the regression setting.

Example 2 (continued)**.** Coefficients in a linear regression model are often interpreted as causal effects of the given variable. Formally, if X_i consists of a scalar *treatment* variable D_i and additional covariates W_i, then we can interpret $E[Y_i|D_i = d', W_i = w] - E[Y_i|D_i = d, W_i = w]$ as the causal effect of changing D_i from d to d' (conditional on $W_i = w$) under the Neyman-Rubin casual model if we assume that D_i is as good as random conditional on W_i (see Imbens and Wooldridge, 2009). One may also be interested in how $E[Y_i|D_i = d, W_i = w]$

varies with d for purely descriptive reasons, even without a causal model. In other words, letting $T_{w,d,d'} = E[Y_i|D_i = d', W_i = w] - E[Y_i|D_i = d, W_i = w]$, we are interested in $T_{w,d,d'}$ for particular values of w, d and d' or perhaps averages over this quantity for different values of w, d and d'.

Several results are available relating $T_{w,d,d'}$ to the best linear predictor pseduo parameter. Suppose that one specifies the regression function $E[Y_i|D_i, W_i] = D_i\beta + \tilde{\psi}(W_i)'\gamma$ for some function $\tilde{\psi}(W_i)'\gamma$. Then the best linear predictor coefficient on D_i (i.e. the probability limit of the OLS estimate of β) can be written as a weighted average of $T_{w,d,d'}$ over different values of w under certain conditions. Basically, one needs the "control" part of the regression function to be "correctly specified" in the sense that $E[Y_i|D_i = d, W_i = w] = \tilde{\psi}(w)'\gamma$ for some γ for value of d, or one must have $E[D_i|W_i] = \tilde{\psi}(w)'\gamma$ for some γ. Such results are discussed in Angrist and Pischke (2008, Section 3.3.1) and include an influential result in Angrist (1998) for the case where D_i is binary. Earlier results relating regression derivatives to best linear predictor coefficients include White (1980b) and Yitzhaki (1996).

The above example shows that there are cases where the best linear predictor and other pseudo-parameters may be related to objects of interest in useful ways. However, these results can still place strong requirements on correct specification of certain parts of the model: in the example above, the $\tilde{\psi}(W_i)'\gamma$ term still has to be correctly specified. See Goldsmith-Pinkham et al. (2022) and references therein for recent discussions of misapplications of the result in Angrist (1998) described above. Another setting where pseudo-parameters have been related to causal parameters of interest is in the IV setting, where an influential result of Imbens and Angrist (1994) relates the pseudo-parameter estimated by certain IV estimaters to averages of treatment effects over certain subgroups. These results also place strong requirements on correct specification of certain aspects of the model when covariates are included; see Blandhol et al. (2022) for a recent discussion.

In general, pseudo-parameters may not bear any clear relation to the object of interest $T(\theta)$ as it is interpreted in the original model. We illustrate this point by example in the next section.

3 The perils of ignoring bias from misspecification

To illustrate the problems that can arise when one ignores misspecification, consider the regression discontinuity (RD) setting, in which a treatment D_i is determined by a cutoff rule involving a scalar variable X_i, called the running variable: $D_i = I(X_i > c)$ for some known

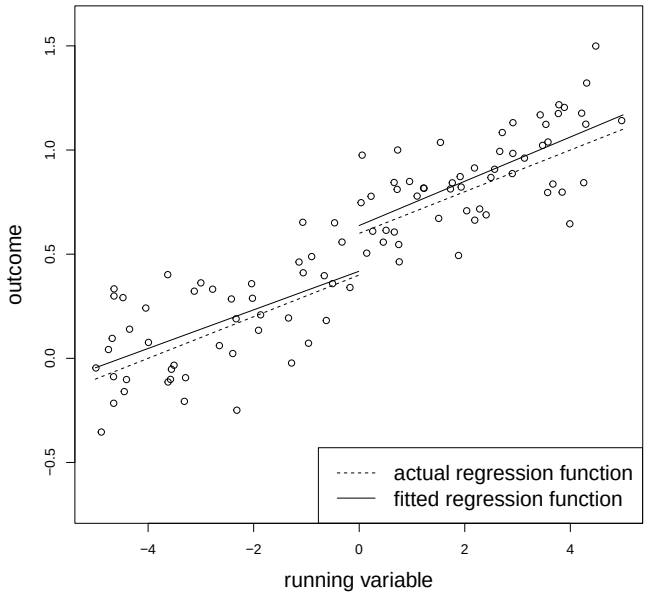

Figure 1: Correctly specified RD

cutoff c. We are interested in the causal effect of the treatment D_i on an outcome variable Y_i. Under continuity assumptions, we can interpret any jump in the regression function $E[Y_i|X_i = x]$ at $x = c$ as an average causal effect of the treatment D_i on the outcome Y_i conditional on $X_i = c$. Details and references can be found in the review article by Imbens and Lemieux (2008).

Issues with bias from misspecified parametric models have been a major concern in empirical papers that use RD designs. Here, we draw from the textbook discussion of these issues in Angrist and Pischke (2008, Ch. 6). Suppose we use a linear model for the regression function $E[Y_i|X_i = x]$ on either side of the cutoff c. This leads to a particular case of the linear model in Example 2:

$$Y_i = (\beta_1 + \beta_2(X_i - c))I(X_i \leq c) + (\beta_3 + \beta_4(X_i - c))I(X_i > c) + U_i, \quad E[U_i|X_i] = 0.$$

We have parameterized the model so that $T(\beta) = \beta_3 - \beta_1$ gives the object of interest: the jump at $x = c$.

What happens if we estimate this model using OLS? Figure 1 shows the OLS fit in a correctly specified model. Here, we find evidence of a positive treatment effect, which turns out to be correct for the simulated model. Figure 2 illustrates what can happen with a

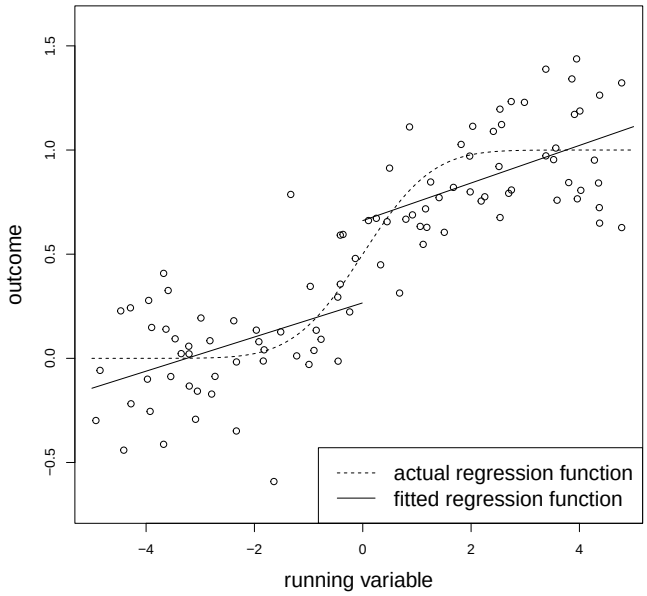

Figure 2: Misspecified RD

misspecified model. Here, the true regression function is continuous, so that the treatment effect estimated in the RD design is zero. Indeed, the regression function is consistent with the treatment effect being zero for all individuals. However, the OLS estimator finds evidence of a positive effect. Furthermore, the spurious effect is statistically significant at conventional levels and sizeable relative to the overall variation in the outcome variable (the point estimate is .39 with a standard error of .096).

The simplicity of this example makes it a useful test case for informal arguments that misspecified parametric models can be expected to lead to conclusions that are approximately correct or useful in some way. Surely the finding of a large positive effect when no effect exists is not an "approximately correct" finding! This example also raises issues with the interpretation of the pseudo-parameter in this setting: the best linear predictor pseudo-parameter suggests a nonzero effect even though the treatment has no effect at all. While pseudo-parameters can be related to objects of causal or scientific interest in some settings, this example illustrates that relying on the pseudo-parameter interpretation of estimands can lead one astray if it is not accompanied by explicit arguments relating the pseudo-parameter to an object of interest.

4 Data analysis in the expanded model

The expanded models in the examples we have covered may appear difficult or nonstandard. Nonetheless, these expanded models are themselves statistical models in the usual sense as defined in textbooks (e.g. Wasserman, 2004, Ch. 6) and used in theories of estimation (Lehmann and Casella, 1998), inference (Lehmann and Romano, 2005) and more general statistical decision problems (going back to Wald (1950); see Berger (1985) for a textbook treatment). In particular, standard ideas about how to form and analyze statistical procedures in a general setting can still be applied. We first review these concepts before discussing how they have been applied in the expanded model approach to misspecification.

4.1 Review of statistical concepts

Just as in more familiar settings, one can form an *estimator* $\hat{T} = \hat{T}(Y)$ of $T(\theta)$ with the goal of making the estimation error $\hat{T} - T(\theta)$ small. One can then analyze the sampling behavior of the error $\hat{T} - T(\theta)$ and report to the reader objects that help to convey the range of possible sampling behavior. Taking a formal decision theoretic perspective, one may choose a *loss function* $\ell(\hat{T}, T)$ and analyze the *risk function* $R(\hat{T}, \theta) = E_\theta \ell(\hat{T}, T(\theta))$. For example, the squared error loss function $\ell(\hat{T}, T) = (\hat{T} - T)^2$ leads to the familiar *mean squared error* (MSE) risk $E_\theta[(\hat{T} - T(\theta))^2]$. To assess the accuracy of the estimator, one may attempt to derive formal bounds on the *worst-case risk* over the parameter space Θ, using asympotic approximations if necessary but ideally reporting exact finite sample bounds. If formal bounds cannot be obtained, one may use Monte Carlo exercises to try to give an idea of the possible sampling behavior of the estimator.

One can also use worst-case risk bounds to compare estimators. An estimator is *minimax* if it minimizes worst-case risk $\sup_{\theta \in \Theta} R_\theta(\hat{T}, T)$ over possible estimators $\hat{T}$. Rather than an exact minimax estimator, one may seek a simple estimator that is highly efficient in the sense that it nearly achieves the minimax bound. One may also take a *Bayesian* approach and choose some prior distribution π over the expanded parameter space. This leads to the *Bayes risk* criterion $\int_{\theta \in \Theta} R_\theta(\hat{T}, T(\theta))\, d\pi(\theta)$.

Standard definitions of a *statistical hypothesis test* and *confidence interval* (CI) also do not necessarily need to be modified. A $100 \cdot (1-\alpha)$ confidence interval $\mathcal{C}$ for $T(\theta)$ is a random set that satisfies $P_\theta(T(\theta) \in \mathcal{C}) \geq 1-\alpha$ for all $\theta \in \Theta$. A hypothesis test $\phi(Y)$ maps the data Y to a zero-one decision to reject or fail to reject some null hypothesis about $T(\theta)$, such as the null hypothesis $H_{T_0} : T(\theta) = T_0$. The test is level α if the rejection probability $P_\theta(\phi(Y) = 1)$

is bounded by α for all θ with $T(\theta) = T_0$. As in more standard settings, one should analyze the *power* $P_\theta(\phi(Y) = 1)$ of the test at alternatives θ where $T(\theta) \neq T_0$ and argue that it is substantially greater than the level α at plausible alternatives θ.

In addition to formal results on the sampling properties of estimators, tests and CIs, one may also use less formal criteria to choose between estimators. For example, one may seek estimators that are transparent and simple to describe. One may also apply statistical decision theory to other decision problems in the expanded model, such as welfare based policy decisions; see Manski (2004), Dehejia (2005).

Remark 4.1. If one adopts a Bayes criterion, one may also ask about misspecification of the prior π. One approach to this issue is to relax the assumption of a single prior by assuming only that the prior is in some given set Γ, leading to the Γ-*minimax criterion* $\sup_{\pi\in\Gamma} \int_{\theta\in\Theta} R_\theta(\hat{T}, T(\theta))\, d\pi(\theta)$. While we do not discuss the question of misspecified priors in this review, we note that prior misspecification is closely related to misspecification of the parameter space: if the parameter space Θ is a subset of a larger set then applying the minimax criterion with parameter space Θ corresponds to the Γ-minimax criterion with Γ given by the set of priors supported on Θ.

4.2 Robust estimation and decision theory

The literature on robust estimation has introduced several concepts that can be understood as applications of standard concepts from statistics and decision theory to the expanded model. The notion of *ambiguity aversion*, often associated with model misspecfication (e.g. Hansen and Sargent, 2024), has been formalized in the axiomatic decision theory literature using minimax and related criteria (Gilboa and Marinacci, 2013). The literature on robust statistics has formulated several notions of "robustness," many of which amount to measuring the robustness of a statistic using its asymptotic worst-case bias or variance (see Donoho and Liu, 1988, and references therein). A related idea in the robust statistics literature is the *breakdown point* of an estimator, which is defined as the smallest expanded model (e.g. the smallest choice of M in Example 1) such that a nontrivial bound can be obtained on the worst-case bias of the given estimator (Huber, 2004, p. 13).

It is helpful to remember that all of these concepts are applications of statistical theory to an expanded model. While concepts such as the minimax criterion are often emphasized in the robust statistics literature, they are also present in more mundane parametric settings. For example, the asymptotic efficiency of maximum likelihood estimators can be formalized using minimax in a particular sequence of local models (van der Vaart, 1998, Ch. 8).

4.3 Complications arising in the expanded model

While the expanded models in our examples are still amenable to the application of the concepts from the theory and practice of statistics described in Section 4.1, they bring up some complications that do not arise in standard parametric settings. These include:

1.) infinite dimensional nuisance parameters, such as the distribution H in Example 1 or the approximation error function $r(\cdot)$ in Example 2

2.) parameter spaces that incorporate bounds or inequalities, such as the bound $M(x)$ on the approximation error $r(x)$ in Example 2, or the bounds on the effect γ of the instruments Z_i in Example 4.

3.) parameters that are *set identified* – for a given distribution P describing the observed data, there may be multiple values of θ or $T(\theta)$ such that $P_\theta = P$.

While set identification can raise conceptual issues, one can still apply the concepts in Section 4.1 when $T(\theta)$ is set identified. This viewpoint has become increasingly common following the influential work of Imbens and Manski (2004) in the context of CIs and several papers by Manski and other authors in the context of policy decisions based on statistical treatment rules (e.g. Manski, 2007).

Conceptual issues aside, set identification may require some modification of the usual methods used to analyze estimators and CIs. For example, if $T(\theta)$ is not point identified, asymptotic analysis of worst-case risk may be complicated by the fact that worst-case bounds on risk will not shrink to a point as the sample size grows. Issues (1) and (2) can also complicate estimation and inference, even when $T(\theta)$ is point identified. The main issue here is that exploiting the bounds defined by the parameter space may lead to biased estimators. Despite these issues, familiar approaches involving approximately normal point estimators and CIs based on these estimators and their standard errors can often be fruitfully applied, with some modifications to take into account bias. We turn to this topic in the next section.

5 Approximately normal estimators

Often, one can construct estimators $\hat{T}$ that are approximately normal in the expanded model:

$$\hat{T} - T(\theta) \overset{d}{\approx} N(\text{bias}(\theta), V(\theta)). \tag{2}$$

Furthermore, an estimate $\hat{V}$ is available of $V(\theta)$ that is accurate enough to be used for inference.[2] These statements can be formalized using appropriate uniform central limit theorems and laws of large numbers that show that remainder terms converge to zero at an appropriate rate, but we do not discuss such formal results here.

In the settings we consider here, the asymptotic bias term $\text{bias}(\theta)$ cannot be consistently estimated, but one can bound this term using the worst-case bias

$$\overline{\text{bias}} = \overline{\text{bias}}(\Theta) = \sup_{\theta \in \Theta} |\,\text{bias}(\theta)|.$$

Thus, to describe the approximate range of possibilities of sampling distributions to the reader, one can report $\overline{\text{bias}}$ along with the standard error. To provide additional information to the reader, one can also report $\overline{\text{bias}}(\Theta)$ for different parameter spaces that allow for different amounts of misspecification from the original model, or one can report $\text{bias}(\theta)$ for particular parameters θ in the expanded model that represent particularly plausible or important deviations from the original model.

Once the worst-case bias has been calculated, it can be used to compute a $100 \cdot (1-\alpha)\%$ CI:

$$\hat{T} \pm [\overline{\text{bias}} + z_{1-\alpha/2}\sqrt{\hat{V}}]$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the $N(0,1)$ distribution. One can also form a less conservative CI by taking into account the fact that the bias cannot be simultaneously equal to $|\overline{\text{bias}}|$ and $-|\overline{\text{bias}}|$, as emphasized by Imbens and Manski (2004). In the case where $V(\theta)$ doesn't depend on θ (or when $V(\theta)$ only depends on certain elements of θ that don't enter into $\text{bias}(\theta)$), this can be done by using a quantile of the $N(\overline{\text{bias}}/\hat{V}, 1)$ distribution as the critical value; see Armstrong and Kolesár (2018). Such CIs that explicitly include bias bounds are sometimes called *bias-aware* CIs (e.g. Noack and Rothe, 2024).

The normal approximation (2) can also be used to choose between estimators. Suppose we have a class of estimators $\hat{T}_w$ where w is some index chosen by the researcher. Suppose the approximation (2) holds for some $\text{bias}(\theta; w)$ and $V(\theta; w)$ for each w. We can then choose the estimator $\hat{T}_w$ that trades off bias and variance in a way that we choose. For example,

[2] As discussed in Remark 2.2, one can often treat certain nuisance parameters such as the conditional distribution of error terms in Example 2 as known for the purpose of this asymptotic approximation. Such nuisance parameters do not need to be included in θ for the purposes of the worst-case bias and variance calculations discussed in this section.

the minimax MSE choice of w minimizes worst-case MSE:

$$\min_w \sup_{\theta \in \Theta} \left\{ \text{bias}(\theta; w)^2 + V(\theta; w)^2 \right\}$$

If we are interested in CIs, we can also directly optimize the length a CI of the form given above.

A concept that can be useful in deriving asymptotic representations of the form (2) is the *influence function*. We will say that a statistic $\hat{T}$ is *asymptotically linear* with influence function $\psi_{\text{IF}}(\cdot)$ if

$$\hat{T} - T(\theta) \approx \frac{1}{n} \sum_{i=1}^{n} \psi_{\text{IF}}(Y_i) \tag{3}$$

in large samples (as with the approximation (2), we do not discuss the formal results that justify this approximation here). The influence function representation (3) leads to the normal approximation (2) with $\text{bias}(\theta) = E_\theta \psi_{\text{IF}}(Y_i)$ and $V(\theta) = \text{var}_\theta(\psi_{\text{IF}}(Y_i))$. The influence function representation can also be helpful in understanding the estimator $\hat{T}$ from a mechanical or algorithmic perspective: $\psi_{\text{IF}}(Y_i)$ measures the contribution or "influence" of observation Y_i on the estimate $\hat{T}$.

In settings where a certain form of linearity holds, an influence function representation (3) will hold *exactly* and can be used to obtain exact finite sample bias and variance calculations. This is the case in Example 2. In nonlinear settings, bias calculations based on the influence function representation (3) will typically hold only as an approximation for small deviations from the original model, an idea that can be formalized using the framework of *local misspecification*. Such approximations are relevant in Examples 1 and 4. We turn to these examples in the next section.

5.1 Examples

Example 1 (continued)**.** In this example, the original model is the $N(\mu, 1)$ location model and the expanded model is the gross error model where $Y_i - \mu \sim F$ where $F = (1-M)\phi + MH$ and H is an arbitrary distribution. In the correctly specified normal location model, the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is asymptotically efficient. Indeed, it is minimax even in finite samples (Lehmann and Casella, 1998, Example 2.8, pp. 324-326). However, in the expanded model, both the bias and variance of $\bar{Y}$ can be made arbitrarily large by taking H to be a point mass at some large value.

To ameliorate this issue, Huber (1964) proposed a class of *M-estimators*, defined as

$$\hat{T} = \arg\min_t \sum_{i=1}^n \rho(Y_i - t)$$

for some function $\rho()$. Huber argued that a particular class of M-estimators, now called *Huber estimators*, have good properties in terms of both the asymptotic bias $\text{bias}(\mu, F)$ and asymptotic variance $V(\mu, F)$ in the gross error model. This class corresponds to taking $\rho(t) = \rho_{\text{Huber},k}(t)$ where $\rho_{\text{Huber},k}(t) = t^2/2$ for $|t| < k$ and $\rho_{\text{Huber},k}(t) = k|t| - k^2/2$ for $|t| \geq k$, where k is a tuning parameter chosen by the user. Note that the Huber estimator is also defined by the first order conditions

$$\sum_{i=1}^n \rho'_{\text{Huber},k}(Y_i - \hat{T}) = 0, \quad \rho'_{\text{Huber},k}(t) = \begin{cases} t & |t| < k \\ 0 & |t| > k \end{cases}$$

Thus, Huber estimators have the effect of trimming outliers.

The class of Huber estimators contain the sample mean and sample median as limiting cases where $k \to \infty$ and $k \to 0$ respectively. In the gross error model, Huber (1964) showed that the Huber estimator with an appropriate choice of k minimizes the worst-case asymptotic variance when the distribution H is constrained to be symmetric. In addition, he showed that the sample median (the limiting case as $k \to 0$) minimizes worst-case asymptotic bias in the gross error model. In subsequent work (Huber, 1965, 1968), Huber developed finite sample optimality properties of this class of estimators for certain estimation and inference problems.

When the $N(\mu, 1)$ model is correctly specified, M-estimators have the influence function representation (3) with influence function

$$\psi_{\text{IF}}(Y_i) = \frac{\rho'(Y_i - \mu)}{\left[\frac{d}{dt}\int \rho'(t - y)\, d\Phi(y)\right]_{t=0}}. \tag{4}$$

This approximation can also be used to approximate the bias of M-estimators when specification error M is small:

$$\text{bias}(\mu, F) \approx E_{\mu,F}\psi_{\text{IF}}(Y_i) = M \int \psi_{\text{IF}}(y)\, dH(y) \quad \text{where } F = (1 - M)\Phi + MH.$$

The absolute value of this quantity is maximized when H places probability one on the value

of y where $|\psi_{\text{IF}}(y)|$ is largest, yielding the approximation

$$\overline{\text{bias}}(\Theta(M)) = \sup_{(\mu,F)\in\Theta(M)} |\,\text{bias}(\mu, F)| \approx M \sup_{y\in\mathbb{R}} |\psi_{\text{IF}}(y)|.$$

to the worst-case asymptotic bias. Due to this approximation, $\sup_{y\in\mathbb{R}} |\psi_{\text{IF}}(y)|$ has been called the *gross error sensitivity* of an estimator with influence function $\psi_{\text{IF}}(y)$; see Huber (2004, p. 14).

Using the influence function to approximate the bias yields insights into why the Huber estimator controls bias from gross error specification. Since $|\rho'(t)|$ is bounded by the user specified parameter k, the influence function (4) is bounded, thereby controlling bias even in the worst case.

Example 2 (continued)**.** A popular approach to estimation in the linear model is the OLS estimator, $\hat{\beta}_{\text{OLS}}$. More generally, one can consider the class of *weighted least squares* (WLS) estimators, which take the form

$$\begin{aligned}\hat{\beta}_{\text{WLS},K_n} &= \arg\min_b \sum_{i=1}^n (Y_i - \psi(X_i)'b)^2 K_n(X_i) \\ &= \left(\sum_{i=1}^n K_n(X_i)\psi(X_i)\psi(X_i)'\right)^{-1} \sum_{i=1}^n K_n(X_i)\psi(X_i)Y_i.\end{aligned}$$

where $K_n(x)$ is a weighting function chosen by the researcher. For concreteness, suppose we are interested in the jth element of β, so that $T(\beta) = e_j'\beta$ where e_j is the jth standard basis vector. Note that the formula for the WLS estimator immediately gives an influence function representation that holds exactly in finite samples.

$$\begin{aligned}&e_j'\hat{\beta}_{\text{WLS},K_n} - e_j'\beta = e_j'\left(\sum_{i=1}^n K_n(X_i)\psi(X_i)\psi(X_i)'\right)^{-1} \sum_{i=1}^n K_n(X_i)\psi(X_i)(Y_i - \psi(X_i)'\beta). \\ &= \frac{1}{n}\sum_{i=1}^n w_n(X_i)(Y_i - \psi(X_i)'\beta) \text{ where} \qquad (5) \\ &w_n(x) = e_j'\hat{\Gamma}_{n,K_n}^{-1} K_n(x)\psi(x), \quad \hat{\Gamma}_{n,K_n} = \frac{1}{n}\sum_{i=1}^n K_n(X_i)\psi(X_i)\psi(X_i)'\end{aligned}$$

Note that (5) is an algebraic identity that holds regardless of β. The influence function takes the form $\psi_{\text{IF}}(X_i, Y_i) = w_n(X_i)(Y_i - \psi(X_i)'\beta)$.[3]

[3]Note that the function $w_n(x)$ depends on the entire sample of covariates $X = X_1, \ldots, X_n$. Replacing

In the baseline model where $E[Y_i|X_i] = \psi(X_i)'\beta$, it follows immediately from (5) that the WLS estimator is unbiased condional on the sample of covariates $X = (X_1, \ldots, X_n)$: $E[e_j'\hat{\beta}_{\text{WLS},K_n}|X] = e_j'\beta$. In the approximately linear model of Sacks and Ylvisaker (1978), we have $E[Y_i|X_i] = \psi(X_i)'\beta + r(X_i)$, which leads to the bias

$$\text{bias}(\beta, r(\cdot); w_n(\cdot)) = E[e_j'\hat{\beta}_{\text{WLS},K_n}|X] - e_j'\beta = \frac{1}{n}\sum_{i=1}^{n} w_n(X_i)r(X_i).$$

Under the approximately linear model, we assume only that $|r(X_i)| \leq M(X_i)$ where $M(X_i)$ is specified by the researcher. This leads to maximum bias conditional on the design points $X = (X_1, \ldots, X_n)$ being given by

$$\overline{\text{bias}}(w_n(\cdot)) = \sup_{\beta \in \mathbb{R}^p, r(\cdot): |r(x)| \leq M(x) \text{ all } x} |\text{bias}(\beta, r(\cdot); w_n(\cdot))| = \frac{1}{n}\sum_{i=1}^{n} |w_n(X_i)M(X_i)| \quad (6)$$

with the maximum being taken when $r(X_i) = M(X_i) \cdot \text{sign}(w_n(X_i))$. The variance (again conditional on $X = (X_1, \ldots, X_n)$) is

$$V(w_n(\cdot)) = \frac{1}{n^2}\sum_{i=1}^{n} w_n(X_i)^2\sigma^2(X_i) \quad \text{where} \quad \sigma^2(x) = E[Y_i|X_i = x].$$

In the baseline model under correct specification, $M(x) = 0$ so that $\overline{\text{bias}}(w_n(\cdot)) = 0$. According to the *Gauss-Markov Theorem*, under homoskedastic errors (i.e. when $\sigma^2(x)$ is constant), the choice of weighting function $K_n(\cdot)$ that minimizes $V(w_n(\cdot))$ subject to the condition that $\overline{\text{bias}}(w_n(\cdot)) = 0$ is the constant weighting function $K_n(x) = 1$, leading to the OLS estimator. Indeed, the Gauss-Markov Theorem makes the stronger statement that OLS is minimum variance unbiased among all *linear estimators*, meaning estimators that take the form $\frac{1}{n}\sum_{i=1}^{n} w_n(X_i)Y_i$ with $w_n(\cdot)$ not necessarily taking the form in (5).

In the expanded model where $r(x)$ may not be zero, unbiasedness is too much to ask for. However, we can trade off bias and variance by minimizing a criterion such as worst-case MSE:

$$\min_{w_n(\cdot)} \overline{\text{bias}}(w_n(\cdot))^2 + V(w_n(\cdot)). \quad (7)$$

$\hat{\Gamma}_{n,K_n}$ with its expectation leads to a influence function representation where $\psi_{\text{IF}}(x, y)$ does not depend on the sample. Here, we avoid this approximation by taking a so-called *fixed design* approach and calculating bias and variance conditional on the sample of covariates $X_1, \ldots, X_n$.

Sacks and Ylvisaker (1978) proposed to generalize the Gauss-Markov Theorem directly by computing a *minimax linear estimator* in the expanded model. In particular, they proposed to choose $w_n(\cdot)$ to minimize worst-case MSE under homoskedasticity. They showed that the solution can be characterized using a Lagrangian and elementary algebraic derivations.

In general, these optimal weights may not take the form of a WLS estimator for any simple weighting function $K_n(\cdot)$. However, in the case where $M(x) = (M/p!)|x - x_0|$ and $\psi(X_i) = (1, X_i - x_0, \ldots, (X_i - x_0)^{p-1})$ so that the approximately linear model follows from a Taylor approximation at some point x_0, a WLS approach with $K_n()$ given by a *kernel function* $K_n(x) = k((x - x_0)/h_n)$ turns out to work well. Here, the kernel function $k(t)$ is taken to be a simple function such that $K_n(x) = k((x - x_0)/h_n)$ downweights observations away from x_0 and h_n is a tuning parameter called the *bandwidth*. Popular choices include the *uniform kernel* $k(t) = I(|t| < 1)$ or the *triangular kernel* $k(t) = \max\{1 - |t|, 0\}$. The resulting estimator is called a *local polynomial estimator*.

Local polynomial estimators with appropriately chosen bandwidth and kernel have been shown to be asympototically optimal or nearly optimal for the MSE minimization problem (7) as $n \to \infty$ under regularity conditions on the density of X_i (Fan, 1993; Cheng et al., 1997). Readers may recognize these results as optimality results for bandwidths, kernels and rates of convergence from the nonparametric statistics literature. Armstrong and Kolesár (2020) provide further references and discussions for such results. Asymptotic results of this form may also be used to compute critical values that take into account maximum bias: Armstrong and Kolesár (2020) show that in the commonly used case of local linear regression ($p = 2$), one can achieve 95% asymptotic coverage by replacing the usual critical value 1.96 with the quantity 2.18. Furthermore, the usual practice of ignoring bias and using the critical value 1.96 yields coverage 92.1% rather than the nominal 95% coverage.

While these results are asympototic, one can always report bias-aware CIs with $\overline{\text{bias}}(w_n(\cdot))$ computed using the exact finite-sample formula (6). If finite sample efficiency is a concern, one can use the exact finite-sample optimal weights proposed by Sacks and Ylvisaker (1978), or one can use the finite sample bias and variance formulas to check the finite sample relative efficiency of simple estimators. Recent applications of these approaches include Armstrong and Kolesár (2018); Kolesár and Rothe (2018); Imbens and Wager (2019).

In addition to being useful for formal calculations of bias and relative efficiency, the influence function representation (5) can also be used as a diagnostic to assess influential observations for WLS estimators without reference to a particular expanded model. Diagnostics based on the influence function include the *leverage* (which corresponds to the

influence function weight for observation i for estimating the fitted value $\psi(X_i)'\beta)$ which appears in textbook discussions of outlier detection for least squares estimators (e.g. Hansen, 2022, Section 3.19). In the regression discontinuity setting, Gelman and Imbens (2017) advocate plotting the influence function weights $w_n(X_i)$ to assess whether the WLS estimator is intuitively reasonable.

Minimax linear estimators have been applied in a variety of settings beyond the approximately linear model. Donoho (1994) presents a general theory and optimality results for this approach along with references to applications. More recent applications include Armstrong and Kolesár (2021a), Kallus (2020) and Hirshberg et al. (2021). Minimax linear estimators and related approaches have also been used in the literature on balancing weights; see Ben-Michael et al. (2021).

Example 4 (continued)**.** Consider the general setting where the baseline model imposes $Eg(W_i, \beta) = 0$ for a function $g()$ specified by the researcher. The expanded model allows for misspecification by imposing $Eg(W_i, \beta) = c$ where c is another unknown parameter, constrained to be in some set $\mathcal{C}$. Suppose we are interested in a differentiable scalar function $h(\beta)$ of the parameter β.

The generalized method of moments (GMM) estimator is given by

$$\hat{\beta}_{\text{GMM},W_n} = \arg\min_b \left(\sum_{i=1}^n g(W_i, b)\right)' W_n \left(\sum_{i=1}^n g(W_i, b)\right)$$

where W_n is a weighting matrix that converges in probability to some matrix W. In the baseline model, the GMM estimator has the influence function representation

$$h(\hat{\beta}_{\text{GMM},W_n}) - h(\beta) \approx \frac{1}{n}\sum_{i=1}^n k_W' g(W_i, \beta) \text{ where } k_W' = -H(\Gamma' W \Gamma)^{-1}\Gamma' W \tag{8}$$

and $\Gamma = \left[\frac{d}{db'} Eg(W_i, b)\right]_{b=\beta}$ is the derivative matrix of $Eg(W_i, b)$ at $b = \beta$. In the baseline model where $c = 0$, this leads to the usual GMM asymptotic variance formula $k_W' \Sigma k_W$ where k_W is given above and $\Sigma = Eg(W_i, \beta)g(W_i, \beta)'$ is the covariance matrix of the moment function $g(W_i, \beta)$.

The influence function representation (8) continues to hold if we consider asymptotics under a sequence of expanded models where $\mathcal{C}$ shrinks at a root-n rate (Newey, 1985). This asymptotic setting, called *local misspecification*, is a useful tool for deriving asymptotic asymptotic approximations where sampling error and specification error are of the same

order of magnitude. In particular, it leads to the normal approximation

$$h(\hat{\beta}_{\mathrm{GMM},W_n}) - h(\beta) \overset{d}{\approx} N(\mathrm{bias}(c; k_W), V(k_W)) \text{ where } \mathrm{bias}(c) = k_W' c,\ V(k_W) = k_W' \Sigma k_W / n.$$

Motivated by this asymptotic bias formula, Andrews et al. (2017) refer to k_W as the *sensitivity* of the estimator $h(\hat{\beta}_{\mathrm{GMM},W_n})$. They advocate reporting k_W along with the GMM estimator so that readers can assess the bias $k_W' c$ under deviations that they find plausible.

By specifying an expanded model where c is in some given set $\mathcal{C}$, one can formalize the idea that the model is "approximately correct." Armstrong and Kolesár (2021b) show how to choose the GMM weighting matrix and corresponding sensitivity k_W in a way that is optimal for a given set $\mathcal{C}$. The calculations involve bias-variance tradeoffs similar to those in the approximately linear model in Example 2 above. The resulting weight functions will depend on the set $\mathcal{C}$: if this set allows for a given component $c_j = Eg(W_i, \beta)$ to be large relative to the bounds on other components of the moment condition, then this moment will receive less weight.

6 Other approaches

In many settings, the approach described in Section 5 of trading off bias and variance to obtain point estimates and CIs works reasonably well. Formal optimality results for such estimators and CIs include Donoho (1994), Armstrong and Kolesár (2018) and Armstrong and Kolesár (2021b) (see also the recent work of Yata (2023) for optimality results on policy decisions based on this class of estimators). In particular, in normal or asympotically normal settings where the expanded model is symmetric around the baseline model and a certain form of linearity or asymptotic linearity holds, bias-aware CIs based on estimators that optimally trade off bias and variance are not only near-minimax among all CIs, but also near-optimal in the more optimistic setting where the baseline model is correctly specified.

In asymmetric settings, one can still choose a single estimator to trade off worst-case bias and variance and base a CI on this estimator. However, it will often make more sense in these settings to base inference on multiple estimators. We discuss this approach in Section 6.1. In settings where characterizing the asymptotic behavior of estimators in the expanded model is difficult, CIs based on test inversion have been proposed. We discuss this approach in Section 6.2.

6.1 Combining bounds on $T(\theta)$

The approach described in Section 5 uses bias-variance tradeoffs to arrive at a single estimator in the expanded model. In some settings, it will be sensible to separately estimate upper and lower bounds for $T(\theta)$. This is the case in Example 3.

Example 3 (continued)**.** In the expanded model, we observe (Y_i, W_i) where W_i is an indicator variable for Y_i^* being observed so that $Y_i = W_i \cdot Y_i^*$. Since Y_i^* is binary, we can obtain an upper and lower bound for the parameter of interest $p = E[Y_i^*]$ by replacing the missing values with 1 and 0 respectively: $\bar{Y}^U = \frac{1}{n}\sum_{i=1}^n [Y_i \cdot W_i + (1 - W_i) \cdot 1]$ and $\bar{Y}^L = \frac{1}{n}\sum_{i=1}^n [Y_i \cdot W_i + (1 - W_i) \cdot 0]$. An upper $100 \cdot (1-\alpha)\%$ CI for p takes the form $\bar{Y}^U + z_{1-\alpha}\,\text{se}(\bar{Y}^*)$ where $\text{se}(\bar{Y}^*)^2$ is the standard error for $\bar{Y}_U$. Similarly, one can form a lower CI from $\bar{Y}^L$. One can then form a two-sided CI by taking intersections of $100 \cdot (1 - \alpha/2)\%$ one-sided CIs, employing a correction along the lines of Imbens and Manski (2004) to avoid conservatism when sample error is small relative to the width of the identified set if one wishes.

In the above example, there is only one natural estimate of an upper and lower bound for the parameter of interest. In other cases, one may want to combine multiple estimators. Here, we describe an approach that Chernozhukov et al. (2013) refer to as an *intersection bounds* approach. Suppose that we have estimators $\hat{T}_1^U, \ldots, \hat{T}_m^U$ for which the asymptotic bias is known to be positive. For example, if we have a lower bound $\underline{\text{bias}}_k(\Theta)$ for the asymptotic bias of each estimator, we can subtract $\underline{\text{bias}}_k(\Theta)$ to obtain an upwardly biased estimator. Let $\hat{V}_1, \ldots, \hat{V}_k$ be variance estimates. An upper $100 \cdot (1-\alpha)\%$ CI can be obtained from any of these estimators as $(-\infty, \hat{T}_j + z_{1-\alpha}\sqrt{\hat{V}_k}]$. If one does not know a priori which of these CIs will provide the best bound, one can take the intersection of these CIs after applying a multiplicity correction, leading to the CI

$$\left(-\infty, \min_{1 \le j \le m} \left\{ \hat{T}_j + \text{cv}_\alpha \sqrt{\hat{V}_j} \right\} \right] \tag{9}$$

where cv_α is a multiplicity corrected critical value. For example, one can use the Bonferroni critical value $\text{cv}_\alpha = z_{1-\alpha/m}$, or one can use a less conservative critical value that takes into account the correlation between the estimates $\hat{T}_1, \ldots, \hat{T}_m$. A lower CI can be obtained analogously with downwardly biased estimators if such estimators are available, and a two-sided CI can be obtained by intersecting the one-sided CIs.

The same approach can be applied to form a CI based on the infimum of upper CIs over

estimators indexed by an infinite set, such as kernel estimators with different locations or bandwidths. The theory used to derive the critical value can be more involved (see, e.g., Chernozhukov et al., 2014), but the idea is the same.

Example 3 (continued)**.** In addition to the no-assumptions bounds discussed previously, Manski (1990) considered the use of additional variablees and assumptions to obtain tighter bounds on the distribution of Y_i^*. Suppose we have an *instrument* Z_i that is independent of Y_i^*, but which shifts the selection probability so that $P(W_i = 1|Z_i = z)$ is nonconstant. If Z_i is a discrete random variable taking on values $1, \ldots, m$, then, for each $j = 1, \ldots, m$ we can obtain an estimate $\hat{T}_j = \frac{\sum_{i:Z_i=j}(Y_iW_i+(1-W_i))}{\#\{i:Z_i=j\}}$ for $E[Y_i^*]$ that is upwardly biased using the sample where $Z_i = j$. These estimates can be used along with their standard errors to obtain an upper CI of the form (9).

This approach will not be feasible if Z_i is continuously distributed or takes on a large number of values relative to the sample size, since the estimates $\hat{T}_j$ will be too noisy. In such settings, one can still form an upwardly biased estimate by coarsening Z_i. For concreteness, suppose Z_i takes on scalar values. Then, for any $s \in \mathbb{R}$ and $t > 0$, the estimator $\hat{T}_{s,t} = \frac{\sum_{i:s<Z_i<s+t}(Y_iW_i+(1-W_i))}{\#\{i:s<Z_i\leq s+t\}}$ will be upwardly biased for $E[Y_i^*]$. Armstrong and Chan (2016) and Chetverikov (2017) consider CIs of the form (9) based on these estimators and related kernel estimators with the supremum taken over a set where s ranges over the whole real line and $t \geq t_n$ for $t_n \to 0$ at an appropriate rate. They show that such CIs are *adaptive*: they are valid under essentially no assumptions on the smoothness of $E[Y_iW_i + (1 - W_i)|Z_i = z]$, while having nearly the same excess length as an oracle CI that uses prior knowledge of the smoothness of $E[Y_iW_i + (1 - W_i)|Z_i = z]$.

The approach used in Armstrong and Chan (2016) and Chetverikov (2017) is related to ideas used in the literature on adaptive inference and adaptive testing using shape restrictions in nonparametric statistics (e.g. Dumbgen and Spokoiny, 2001). A general theory of adaptive inference in convex parameter spaces has been developed by Cai and Low (2004). We end this section with a discussion of adaptive inference under monotonicity in the context of our Example 2.

Example 2 (continued)**.** As discussed in Section 5.1, one can use the bound $|r(x)| \leq M(x)$ in the expanded model to form estimators that optimally trade off bias and variance, and to construct CIs based on these estimators. In the case where $\psi(X_i) = (1, X_i - x_0, \ldots, (X_i - x_0)^{p-1})$ and the bound $M(x) = (M/p!)|x-x_0|^p$ follows from a bound on Taylor approximation error, this amounts to using the bound M on the pth derivative to choose the bandwidth for a local polynomial estimator and to bound the bias of this estimator.

The goal of *adaptive inference* is to avoid the need for the user to specify M and p explicitly. Formally, one seeks a CI that (1) has coverage $1-\alpha$ regardless of M and p, or for a very conservative choice of M and p and (2) is nearly as short as a CI that uses knowledge of M and p to choose the estimator and bound its bias. Unfortunately, results going back to Low (1997) show that it is impossible to construct adaptive CIs in this setting: coverage under a particular choice of M or p requires that the length of the CI reflects this a priori choice even if it "turns out" that the regression function allows for a less conservative choice of M or p.

To avoid these impossibility results, one must impose additional restrictions. One possibility is a *shape restriction*, such as a monotonicity or concavity restriction on the regression function $E[Y_i|X_i = x]$. Consider the case where $p = 1$, so that the expanded model gives a bound on the first derivative of the regression function: $E[Y_i|X_i = x] = \beta + r(x)$ where $r(x) = M \cdot |x - x_0|$ and $\beta = E[Y_i|X_i = x_0]$. The local polynomial estimator described in Section 5.1 with $p = 1$ is the local constant or *Nadaraya-Watson* estimator $\hat{T} = \arg\min_b \sum_{i=1}^n K((X_i - x_0)/h_n)(Y_i - b)^2 = \frac{\sum_{i=1}^n Y_i K((X_i - x_0)/h_n)}{\sum_{i=1}^n K((X_i - x_0)/h_n)}$. This approach requires prior knowledge of the bound M on the first derivative of the regression function in order to choose h_n and bound the bias. As discussed above, impossibility results on adaptive inference (Low, 1997; Armstrong and Kolesár, 2018) show that prior knowledge of the bound M cannot be avoided. However, if one is willing to impose that the regression function $x \mapsto E[Y_i|X_i = x]$ is nondecreasing, one can salvage the possibility of adaptive inference. For a nonnegative kernel function $K(\cdot)$, let

$$\hat{T}_{U,h} = \frac{\sum_{i:X_i \geq x_0} Y_i K((X_i - x_0)/h_n)}{\sum_{i:X_i \geq x_0} K((X_i - x_0)/h_n)}, \quad \hat{T}_{L,h} = \frac{\sum_{i:X_i \leq x_0} Y_i K((X_i - x_0)/h_n)}{\sum_{i:X_i \leq x_0} K((X_i - x_0)/h_n)}.$$

Then $\hat{T}_{U,h}$ and $\hat{T}_{L,h}$ are respectively biased upward and downward for any h regardless of the bound M on the first derivative (or even if no such bound holds). One can then combine CIs based on these estimators for different bandwidths h or use data-driven choices of h to obtain adaptive CIs. See Dumbgen (2003), Cai et al. (2013) and Armstrong (2015) for examples of this approach.

6.2 Test inversion

In some settings, it may be difficult to find estimators $\hat{T}$ and characterize their bias and variance. One approach to inference in such settings is to base confidence sets on hypothesis

tests for the null $H_{\theta_0} : \theta = \theta_0$. Let $\hat{S}(\theta_0)$ be a test statistics and $\widehat{\text{cv}}(\theta_0)$ a critical value such that $P_\theta(\hat{S}(\theta) > \widehat{\text{cv}}(\theta)) < \alpha$ (or such that this inequality holds approximately in large samples). Then the set $\{T(\theta_0) : \hat{S}(\theta_0) > \widehat{\text{cv}}(\theta_0)\}$ is a $100 \cdot (1-\alpha)\%$ confidence set for $T(\theta)$.

Example 4 (continued)**.** The *moment inequalities* literature has considered models where a parameter β satisfies $Eg(W_i, \beta) \leq 0$ for a function $g()$ specified by the researcher. Here $g(W_i, \beta)$ is a $\mathbb{R}^p$ valued function and inequality is taken elementwise. This corresponds to the misspecified GMM example where $Eg(W_i, \beta) = c$, but instead of bounding the magnitude of c, we assume that $c \leq 0$ (note, however, that we can incorporate lower and upper bounds on a moment, say $a_1 \leq Em(W_i, \beta) \leq a_2$, by defining $g(W_i, \beta)$ to include the components $(a_1 - m(W_i, \beta), m(W_i, \beta) - a_2)$).

A common approach in this literature is to form a test statistic $\hat{S}(\beta_0)$ that depends on the vector of sample moments $\frac{1}{n}\sum_{i=1}^n g(W_i, \beta_0)$ and is increasing in each component of this vector. For example, one can take $S(\beta_0) = \max_{1 \leq j \leq p} \frac{1}{n}\sum_{i=1}^n g(W_i, \beta_0)$. One then obtains a CI for a parameter $T(\beta)$ by collecting the values of β that this test fails to reject: $\{T(\beta_0) : S(\beta_0) \leq \widehat{\text{cv}}(\beta_0)\}$. See Canay et al. (2023) for a recent review of this approach.

The CIs described in Section 5 take the form $\hat{T} \pm \hat{\chi}$, where $\hat{\chi}$ is a constant that depends on the standard error of $\hat{T}$ and a bound on its bias. These CIs are called *fixed length CIs*, since their length is fixed in an asymptotic setting where the variance of $\hat{T}$ can be treated as known. This feature makes it easy to assess the accuracy of the estimator and of inferences drawn from this CI.

In contrast, CIs based on test inversion do not in general provide a measure of accuracy based on their ex-post length. As an extreme example of this phenomenon, consider a CI that is either equal to $(-\infty, \infty)$ or a very small set depending on the outcome of a coin flip that yields heads with probability α. When reporting such CIs, one should ideally include a *statistical power analysis* showing that the tests used to form the CI are powerful enough to rule out relevant parameter values.

While CIs that are based on test inversion without reference to particular estimators can be difficult to analyze, ideas related to the bias-variance tradeoffs discussed in Section 5 often turn out to be relevant in settings involving inequalities or bounds. Armstrong (2014) derives rates of convergence of confidence sets in set identified models based on conditional moment inequalities (similar to the example described above, but using a vector of conditional moments: $E[g(W_i, \beta)|X_i = x] \geq 0$ all x). The derivations involve bias-variance tradeoffs and rates of convergence that are slower than root-n and depend on smoothness

conditions, similar to those in the nonparametric statistics literature discussed in the context of Example 2 above.

7 Choosing the expanded model

The conclusions of an analysis based on an expanded model will depend on how this expanded model is defined, including the relative magnitude of different sources of misspecification. This section discusses some proposals from the literature for defining the expanded model.

7.1 Specification tests and choice of the bound

Often, the researcher will be required to take a stance on a bound M on the magnitude of misspecification. Formally, the expanded model takes the form $\Theta(M)$ for some constant $M \geq 0$, with the baseline model obtained as the case where $M = 0$. This is the case in Example 1 (M is the bound on the gross error probability), Example 2 (M is the bound on the pth derivative used to obtain the Taylor approximation bound) and in some applications of Example 4 (one places a bound $\|\gamma\| \leq M$ for some norm $\|\cdot\|$).

Ideally, one would like to have a data-driven way of choosing M, perhaps based on specification tests or validation studies. The idea would be to choose M to be the smallest value such that the expanded model $\Theta(M)$ "fits the data" in some way. One way of formalizing this idea would be form a CI that is *adaptive* to M. Such a CI would be have nominal coverage over $\Theta(\infty)$ or over $\Theta(\overline{M})$ for some very conservative choice $\overline{M}$, while being more informative (i.e. substantially smaller) when M is smaller. Unfortunately, it can be shown formally that this goal is not possible in the settings we consider here (Low, 1997; Armstrong and Kolesár, 2018).

Even in settings where adaptive *inference* is not possible, it is often possible to find an adaptive *estimator*: that is, a single estimator $\hat{T}$ that is simultaneously near-minimax over $\Theta(M)$ for a wide range of values of M. In nonparametric estimation settings, *Lepski's method* (Lepskii, 1991) is an approach to choosing the bandwidths and related tuning parameters that can be used for adaptive estimation. Adaptive estimation in low dimensional settings such as the misspecified IV problem (Example 4) has been considered by Bickel (1984) and Armstrong et al. (2023).

Part of the reason for these impossibility results is that it is not possible to get a data-driven upper bound for the magnitude M of misspecification in these examples. In many

cases, however, one can get a data-driven *lower* bound using specification tests. One possibility is to report results for a range of values of M as a form of sensitivity analysis, along with a lower bound obtained from specification tests.

Example 4 (continued)**.** In the GMM model $Eg(W_i, \beta) = 0$, one can test the null that the data generating process is described by this baseline model using a test for overidentifying restrictions based on the minimized GMM objective function (Newey and McFadden, 1994, Section 9.5). One can generalize this approach to test the null hypothesis that the data generating process is described by the expanded model $Eg(W_i, \beta) = c$ with parameter space $\Theta(M) = \{(\beta', c')' : \|c\| \leq M\}$ where $\|\cdot\|$ is a norm such as the ℓ_p norm $\|c\| = \left(\sum_j |c_j|^p\right)^{1/p}$ (Armstrong and Kolesár, 2021b, Appendix B). One can then report the lower CI $[\underline{\hat{M}}, \infty)$ for M by taking $\underline{\hat{M}}$ to be the largest value of M such that the test fails to reject.

7.2 Data-driven bounds using auxiliary assumptions

One approach to obtaining a data-driven upper bound on the magnitude of misspecification is to introduce auxiliary assumptions. A common approach is to impose assumptions that rule out parameters θ that are only in $\Theta(M)$ when M is large, but which are difficult to distinguish from parameters that are in $\Theta(M)$ for smaller M. The basic idea can be described as follows: given parameter spaces $\Theta(M)$ indexed by $M \geq 0$, assume that $\theta \in \Theta(M) \cap \Theta_{\text{aux}}$ for some M where Θ_{aux} is a parameter space that introduces additional assumptions such that the hypotheses

$$H_{M_0} : \theta \in \Theta(M_0) \cap \Theta_{\text{aux}} \text{ vs } H_{M_1} : \theta \in \Theta(M_1) \cap \Theta_{\text{aux}} \tag{10}$$

are easy to distinguish when M_0 and M_1 are far apart, in the sense that a statistical hypothesis test for H_{M_0} with uniformly high power over H_{M_1} exists. Adding such auxiliary assumptions restores the possibility of obtaining a useful data-driven upper bound on M.

In the nonparametric regression setting (Example 2), a sizeable literature has used *self-similarity* conditions. These conditions, originally proposed by Giné and Nickl (2010), allow for the automatic choice of M and p by imposing conditions that rule out nonsmooth functions that are difficult to detect. Hoffmann and Nickl (2011) relate these conditions to hypothesis testing problems of the form (10) (here adaptation and the corresponding hypothesis testing problem are over the order p of the derivative as well as the bound M). While self-similarity conditions allow for adaptation to M and p, they introduce additional tuning parameters that are themselves subject to impossibility of adaptation results and

therefore must be specified by the researcher; see Armstrong (2021). Another approach to choosing the derivative bound M is to use a rule of thumb relating the smoothness bound M near the point where the regression function is being estimated to a global polynomial approximation. This approach is used by Armstrong and Kolesár (2020) to formally justify bandwidth selection rules based on global polynomial estimates similar to those proposed in Fan and Gijbels (1996, Chapter 4.2).

The literature on sensitivity analysis to IV exclusion restrictions (Example 4) has also introduced assumptions of this form. A direct approach (applied, for example, by Masten and Poirier (2021)) is to simply assume that the smallest value of M that one would fail to reject in the population is in fact a valid upper bound.

One approach that has gained recent popularity is to impose assumptions relating omitted variables bias from unobserved variables in OLS to omitted variables bias from observed variables. This idea has been used informally to justify the practice (suggested in an influential paper by Leamer, 1983) of concluding that a result is robust if there is little change in the OLS coefficient when additional covariates are added; however, writing down formal conditions and procedures can be tricky: see Altonji et al. (2005) and the follow up papers by Oster (2019), Masten and Poirier (2024) and Diegert et al. (2025).

7.3 Placebo tests

One approach to model validation is to estimate the effect of a policy change using a part of the data where no policy change occured. Such estimates are sometimes called *placebo* estimates. The idea is that, if the model specification is correct, the estimate should not be statistically different from zero.

Without further assumptions, this approach is subject to the same issues and formal impossibility results described above: the placebo estimates can give a lower bound on the magnitude M of misspecification, but not an upper bound. One way of formally justify placebo estimators is to assume that the location or part of the data set where the policy change occurs is randomized. This allows the placebo estimates to be used in a *Fisher randomization test* of the *sharp null* that the policy has no effect at all. Ganong and Jäger (2018) take this approach in the regression discontinuity setting discussed in Section 3 above (Ganong and Jäger (2018) focus on the related regression kink setting, but the approach applies to regression discontinuity as well). They propose to form placebo cutoffs $c_1, \ldots, c_J$ away from the actual cutoff c and to use the same methods as those used to form the original estimate $\hat{T}$ to form placebo estimates $\hat{T}_{\text{placebo},1}, \ldots, \hat{T}_{\text{placebo},J}$. One can then test the

null hypothesis of no treatment effect by comparing $|\hat{T}|$ to the magnitude of the placebo estimates $|\hat{T}_{\text{placebo},1}|, \ldots, |\hat{T}_{\text{placebo},J}|$ and making an assumption that the cutoff point c was randomly drawn with a given distribution. For example, if the distribution is uniform over the placebo points, one uses the $1-\alpha$ quantile of the estimates as a critical value.

The assumption that the discontinuity point is drawn randomly with a known distribution may be strong in some applications. On the other hand, the alternative of explicitly choosing the smoothness bound M or imposing auxiliary assumtions as discussed in Section 7.2 may not be palatable either.

7.4 Defining misspecification using statistical distances

In Example 1, we started with the baseline $N(\mu, 1)$ model and considered the expanded model where $Y_i \sim F$ with $d_{\text{g.e.}}(F; N(\mu, 1)) \leq M$ where $d_{\text{g.e.}}(F, N(\mu, 1))$ is the smallest value of $\tilde{M}$ such that we can write F as a mixture where one draws from a $N(\mu, 1)$ distribution with probability $\tilde{M}$ and from an arbitrary distribution H with probability $1 - \tilde{M}$. More generally, given a baseline model P_β with parameter space $\beta \in B$ and a function $d(P; Q)$ that measures the distance between probability distributions P and Q, one can define the expanded model with parameter space $\{(\beta, F) : \beta \in \beta, d(F, P_\beta) \leq M\}$ where the data follow the distribution $P_{\beta,F} = F$. References that take this approach were given when Example 1 was introduced in Section 2.3.

While the gross error assumption $d_{\text{g.e.}}(F; P_\beta) \leq M$ has a clear interpretation as a bound M on the probability of data contamination, the interpretation of expanded models based on other notions of distance can be less clear. One popular choice is the class of distances defined as $d_\phi(P; Q) = E_Q \phi(\frac{dP}{dQ})$ where $\frac{dP}{dQ}$ is the likelihood ratio and $\phi()$ is a convex function satisfying certain regularity conditions. Recent papers including Andrews et al. (2020); Bonhomme and Weidner (2022); Christensen and Connault (2023) have considered expanded models of this form, partly motivated by the work of Hansen and Sargent (2001). Taking this approach with the GMM model $Eg(W_i, \theta) = 0$ as the baseline model (Example 4), Andrews et al. (2020) show that this class of distances leads to an expanded model where $Eg(W_i, \theta) = c$ and the parameter space for c is approximated by the set $c'\Sigma^{-1}c \leq M$ when the bound M on the distance d_ϕ is small, where Σ is the variance matrix of the moment $g(W_i, \theta)$. Thus, the expanded model places a bound on the misspecification of each moment proportional to the variance of the moment. This may or may not correspond to bounds on failure of exogeneity restrictions based on economic intuition.

8 Discussion

In this review, we have covered papers from three seemingly distinct literatures: (1) misspecification and robust estimation (2) nonparametric statistics and (3) set identification. We have noted that all three literatures ultimately use an expanded parameter space to motivate and analyze statistical procedures such as estimators and CIs. Furthermore, in all three literatures, efficiency comparisons of estimators and CIs often boil down to trading off bias and variance. Nonetheless, while these literatures have drawn from each other at times, much of their historical development has been separate. This section discusses some of the historical connections between these literatures.

8.1 Set identification in robust estimation and nonparametrics

The econometrics literature has ultimately embraced the idea that standard statistical procedures and concepts can be usefully applied to set identified parameters without modifying the basic definition of these concepts (as argued in the context of CIs by Imbens and Manski (2004)). It is interesting to note that, while the results of Huber (1964) and Sacks and Ylvisaker (1978) allow for set identified parameters, both papers express hesitance about applying their results to set identified parameters.

In the contaminated normal model (Example 1), Huber (1964) notes in the first page of the article that the parameter of interest $T(\mu, F) = \mu$ (the mean of the normal distribution that is observed after data contamination) is "not uniquely determined" (i.e. not point identified) and states that this causes "some inconvenience." He then suggests that one can "remove this difficulty" by assuming that the contamination distribution is symmetric around μ (thereby making μ point identified) or by redefining the parameter of interest as one that is point identified (i.e. considering a pseudo-parameter). Despite this hesitance to consider statistical properties of estimators in set identified models, Huber (1964) ultimately does so in Section 7, where he derives bounds on the bias of estimators for $T(\mu, F) = \mu$ without imposing symmetry or other conditions to achieve point identification. Problems involving hypothesis testing and confidence bounds for the set identified parameter μ in this setting were also considered in later work (Huber, 1968).

Sacks and Ylvisaker (1978) introduce a condition (given by Equation (2.3) in their paper) that is sufficient for the parameters β in the approximately linear model to be point identified. They consider a fixed design setup (X_i's are nonrandom), so there is no formal role for a distribution of X_i, but the condition essentially says that β is identified if X_i is drawn from

a distribution with support $\mathcal{X}$ for some set $\mathcal{X}$. However, the authors later note that this condition is in fact not needed for their results (Remark 6, p. 1128). They state that the identification condition is used "to permit an unequivocal interpretation of the parameters and therefore of the estimates." They then note that one can often define $\mathcal{X}$ in a way such that identification holds even if the observed X_i's are far from such a set $\mathcal{X}$. However, this approach "may introduce fictitious treatments or locations and may be far from realistic so that interpretation of the parameters will remain elusive."

In later work, Knafl et al. (1982a,b) introduce confidence intervals for linear transformations of β in the approximately linear model. As with the minimax estimation results in Sacks and Ylvisaker (1978), the results on coverage and optimality of these CIs do not require point identification: the proposed CIs are valid for the set identified parameter in the sense of Imbens and Manski (2004). This subsequent work does not appear to contain any discussion of the issue of point vs set identification.

8.2 Nonparametrics and robust estimation

In the introductory paragraph of their paper, Sacks and Ylvisaker (1978) phrase their problem as one of robustness to misspecification of the regression function analogous the problems of robustness to distributional assumptions considered by Huber (1964). However, their results ultimately contributed to the literature on nonparametric estimation.[4] Other work by these authors (e.g. Sacks and Ylvisaker, 1981) applied these ideas to other nonparametric estimation problems such as density estimation, yielding optimality results for kernel based nonparametric procedures.

Nonparametric methods motivated by Sacks and Ylvisaker's ideas, such as local polynomial estimators, have become popular in the applied literature on the RD setting discussed in Section 3. Although this problem is not mentioned in their original papers, Sacks and Ylvisaker were originally motivated by the RD problem and discussions with the originator of the method, Donald Campbell; see Cook (2008) and Sacks and Ylvisaker (2012).

8.3 Recent developments in the econometrics literature

Problems where the parameter space involves bounds or inequalities have been of interest in the econometrics literature on moment inequalities and set identified models discussed above.

[4]Given that nonparametric methods motivated by Sacks and Ylvisaker's results have become popular in the RD literature, it

Such problems are often amenable to the use of ideas from the nonparametric statistics literature involving bias-variance tradoffs. As mentioned above, efficiency comparisons of CIs in conditional moment inequality models follow from such an approach (Armstrong, 2014). Rambachan and Roth (2023, Corollary 3.3 and Lemma A.8) use a result on convex testing from the nonparametric testing literature (Ingster and Suslina, 2003, Section 2.4.3) to derive the power envelope for inference on the parameter of interest in a class of set identified models. The same result was used by Armstrong and Kolesár (2018) to derive efficiency bounds in settings that include the approximately linear model in Example 2. (A related power envelope result from Romano et al. (2014, Theorem S.1) can also be derived as a special case of this convex testing result). Ideas developed in the literature on robust estimation problems such as Example 1 have also been fruitfully applied in the recent econometrics literature. Kaido and Zhang (2019) uses results on robust testing from Huber and Strassen (1973) to derive results on optimal tests in a class of incomplete models used in the literature on structural estimation of games.

The general theory of optimal estimation and inference in convex parameter spaces developed in Ibragimov and Khas'minskii (1985), Donoho (1994), Cai and Low (2004) and Armstrong and Kolesár (2018) applies not only to the nonparametric regression setting in Example 2 but also to many problems of interest in the recent econometrics literature involving bounds and set identification. Related ideas in the literatures on nonparametric testing and shape constrained inference have also proved useful in such settings (as in the papers by Armstrong and Chan (2016), Chetverikov (2017) and Armstrong (2014) mentioned above). Econometricians working on problems involving bounds and set identification should familiarize themselves with this previous work, given that it applies directly to many problems of this form.

9 Conclusion

In this article, we have reviewed an approach to misspecification in which the original misspecified model is embedded in an expanded model. The expanded model is intended to be "correctly specified" in the sense that it is defensible as providing an adequate description of reality. We have noted that this approach is commonly used not only in articles that explicitly frame their contribution in terms of misspecification, but also in the literature on nonparametric estimation and on set identified models. Indeed, other than papers that focus exclusively on pseudo-parameters (as defined in Section 2.5), most, if not all, papers in the

literature on misspecification consider expanded models of some form.

The expanded model approach to misspecification gives a concrete answer to the question of how to interpret statistical procedures under misspecification. To evaluate an estimator, CI or other procedure motivated by the original model, one simply examines its performance in the expanded model. On the other hand, the burden is still on the researcher to formalize the claim that the model is a useful approximation by proposing an expanded model and using it to evaluate the performance (e.g. size and power of tests) of the procedures used in the researcher's empirical analysis. As we discussed in Section 7, deciding on an expanded model and defending it can be a nontrivial task.

Let us conclude by turning back to the issues of misspecification in empirical practice discussed at the beginning of this article. Is it indeed the case that empirical researchers invoke models as an approximation without explaining what this means? How common is it for empirical researchers in economics to formally account for misspecification by adjusting their estimators and CIs based on an expanded model that accounts for misspecification in a convincing way? While empirical economics as a whole is certainly not perfect in this regard, some empirical papers are reasonably clear about using a clearly defined expanded parameter space to motivate estimators and CIs. The recommendation to use local polynomial estimators to account for Taylor approximation error, stemming from the analysis of Sacks and Ylvisaker (1978) and discussed in Section 5.1 above, has been widely adopted in the RD literature. While the formal analysis of Huber is rarely, if ever, invoked in applied work, the general idea that the median is insensitive to outliers has motivated the use of procedures such as quantile regression in empirical work.

My conjecture is that ad hoc procedures in empirical work that restrict attention to certain parts of a data set can in many cases be motivated by a formal analysis of misspecification using an expanded model. For example, event studies are often estimated using only a few time periods before and after the given event rather than a long time series. One might motivate this practice formally by writing down an expanded model (along the lines of Manski and Pepper (2018) or Rambachan and Roth (2023)) where the two way fixed effects model motivating the event study design holds only approximately, with approximation error that gets worse over longer time horizons. Formalizing these ad hoc approaches may be a useful direction for future research.

While applied researchers may be more adept at formally accounting for misspecification than theoretical econometricians give them credit for, there is clearly room for improvement for providing clear and convincing arguments that estimators and CIs used in empirical

work properly account for misspecification. Even in settings where recommendations from a formal analysis of an expanded model are used to motivate estimators in applied work, it is less common in applied work to make a reasoned argument that a particular expanded parameter space (including the bound M on the magnitude of misspecification that is needed for the conclusions of the analysis to hold) is an adequate description of the situation at hand. Formalizing some of the ideas regarding the use of specification tests and placebo estimates to bound the magnitude of misspecification (as discussed in Sections 7.2 and 7.3) may help with this task. How to choose the expanded parameter space and argue that a particular expanded parameter space is adequate in a given setting are important problems that may need to be tackled on a case by case basis. Progress on this front, perhaps through collaboration between theoretical econometricians and applied researchers, would improve the credibility of empirical research.

References

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," *Journal of Human Resources*, XL, 791–821.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 132, 1553–1592.

——— (2020): "On the Informativeness of Descriptive Statistics for Structural Estimates," *Econometrica*, 88, 2231–2258.

Angrist, J. D. (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66, 249–288, publisher: [Wiley, Econometric Society].

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

Armstrong, T. (2015): "Adaptive testing on a regression function at a point," *The Annals of Statistics*, 43, 2086–2101.

Armstrong, T., T. Kitagawa, and A. Tetenov (2025): "Statistical Decision Theory and Empirical Practice," .

Armstrong, T. B. (2014): "Weighted KS statistics for inference on conditional moment inequalities," *Journal of Econometrics*, 181, 92–116.

——— (2021): "Adaptation bounds for confidence bands under self-similarity," *Bernoulli*, 27, 1348–1370.

Armstrong, T. B. and H. P. Chan (2016): "Multiscale adaptive inference on conditional moment inequalities," *Journal of Econometrics*, 194, 24–43.

Armstrong, T. B., P. Kline, and L. Sun (2023): "Adapting to Misspecification," .

Armstrong, T. B. and M. Kolesár (2018): "Optimal Inference in a Class of Regression Models," *Econometrica*, 86, 655–683.

——— (2020): "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics*, 11, 1–39.

——— (2021a): "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness," *Econometrica*, 89, 1141–1177.

——— (2021b): "Sensitivity analysis using approximate moment condition models," *Quantitative Economics*, 12, 77–108–108.

Ben-Michael, E., A. Feller, D. A. Hirshberg, and J. R. Zubizarreta (2021): "The Balancing Act in Causal Inference," .

Berger, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis*, New York: Springer, 2nd ed. 1985. corr. 3rd printing 1993 edition ed.

Bickel, P. J. (1984): "Parametric Robustness: Small Biases can be Worthwhile," *The Annals of Statistics*, 12, 864–879, publisher: Institute of Mathematical Statistics.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022): "When is TSLS Actually LATE?" .

Bonhomme, S. and M. Weidner (2022): "Minimizing sensitivity to model misspecification," *Quantitative Economics*, 13, 907–954, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE1930.

Cai, T. T. and M. G. Low (2004): "An Adaptation Theory for Nonparametric Confidence Intervals," *The Annals of Statistics*, 32, 1805–1840.

Cai, T. T., M. G. Low, and Y. Xia (2013): "Adaptive confidence intervals for regression functions under shape constraints," *The Annals of Statistics*, 41, 722–750.

Canay, I. A., G. Illanes, and A. Velez (2023): "A User's guide for inference in models defined by moment inequalities," *Journal of Econometrics*, 105558.

Cheng, M.-Y., J. Fan, and J. S. Marron (1997): "On automatic boundary corrections," *The Annals of Statistics*, 25, 1691–1708.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014): "Gaussian approximation of suprema of empirical processes," *The Annals of Statistics*, 42, 1564–1597.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737.

Chetverikov, D. (2017): "Adaptive Tests of Conditional Moment Inequalities," *Econometric Theory*, 1–42.

Christensen, T. and B. Connault (2023): "Counterfactual Sensitivity and Robustness," *Econometrica*, 91, 263–298, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA17232.

Ciliberto, F. and E. Tamer (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77, 1791–1828.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2010): "Plausibly Exogenous," *Review of Economics and Statistics*, 94, 260–272.

Cook, T. D. (2008): ""Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics," *Journal of Econometrics*, 142, 636–654.

Dehejia, R. H. (2005): "Program evaluation as a decision problem," *Journal of Econometrics*, 125, 141–173.

Diegert, P., M. A. Masten, and A. Poirier (2025): "Assessing Omitted Variable Bias when the Controls are Endogenous," .

Donoho, D. L. (1994): "Statistical Estimation and Optimal Recovery," *The Annals of Statistics*, 22, 238–270.

Donoho, D. L. and R. C. Liu (1988): "The "Automatic" Robustness of Minimum Distance Functionals," *The Annals of Statistics*, 16, 552–586.

Dumbgen, L. (2003): "Optimal confidence bands for shape-restricted curves," *Bernoulli*, 9, 423–449.

Dumbgen, L. and V. G. Spokoiny (2001): "Multiscale Testing of Qualitative Hypotheses," *The Annals of Statistics*, 29, 124–152.

Fan, J. (1993): "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216.

Fan, J. and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, CRC Press.

Ganong, P. and S. Jäger (2018): "A Permutation Test for the Regression Kink Design," *Journal of the American Statistical Association*, 113, 494–504.

Gelman, A. and G. Imbens (2017): "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 1–10.

Gilboa, I. and M. Marinacci (2013): "Ambiguity and the Bayesian Paradigm," in *Advances in Economics and Econometrics: Volume 1, Economic Theory: Tenth World Congress*, Cambridge University Press, vol. 49, 179.

Giné, E. and R. Nickl (2010): "Confidence bands in density estimation," *The Annals of Statistics*, 38, 1122–1170.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2022): "Contamination Bias in Linear Regressions," .

Grama, I. and M. Nussbaum (2002): "Asymptotic equivalence for nonparametric regression," *HAL*, 2002.

Hansen, B. (2022): *Econometrics*, Princeton: Princeton University Press.

Hansen, L. and T. J. Sargent (2001): "Robust Control and Model Uncertainty," *American Economic Review*, 91, 60–66.

Hansen, L. P. and T. J. Sargent (2024): "Risk, ambiguity, and misspecification: Decision theory, robust control, and statistics," *Journal of Applied Econometrics*, 39, 969–999.

Hirano, K. and J. R. Porter (2020): "Chapter 4 - Asymptotic analysis of statistical decision rules in econometrics," in *Handbook of Econometrics*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, Elsevier, vol. 7 of *Handbook of Econometrics, Volume 7A*, 283–354.

Hirshberg, D. A., A. Maleki, and J. R. Zubizarreta (2021): "Minimax Linear Estimation of the Retargeted Mean," *arXiv:1901.10296 [math, stat]*.

Hoffmann, M. and R. Nickl (2011): "On adaptive inference and confidence bands," *The Annals of Statistics*, 39, 2383–2409.

Huber, P. J. (1964): "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101.

——— (1965): "A Robust Version of the Probability Ratio Test," *The Annals of Mathematical Statistics*, 36, 1753–1758.

——— (1968): "Robust confidence limits," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 10, 269–278.

——— (2004): *Robust Statistics*, John Wiley & Sons, google-Books-ID: e62RhdqIdMkC.

Huber, P. J. and V. Strassen (1973): "Minimax Tests and the Neyman-Pearson Lemma for Capacities," *The Annals of Statistics*, 1, 251–263.

Ibragimov, I. and R. Khas'minskii (1985): "On Nonparametric Estimation of the Value of a Linear Functional in Gaussian White Noise," *Theory of Probability & Its Applications*, 29, 18–32.

Imbens, G. and S. Wager (2019): "Optimized Regression Discontinuity Designs," *The Review of Economics and Statistics*, 101, 264–278.

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

Imbens, G. W. and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142, 615–635.

Imbens, G. W. and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

Imbens, G. W. and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Ingster, Y. and I. A. Suslina (2003): *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, Springer.

Kaido, H. and Y. Zhang (2019): "Robust Likelihood Ratio Tests for Incomplete Economic Models," *arXiv:1910.04610 [econ, math, stat]*.

Kallus, N. (2020): "Generalized optimal matching methods for causal inference." *J. Mach. Learn. Res.*, 21, 62–1.

Kitamura, Y., T. Otsu, and K. Evdokimov (2013): "Robustness, Infinitesimal Neighborhoods, and Moment Restrictions," *Econometrica*, 81, 1185–1201.

Knafl, G., J. Sacks, and D. Ylvisaker (1982a): "Model robust confidence intervals," *Journal of Statistical Planning and Inference*, 6, 319–334.

——— (1982b): "Model Robust Confidence Intervals II," in *Statistical Decision Theory and Related Topics III*, ed. by S. S. Gupta and J. O. Berger, Academic Press, vol. II, 87–102.

Kolesár, M. and C. Rothe (2018): "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review*, 108, 2277–2304.

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

Lehmann, E. L. and G. Casella (1998): *Theory of Point Estimation*, New York: Springer, 2nd edition ed.

Lehmann, E. L. and J. P. Romano (2005): *Testing statistical hypotheses*, Springer.

Lepskii, O. (1991): "On a Problem of Adaptive Estimation in Gaussian White Noise," *Theory of Probability & Its Applications*, 35, 454–466.

Low, M. G. (1997): "On nonparametric confidence intervals," *The Annals of Statistics*, 25, 2547–2554.

Manski, C. F. (1989): "Anatomy of the Selection Problem," *The Journal of Human Resources*, 24, 343–360.

——— (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323.

——— (2003): *Partial Identification of Probability Distributions*, Springer Series in Statistics, New York: Springer-Verlag.

——— (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246.

——— (2007): "Minimax-regret treatment choice with missing outcome data," *Journal of Econometrics*, 139, 105–115.

——— (2021): "Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald," *Econometrica*, 89, 2827–2853, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA17985.

Manski, C. F. and J. V. Pepper (2018): "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *The Review of Economics and Statistics*, 100, 232–244.

Masten, M. A. and A. Poirier (2021): "Salvaging Falsified Instrumental Variable Models," *Econometrica*, 89, 1449–1469.

——— (2024): "The Effect of Omitted Variables on the Sign of Regression Coefficients," .

Newey, W. K. (1985): "Generalized method of moments specification testing," *Journal of Econometrics*, 29, 229–256.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. F. E. a. D. L. McFadden, Elsevier, vol. 4, 2111–2245.

Noack, C. and C. Rothe (2024): "Bias-Aware Inference in Fuzzy Regression Discontinuity Designs," *Econometrica*, 92, 687–711.

Oster, E. (2019): "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37, 187–204, publisher: Taylor & Francis.

Rambachan, A. and J. Roth (2023): "A More Credible Approach to Parallel Trends," *The Review of Economic Studies*, rdad018.

Romano, J. P., A. M. Shaikh, and M. Wolf (2014): "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, 82, 1979–2002.

Sacks, J. and D. Ylvisaker (1978): "Linear Estimation for Approximately Linear Models," *The Annals of Statistics*, 6, 1122–1137.

——— (1981): "Asymptotically Optimum Kernels for Density Estimation at a Point," *The Annals of Statistics*, 9, 334–346.

——— (2012): "After 50+ Years in Statistics, An Exchange," *Statistical Science*, 27, 308–318.

Savage, L. J. (1954): *The Foundations of Statistics*, John Wiley & Sons.

Stoye, J. (2012): "New Perspectives on Statistical Decisions Under Ambiguity," *Annual Review of Economics*, 4, 257–282, _eprint: https://doi.org/10.1146/annurev-economics-080511-110959.

Tamer, E. (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge, UK ; New York, NY, USA: Cambridge University Press.

Wald, A. (1950): *Statistical decision functions.*, Wiley publications in statistics, New York: Wiley.

Wasserman, L. (2004): *All of Statistics: A Concise Course in Statistical Inference*, New York: Springer.

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

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

Yata, K. (2023): "Optimal Decision Rules Under Partial Identification," ArXiv:2111.04926 [econ, math, stat].

YITZHAKI, S. (1996): "On Using Linear Regressions in Welfare Economics," *Journal of Business & Economic Statistics*, 14, 478–486.