On the Informativeness of Descriptive Statistics for Structural Estimates

Isaiah Andrews, *Harvard University and NBER*[*]
Matthew Gentzkow, *Stanford University and NBER*
Jesse M. Shapiro, *Brown University and NBER*

October 2019

Abstract

We propose a way to formalize the relationship between descriptive analysis and structural estimation. A researcher reports an estimate $\hat{c}$ of a structural quantity of interest c that is valid under some model. The researcher also reports descriptive statistics $\hat{\gamma}$ that estimate features γ of the distribution of the data, and highlights the economic relationship between γ and c under the model. We compare the bound on the absolute bias of $\hat{c}$ across all models in a local neighborhood of the assumed model with the bound across a subset of these models under which the assumed relationship between γ and c is correct. Our main result shows that the ratio of these tight bounds depends only on a quantity we call the *informativeness* of $\hat{\gamma}$ for $\hat{c}$. Informativeness can be easily estimated even for complex models. We recommend that researchers report estimated informativeness alongside their descriptive analyses, and we illustrate with applications to three recent papers.

[*]E-mail: iandrews@fas.harvard.edu, gentzkow@stanford.edu, jesse_shapiro_1@brown.edu. We acknowledge funding from the National Science Foundation (DGE-1654234), the Brown University Population Studies and Training Center, the Stanford Institute for Economic Policy Research (SIEPR), the Alfred P. Sloan Foundation, and the Silverman (1968) Family Career Development Chair at MIT. We thank Tim Armstrong, Matias Cattaneo, Gary Chamberlain, Liran Einav, Nathan Hendren, Yuichi Kitamura, Adam McCloskey, Costas Meghir, Ariel Pakes, Ashesh Rambachan, Eric Renault, Jon Roth, Susanne Schennach, and participants at the Radcliffe Institute Conference on Statistics When the Model is Wrong, the Fisher-Schultz Lecture, the HBS Conference on Economic Models of Competition and Collusion, the University of Chicago Becker Applied Economics Workshop, the UCL Advances in Econometrics Conference, the Harvard-MIT IO Workshop, the BFI Conference on Robustness in Economics and Econometrics (especially discussant Jinyong Hahn), the Cornell Econometrics-IO Workshop, and the Johns Hopkins Applied Micro Workshop, for their comments and suggestions. We thank Nathan Hendren for assistance in working with his code and data. We thank our dedicated research assistants for their contributions to this project.

1 Introduction

Empirical researchers often present descriptive statistics alongside structural estimates that answer policy or counterfactual questions of interest. One leading case is where the structural model is estimated on data from a randomized experiment, and the descriptive statistics are treatment-control differences (e.g., Attanasio et al. 2012a; Duflo et al. 2012; Alatas et al. 2016). A second leading case is where the structural model is estimated on observational data, and the descriptive statistics are regression coefficients or correlations that capture important relationships (e.g., Gentzkow 2007a; Einav et al. 2013; Gentzkow et al. 2014; Morten 2019). Researchers often provide a heuristic argument that links the descriptive statistics to key structural estimates, sometimes framing this as an informal analysis of identification.[1]

Such descriptive analysis has the potential to make structural estimation more credible. Structural models are often criticized for lacking transparency, with large numbers of assumptions and a high level of complexity making it difficult for readers to evaluate how the results might change under plausible forms of misspecification (Heckman 2010; Angrist and Pischke 2010). If a particular result were mainly driven by some intuitive descriptive features of the data, a reader could focus on evaluating the assumptions that link those features to the result.

In this paper, we propose a way to make this logic precise. A researcher specifies an economic model that relates a scalar quantity of interest c to the distribution F of some data. The researcher computes an estimate $\hat{c}$ of c and a vector $\hat{\gamma}$ of descriptive statistics that estimate some features $\gamma(F)$ of the distribution F. The researcher asserts that γ is informative about c under the model (e.g., that γ partially or point identifies c), and argues that the economic assumptions linking the two are plausible. To what extent should accepting this argument increase a reader's confidence in $\hat{c}$?

We answer these questions focusing on the case where misspecification is local to the researcher's assumed model in an appropriate sense. To outline our proposal, let $\mathcal{F}^0(c)$ denote the set of distributions consistent with a given quantity c under the base model assumed by the researcher. This mapping captures all of the model's assumptions relevant to learning c. We assume that the sets $\mathcal{F}^0(c)$ and $\mathcal{F}^0(c')$ are disjoint for any $c \neq c'$, so the quantity c is identified under the base model, and that $\hat{c}$ is unbiased under the base model, either exactly or asymptotically.

To allow for misspecification, we assume that for any given c the true distribution could lie in a larger set $\mathcal{F}^N(c) \supseteq \mathcal{F}^0(c)$. Each $\tilde{F} \in \mathcal{F}^N(c)$ lies in a neighborhood $\mathcal{N}(F)$ of an F consistent with the base model, so that $\mathcal{F}^N(c) = \cup_{F \in \mathcal{F}^0(c)} \left\{ \tilde{F} \in \mathcal{N}(F) \right\}$. The sets $\mathcal{F}^N(c)$ and $\mathcal{F}^N(c')$ may

[1] See, for example, Fetter and Lockwood (2018, pp. 2200-2201), Spenkuch et al. (2018, pp. 1992-1993), and the examples discussed in Andrews et al. (2017).

overlap for $c \neq c'$, so the quantity c may not be identified once we allow for misspecification, and the estimator $\hat{c}$ may be biased. Let b^N denote the largest possible absolute bias in $\hat{c}$ that can arise under $\mathcal{F}^N(\cdot)$, where this bound may be infinite.

We wish to know how much the scope for bias is reduced if we accept the author's interpretation of the descriptive analysis. To operationalize this, we ask what happens when the true distribution lies in a restricted set of distributions that leave γ unchanged:

$$\mathcal{F}^R(c) = \cup_{F \in \mathcal{F}^0(c)} \left\{ \tilde{F} \in \mathcal{N}(F) : \gamma\left(\tilde{F}\right) = \gamma(F) \right\}.$$

This again requires that the true distribution $\tilde{F}$ lie in a neighborhood of some distribution F consistent with c under the base model, but adds the requirement that the descriptive features $\gamma\left(\tilde{F}\right)$ implied by $\tilde{F}$ match those implied by F.[2] Let b^R denote the largest absolute bias in $\hat{c}$ that can arise under $\mathcal{F}^R(\cdot)$. Because restricting the form of misspecification can only lessen the scope for bias, we know $b^R \leq b^N$.

We focus on characterizing the ratio b^R/b^N, which we take as a summary of the extent to which the structural estimate can borrow credibility from the descriptive analysis. We first provide an exact characterization of b^R/b^N in a linear model with normal errors. We then provide an approximate characterization of b^R/b^N in more general nonlinear models. We do this via a local asymptotic analysis, in which we study sequences of distributions local to a fixed data generating process, and define asymptotic analogues of b^R and b^N that correspond to the maximal asymptotic bias of $\hat{c}$ over the sets of sequences we consider. Throughout, we define the neighborhoods we consider so that they coincide, either exactly or asymptotically, with neighborhoods based on standard measures of statistical distance.

Our main results show that under these conditions the ratio b^R/b^N (or its asymptotic analogue) is equal to $\sqrt{1-\Delta}$, where Δ is a scalar which we call the *informativeness* of the descriptive statistics $\hat{\gamma}$ for the structural estimate $\hat{c}$. Informativeness is the R^2 from a regression of the structural estimate on the descriptive statistics when both are drawn from their joint (asymptotic) distribution. Intuitively, when informativeness is high, $\hat{\gamma}$ captures most of the variation in the data that determines $\hat{c}$, so knowing that the former is correctly described by the model significantly reduces the scope for bias in the latter. When informativeness is low, $\hat{c}$ is mainly determined by features of the data orthogonal to $\hat{\gamma}$, so confidence in the model's description of $\hat{\gamma}$ does not meaningfully

[2] Imposing that $\tilde{F} \in \mathcal{F}^R(c)$ is stronger than imposing that (i) $\tilde{F}$ lies in a neighborhood of some F consistent with c under the base model and (ii) $\gamma\left(\tilde{F}\right)$ is consistent with c under the base model. Imposing only (i) and (ii) means that

$$\tilde{F} \in \left(\mathcal{F}^N(c) \cap \left(\cup_{F^* \in \mathcal{F}^0(c)} \left\{ \tilde{F} : \gamma\left(\tilde{F}\right) = \gamma\left(F^*\right) \right\} \right) \right) \supseteq \mathcal{F}^R(c),$$

and so implies an absolute worst-case bias weakly greater than that under $\mathcal{F}^R(c)$.

reduce the scope for bias in $\hat{c}$. We propose informativeness as a way to formalize the colloquial notion of the extent to which $\hat{\gamma}$ "drives" $\hat{c}$.

Informativeness can be estimated at low cost even for computationally challenging models. We show that a consistent estimator of Δ can be obtained from manipulation of the estimated influence functions of $\hat{c}$ and $\hat{\gamma}$. In the large range of settings in which estimated influence functions are available from the calculations used to obtain $\hat{c}$ and $\hat{\gamma}$, the additional computation required to estimate Δ is trivial. We recommend that researchers report an estimate of informativeness whenever they present descriptive evidence as support for structural estimates.

As a concrete illustration, consider an example similar to Attanasio et al. (2012a), where a parametric structural model is estimated by maximum likelihood using individual-level data from a randomized experiment. Suppose the randomized treatment is an incentive offered to parents to keep their children in school. The descriptive statistics $\hat{\gamma}$ are treatment-control differences in school attendance for households randomized to different levels of the incentive. The quantity of interest c is the impact of a counterfactual policy that would change the design of the incentives. The researchers argue that the treatment-control differences are informative about c and that the economic assumptions underlying the extrapolation to the counterfactual policy are reasonable. The estimate $\hat{c}$ is related to $\hat{\gamma}$, but because it is estimated by maximum likelihood it may depend on other features of the data as well. Our measure of informativeness captures the extent to which $\hat{c}$ depends on $\hat{\gamma}$. If informativeness is close to one, a reader interested in worst-case bias can focus on evaluating the assumptions that govern $\hat{\gamma}$ and relate it to $\hat{c}$. If informativeness is far from one, the reader would want to focus on evaluating assumptions that govern features orthogonal to $\hat{\gamma}$.

Our results are related to Andrews et al. (2017). In that paper, we propose a measure Λ of the *sensitivity* of a parameter estimate $\hat{c}$ to a vector of statistics $\hat{\gamma}$, focusing on the case where $\hat{\gamma}$ are estimation moments that fully determine the estimator $\hat{c}$ (and so $\Delta = 1$). Here, we propose a complementary measure of the extent to which a vector of descriptive statistics $\hat{\gamma}$ determines the value of $\hat{c}$.[3] In an extension, we generalize our main result to accommodate the setting of Andrews et al. (2017) and so provide a unified treatment of sensitivity and informativeness.

We implement our proposal for three recent papers in economics, each of which reports or discusses descriptive statistics alongside structural estimates. In the first application, to Attanasio et al. (2012a), the quantity c of interest is the effect of a counterfactual redesign of the PROGRESA cash transfer program, and the descriptive statistics $\hat{\gamma}$ are sample treatment-control differences for different groups of children, similar to the example above. In the second application, to Gentzkow (2007a), the quantity c of interest is the effect of removing the online edition of the *Washington Post*

[3] Our work here draws on the analysis of "sensitivity to descriptive statistics" in Gentzkow and Shapiro (2015).

on readership of the print edition, and the descriptive statistics $\hat{\gamma}$ are linear regression coefficients. In the third application, to Hendren (2013a), the quantity c of interest is a key parameter governing the existence of insurance markets, and the descriptive statistics $\hat{\gamma}$ summarize the joint distribution of self-reported probabilities of loss events and the realizations of these events. Our choices of $\hat{\gamma}$ are guided by the authors' discussions of the relationship between their structural analysis and intuitive features of the data. In each case, we report an estimate of Δ for various combinations of $\hat{c}$ and $\hat{\gamma}$, and we discuss the implications for a reader's confidence in the conclusions. These applications illustrate how estimates of Δ can be presented and discussed in applied research.

Important limitations of our analysis include the use of asymptotic approximations to describe the behavior of estimators, and the use of a purely statistical notion of distance to define sets of alternative models. Ideally one would like to use exact finite-sample properties to characterize the credibility of estimators, and economic knowledge to define sets of alternative models. We are not aware of convenient procedures that achieve this ideal in the generality that we consider. We therefore propose the use of informativeness as a practical option to improve the precision of discussions of the connection between descriptive statistics and structural estimates in applied research.

In a related paper, Mukhin (2018) derives informativeness and sensitivity from a statistical-geometric perspective, and notes strong connections to semiparameteric efficiency theory. Mukhin also shows how to derive sensitivity and informativeness measures based on alternative metrics for the distance between distributions, and discusses the use of these measures for local counterfactual analysis.

Our work is also closely related to the large literature on local misspecification (e.g., Newey 1985; Conley et al. 2012; Andrews et al. 2017). Much of this literature focuses on testing and confidence set construction (e.g. Berkowitz et al. 2008; Guggenberger 2012; Armstrong and Kolesar, 2019) or robust estimation (e.g., Rieder 1994; Kitamura et al. 2013; Bonhomme and Weidner 2018). Rieder (1994) studies the choice of target parameters and proposes optimal robust testing and estimation procedures under forms of local misspecification including the one that we consider here. Bonhomme and Weidner (2018) derive minimax robust estimators and accompanying confidence intervals for economic parameters of interest under a form of local misspecification closely related to the one we study. Armstrong and Kolesar (2019) consider a class of ways in which the model may be locally misspecified that nests the one we consider, derive minimax optimal confidence sets, and show that there is limited scope to improve on their procedures by "estimating" the degree of misspecification, motivating a sensitivity analysis. In contrast to this literature, we focus on characterizing the relationship between a set of descriptive statistics and a given structural estimator, with the goal

of allowing consumers of research to sharpen their opinions about the reliability of the researcher's conclusions.

Our use of statistical distance to characterize the degree of misspecification relates to a number of recent papers. Our results cover the Cressie-Read (1984) family, which nests widely studied measures including the Kullback-Leibler divergence, Hellinger divergence, and many others, up to a monotone transformation. Kullback-Leibler divergence has been used to measure the degree of misspecification by, for example, Hansen and Sargent (2001), Hansen and Sargent (2005), Hansen et al. (2006), Hansen and Sargent (2016), and Bonhomme and Weidner (2018). Hellinger divergence has been used by, for example, Kitamura et al. (2013).

Finally, our work relates to discussions about the appropriate role of descriptive statistics in structural econometric analysis (e.g., Pakes 2014).[4] It is common in applied research to describe the data features that "primarily identify" structural parameters or "drive" estimates of those parameters.[5] As Keane (2010) and others have noted, such statements are not directly related to the formal notion of identification in econometrics. Their intended meaning is therefore up for grabs. If researchers are prepared to reinterpret these as statements linking correct specification of descriptive statistics to confidence in related structural estimates, then our approach provides a way to sharpen and quantify these statements at low cost to researchers.

2 Informativeness in a Linear Normal Model

Suppose we observe a vector $Y \in \mathbb{R}^k$, where a base model implies that

$$Y \sim N(X\eta, \Omega) \tag{1}$$

under $F(\eta)$, for $\eta \in \mathbb{R}^p$ an unknown parameter and X and Ω known, nonrandom matrices with full column rank. The scalar quantity of interest c is a linear function $c(\eta) = L'\eta$ of the model parameters, and the p_γ-dimensional vector of descriptive statistics $\hat{\gamma} = \Gamma' Y$ is a linear function of the data.

For any matrix M with $MX = I_p$, the estimators $\hat{\eta} = MY$ and $\hat{c} = L'\hat{\eta}$ are unbiased (e.g., $M = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}$ for maximum likelihood). For any such estimator, define $C' = L'M$.

[4] See also Dridi et al. (2007) and Nakamura and Steinsson (2018) for discussion of the appropriate choice of moments to match when fitting macroeconomic models.

[5] Andrews et al. (2017, footnotes 2 and 3) provide examples.

Under the model

$$\begin{pmatrix} \hat{c} \\ \hat{\gamma} \end{pmatrix} \sim N\left(\begin{pmatrix} L'\eta \\ \Gamma' X\eta \end{pmatrix}, \Sigma\right) \text{ for } \Sigma = \begin{pmatrix} \sigma_c^2 & \Sigma_{c\gamma} \\ \Sigma_{\gamma c} & \Sigma_{\gamma\gamma} \end{pmatrix} = \begin{pmatrix} C'\Omega C & C'\Omega\Gamma \\ \Gamma'\Omega C & \Gamma'\Omega\Gamma \end{pmatrix}.$$

We assume that $\sigma_c^2 > 0$, and that $\Sigma_{\gamma\gamma}$ has full rank.

We are concerned that the linear model (1) may be misspecified. To formalize this concern, we embed it in a larger nesting model. Elements of the nesting model are indexed by (η, ζ) for $\zeta \in \mathbb{R}^k$, where $Y \sim N(X\eta + \zeta, \Omega)$ under $F(\eta, \zeta)$. The parameter of interest is still $c = c(\eta)$, but the nesting model allows us to consider a wider range of (c, F) pairs than allowed by the base model. Without further restrictions this nesting model allows all (c, F) pairs and so imposes no specific relationship between the mean of Y and the quantity of interest c. In particular, the nesting model allows that the mean $E[Y_j]$ of the jth row of Y is nonlinear in X_j.

If Y follows distribution F allowed by the nesting model, then

$$\begin{pmatrix} \hat{c} \\ \hat{\gamma} \end{pmatrix} \sim N\left(\begin{pmatrix} C'E_F[Y] \\ \Gamma'E_F[Y] \end{pmatrix}, \Sigma\right). \tag{2}$$

Note that the mean $E_F[\hat{c}]$ of $\hat{c}$ need not correspond to the parameter of interest. Indeed, without further restrictions the parameter of interest is totally unidentified under the nesting model.

We proceed by restricting the deviation of the data generating process from the base model. As in the introduction, we let $\mathcal{F}^0(c)$ denote the set of distributions F consistent with quantity of interest c under the base model,

$$\mathcal{F}^0(c) = \{N(X\eta, \Omega) : \eta \in \mathbb{R}^p, c(\eta) = c\}.$$

That is, $\mathcal{F}^0(c)$ is the set of normal distributions with variance Ω and mean $X\eta$ for η such that $c(\eta) = L'\eta = c$.

Next, we consider neighborhoods $\mathcal{N}(F)$ of the form

$$\mathcal{N}(F) = \left\{N(X\eta + \zeta, \Omega) : \eta \in \mathbb{R}^p, F(\eta) = F, \zeta \in \mathbb{R}^k, \|\zeta\|_{\Omega^{-1}} \leq \mu\right\}, \tag{3}$$

for $\|V\|_A = \sqrt{V'AV}$ and $\mu > 0$ a known constant that indexes the degree of misspecification. Intuitively, $\mathcal{N}(F(\eta))$ is the set of distributions $N(\delta, \Omega)$ with $\|\delta - X\eta\|_{\Omega^{-1}} \leq \mu$. Such neighborhoods arise from restrictions on any Cressie-Read (1984) divergence, including Kullback-Leibler divergence and Hellinger distance. Such neighborhoods also arise if we define $\mathcal{N}(F)$ as the set of normal

distributions $\tilde{F}$ with variance Ω such that all size-α tests have power to distinguish F and $\tilde{F}$ bounded by a known constant.

Taking the union of such neighborhoods over $F \in \mathcal{F}^0(c)$ yields the expanded set

$$(4) \quad \mathcal{F}^N(c) = \bigcup_{F \in \mathcal{F}^0(c)} \left\{ \tilde{F} \in \mathcal{N}(F) \right\} = \left\{ N(X\eta + \zeta, \Omega) : \eta \in \mathbb{R}^p, \zeta \in \mathbb{R}^k, \|\zeta\|_{\Omega^{-1}} \leq \mu, c(\eta) = c \right\}.$$

The quantity of interest c is not in general point-identified under $\mathcal{F}^N(\cdot)$, since the sets $\mathcal{F}^N(c)$ and $\mathcal{F}^N(c')$ can overlap for $c \neq c'$, and the estimator $\hat{c}$ will in general be biased.

Finally, the subset $\mathcal{F}^R(c)$ of $\mathcal{F}^N(c)$ consistent with the restriction that the descriptive statistics $\hat{\gamma}$ are correctly described by the model is

$$\mathcal{F}^R(c) = \bigcup_{F \in \mathcal{F}^0(c)} \left\{ \tilde{F} \in \mathcal{N}(F) : \gamma\left(\tilde{F}\right) = \gamma(F) \right\}$$

$$= \left\{ N(X\eta + \zeta, \Omega) : \eta \in \mathbb{R}^p, \zeta \in \mathbb{R}^k, \|\zeta\|_{\Omega^{-1}} \leq \mu, \Gamma'\zeta = 0, c(\eta) = c \right\},$$

where $\gamma(F) = \Gamma' E_F[Y]$ is the mean of $\hat{\gamma}$ under F. Intuitively, $\mathcal{F}^R(\cdot)$ limits attention to forms of misspecification that don't change the distribution of $\hat{\gamma}$ and, consequently, preserves any relationship between c and $\hat{\gamma}$ present under the base model.

To quantify the extent of possible bias, let us define the worst-case bias under $\mathcal{F}^N(\cdot)$ as

$$b_N = \sup_c \sup_{F \in \mathcal{F}^N(c)} |E_F[\hat{c} - c]|,$$

and worst-case bias under $\mathcal{F}^R(\cdot)$ as

$$b_R = \sup_c \sup_{F \in \mathcal{F}^R(c)} |E_F[\hat{c} - c]|.$$

Our main result in this section characterizes the ratio of b_R to b_N. This ratio measures the extent to which the knowledge that $\hat{\gamma}$ is correctly described by the model reduces the scope for bias in $\hat{c}$, and is related to a quantity we term informativeness.

Definition. *The* **informativeness** *of $\hat{\gamma}$ for $\hat{c}$ is*

$$\Delta = \frac{\Sigma_{c\gamma}\Sigma_{\gamma\gamma}^{-1}\Sigma_{\gamma c}}{\sigma_c^2} \in [0, 1].$$

Informativeness is the R^2 from the population regression of $\hat{c}$ on $\hat{\gamma}$ under their joint distribution. Intuitively, it measures the extent to which variation in the descriptive statistics $\hat{\gamma}$ explains variation

in the structural estimator $\hat{c}$.

Proposition 1. *For any estimator $\hat{c}$ such that (2) holds, the set of possible biases under $\mathcal{F}^N(\cdot)$ is*

$$\left\{E_F[\hat{c}-c] : F \in \mathcal{F}^N(c)\right\} = [-\mu\sigma_c, \mu\sigma_c]$$

for any c, while the set of possible biases under $\mathcal{F}^R(\cdot)$ is

$$\left\{E_F[\hat{c}-c] : F \in \mathcal{F}^R(c)\right\} = \left[-\mu\sigma_c\sqrt{1-\Delta}, \mu\sigma_c\sqrt{1-\Delta}\right]$$

for any c. Hence,

$$\frac{b_R}{b_N} = \sqrt{1-\Delta}.$$

This result shows that the informativeness Δ governs the extent to which imposing restricted misspecification $F \in \mathcal{F}^R(c)$, rather than unrestricted misspecification $F \in \mathcal{F}^N(c)$, limits the worst-case bias. It also shows that the worst-case bias is the same for all c, and that any absolute bias smaller than the worst case is achievable.

We now sketch the proof of Proposition 1 in the special case where Σ has full rank. A complete proof is given in Appendix A.1.

Sketch of Proof Under all of the models we consider, the distribution of $(\hat{c}, \hat{\gamma})$ is normal with variance Σ and some mean $(\overline{c}, \overline{\gamma}) \in \mathbb{R}^{p_\gamma+1}$. The models differ in the restrictions they impose on the possible values of $(\overline{c}, \overline{\gamma})$.

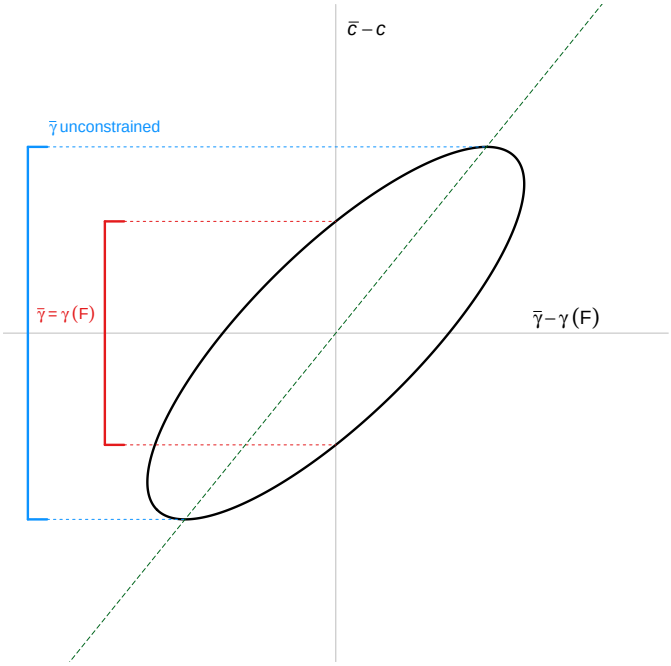
Say that the quantity of interest is $c \in \mathbb{R}$. Then under the base model we have that $F \in \mathcal{F}^0(c)$, and the set of possible values of $(\overline{c}, \overline{\gamma})$ consists of all those with $(\overline{c}-c, \overline{\gamma}-\gamma(\eta)) = (0,0)$ for some $\eta \in \mathbb{R}^p$ such that $c(\eta) = c$, where $\gamma(\eta) = \gamma(F(\eta))$.

Under the expanded model that allows misspecification we have that $F \in \mathcal{F}^N(c)$, and the set of possible values of $(\overline{c}, \overline{\gamma})$ expands to all those with $\|(\overline{c}-c, \overline{\gamma}-\gamma(\eta))\|_{\Sigma^{-1}} \leq \mu$ for some $\eta \in \mathbb{R}^p$ such that $c(\eta) = c$. Intuitively, this corresponds to taking a $\|\cdot\|_{\Sigma^{-1}}$-ball of radius μ around the set of means in the correctly specified case.

Finally, under the restricted model that imposes correct specification of $\hat{\gamma}$, in the sense that $F \in \mathcal{F}^R(c)$, the set of possible values of $(\overline{c}, \overline{\gamma})$ consists of all those with $\|(\overline{c}-c, 0)\|_{\Sigma^{-1}} \leq \mu$ for some $\eta \in \mathbb{R}^p$ such that $c(\eta) = c$ and $\gamma(\eta) = \overline{\gamma}$. The result then follows by calculating the range of values for $E_F[\hat{c}-c]$ implied by the appropriate sets.

Figure 1 shows the set of $(\overline{c}-c, \overline{\gamma}-\gamma(F))$ pairs consistent with a neighborhood $\mathcal{N}(F)$, $F \in \mathcal{F}^0(c)$, in the case where γ is a scalar. This set is an ellipsoid centered at zero. The range of

Figure 1: Possible values of $(\overline{c}-c, \overline{\gamma}-\gamma(F))$ under misspecification



Notes: The ellipsoid depicts the set of $(\overline{c}-c, \overline{\gamma}-\gamma(F))$ pairs consistent with a neighborhood $\mathcal{N}(F)$ of some $F \in \mathcal{F}^0(c)$ in the case where γ is a scalar. The interval labeled "$\overline{\gamma}$ unconstrained" characterizes the set of all values of the bias $\overline{c}-c$ consistent with $\mathcal{N}(F)$. The interval labeled "$\overline{\gamma}=\gamma(F)$" characterizes the set of all values of $\overline{c}-c$ consistent with $\left\{\tilde{F} \in \mathcal{N}(F) : \overline{\gamma}=\gamma(F)\right\}$. The diagonal line through the ellipsoid connects the points on the ellipsoid that achieve the minimum and maximum values of $\overline{c}-c$ consistent with $\mathcal{N}(F)$.

biases $\overline{c}-c$ consistent with $\mathcal{N}(F)$ is given by the light blue (outer) interval in the figure, which by Proposition 1 is equal to $[-\mu\sigma_c, \mu\sigma_c]$. The subset of these biases consistent with $\mathcal{N}(F)$ and $\bar{\gamma}=\gamma(F)$ is given by the red (inner) interval in the figure, which by Proposition 1 is equal to $\left[-\mu\sigma_c\sqrt{1-\Delta}, \mu\sigma_c\sqrt{1-\Delta}\right]$. As the figure makes clear, restricting the range of possible biases in $\overline{\gamma}$ reduces the scope for bias in $\overline{c}$. This is more true when informativeness Δ is high (corresponding to a tighter ellipse) and less true when Δ is low (corresponding to a wider ellipse).

2.1 Discussion

2.1.1 Relationship to Analysis of Identification

Our analysis is distinct from an analysis of identification. In particular, we focus on the behavior of a particular estimator under misspecification, taking as given that c is identified under the base

model. This is distinct from asking whether the identification of c is parametric or nonparametric, and from asking how the identified set changes under misspecification. To see the latter point, consider a case where $\hat{\gamma}$ is an unbiased estimator of c under the base model, but differs from $\hat{c}$.[6] An analysis of identification would conclude that c is point-identified under $\mathcal{F}^R(\cdot)$, whereas our analysis would conclude that there is potential for bias in the estimator $\hat{c}$ under $\mathcal{F}^R(\cdot)$.

We can connect our analysis to an analysis of identification if we consider identification from the distribution of $\hat{c}$ alone. In particular, Proposition 1 implies that if we assume only that $F \in \mathcal{F}^N(c)$, the quantity c is partially identified from the distribution of $\hat{c}$, with $b_N = \mu\sigma_c$ giving the radius of the identified set $[-\mu\sigma_c, \mu\sigma_c]$, whereas if we further assume that $F \in \mathcal{F}^R(c)$, the resulting identified set has radius $b_R = \mu\sigma_c\sqrt{1-\Delta}$. Under this interpretation, the ratio b_R/b_N measures how much the identified set shrinks when we impose that γ is correctly described by the base model.

Another connection to an analysis of identification is possible if we consider identification of c from γ. Under both $\mathcal{F}^0(\cdot)$ and $\mathcal{F}^R(\cdot)$, the identified set for c based on γ is

$$\mathcal{C}(\gamma) = \left\{c : \gamma(F) = \gamma \text{ for some } F \in \mathcal{F}^0(c)\right\}.$$

If $\mathcal{C}(\gamma)$ is a singleton, it may be possible to form a point estimate $\hat{c}_R$ of c, say by indirect inference, that depends on the data only through $\hat{\gamma}$. Such an estimate will typically be asymptotically unbiased under both $\mathcal{F}^0(\cdot)$ and $\mathcal{F}^R(\cdot)$, whereas other estimators $\hat{c}_0$ (e.g., the MLE under $\mathcal{F}^0(\cdot)$) that are unbiased under $\mathcal{F}^0(\cdot)$ may be biased under $\mathcal{F}^R(\cdot)$. In such cases, an alternative to reporting an estimate of Δ may be for researchers to report the estimate $\hat{c}_R$, either instead of or in addition to the estimate $\hat{c}_0$.

Basing estimation of c entirely on $\hat{\gamma}$ is not always a complete solution, however. The object of interest c may not be point-identified from γ alone, and the identified set $C(\gamma)$ may be large. Even when c is point-identified from γ, the estimate $\hat{c}_0$ may be much more precise than the estimate $\hat{c}_R$. In such a case, reporting an estimate of Δ alongside the estimate $\hat{c}_0$ may be useful in allowing readers to judge the scope for bias in the more precise estimate $\hat{c}_0$ under $\mathcal{F}^R(\cdot)$. This practice may be especially useful when readers are likely to differ in their confidence in $\mathcal{F}^0(\cdot)$ and $\mathcal{F}^R(\cdot)$.

2.1.2 Interpretation and Limitations

We pause here to discuss some other aspects and limitations of our approach.

First, our analysis focuses on bounding the absolute bias of the estimator $\hat{c}$. Since the variance of $\hat{c}$ is unaffected by misspecification, there is a one-to-one relationship between absolute bias and

[6] For instance, $\hat{\gamma}$ could be an estimator based on matching a statistically non-sufficient set of moments, while $\hat{c}$ might be the maximum likelihood estimator.

MSE. So, for fixed μ, Δ governs the extent to which restricting attention to $\mathcal{F}^R$ reduces the maximal MSE for $\hat{c}$. Unlike for absolute bias, however, the ratio of worst-case MSEs under $\mathcal{F}^N(\cdot)$ and $\mathcal{F}^R(\cdot)$ depends in general on μ.

Second, assuming $F \in \mathcal{F}^R(c)$ requires that the model-implied relationship between c and γ be correct local to each point in the base model. This is more restrictive than assuming that the true (c, γ) pair is consistent with the base model,

$$F \in \left(\mathcal{F}^N(c) \cap \left(\cup_{F^* \in \mathcal{F}^0(c)} \left\{\tilde{F} : \gamma\left(\tilde{F}\right) = \gamma\left(F^*\right)\right\}\right)\right). \tag{5}$$

Requiring $F \in \mathcal{F}^R$ can reduce worst-case bias even in settings where all (c, γ) pairs are possible under the base model. By contrast, in such settings (5) is equivalent to $F \in \mathcal{F}^N(c)$, and so does not reduce the scope for bias relative to $\mathcal{F}^N(c)$. More generally, the ratio of worst case bias under (5) to worst-case bias under $F \in \mathcal{F}^N(c)$ is bounded below by $\sqrt{1-\Delta}$.

Finally, we see the use of statistical distance to define the neighborhoods $\mathcal{N}(F)$ as a key potential limitation of our analysis. While defining neighborhoods in this way provides a practical default for many situations, it also means that the informativeness Δ depends on the sampling process that generates the data. To illustrate, suppose we are interested in estimating the average treatment effect c of some policy, that $\hat{c}$ is a treatment-control difference from an RCT, and that $\hat{\gamma}$ is the control group mean from the same RCT. If the control group is much larger than the treatment group, variability in $\hat{c}$ will primarily be driven by the treatment group, and the informativeness of $\hat{\gamma}$ for $\hat{c}$ will be low. If, on the other hand, the control group is much smaller than the treatment group, variability in $\hat{c}$ will primarily be driven by the control group, and the informativeness of $\hat{\gamma}$ for $\hat{c}$ will be high. Thus, the informativeness of the control group mean for the average treatment effect estimate in this setting depends on features of the experimental design, rather than solely on economic objects such as the distribution of potential outcomes.

2.2 Example

To fix ideas, suppose that a researcher observes i.i.d. data from a randomized evaluation of a conditional cash transfer program. Households are uniformly randomized between three subsidy levels, $s \in \{0, 2, 3\}$, receiving a payment of size s if their children attend school regularly. We think of those receiving $s = 0$ as the control group. The observed outcomes are average school attendance of children assigned subsidy s, which we denote Y_s. The key quantity of interest c is the mean attendance that would be obtained at a counterfactual subsidy level s^*.

The base model assumes that Y_s is normally distributed with variance ω^2 and a mean that

changes linearly with the subsidy, $E[Y_s] = \eta_1 + \eta_2 s$. Hence, under the base model $Y = (Y_0, Y_2, Y_3)'$ follows the normal model (1) with

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \ \Omega = \omega^2 I_3.$$

For a given distribution F, the vector $E_F[Y]$ has three elements, giving the expected values of Y_s under F at $s = 0$, 2,and 3, respectively.

Under the base model, the key quantity of interest (average attendance at subsidy s^*) is $c = L'\eta$ for $L = (1, s^*)'$. The researcher estimates c by maximum likelihood, which in this setting yields the ordinary least squares extrapolation:

$$\hat{c} = L' \left(X'X\right)^{-1} X'Y.$$

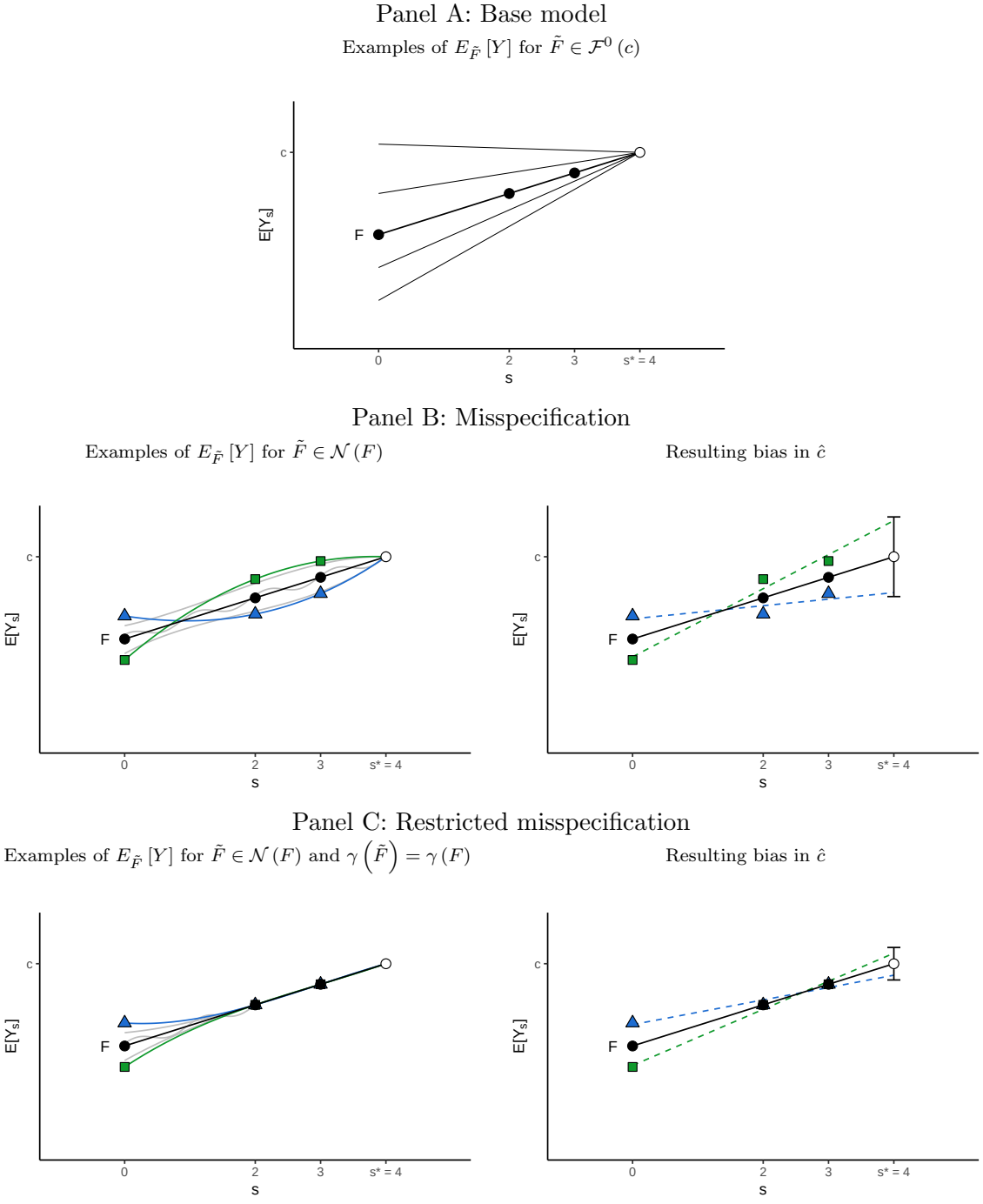
In the illustration that follows, we focus on the case where $\eta = (2, 0.2)$ and $s^* = 4$. This implies $c = 2 + 0.2 \cdot 4 = 2.8$. We also fix $\omega^2 = 0.1$.

We can identify each distribution F with its mean $E_F[Y]$ and plot the sets we consider. Panel A of Figure 2 shows elements of the set $\mathcal{F}^0(c)$ of distributions consistent with the base model for the given s^* and c, highlighting one particular element F. The set $\mathcal{F}^0(c)$ contains all linear functions passing through the point (s^*, c). We show these as continuous functions of s, but in the example only the values of Y at $s \in \{0, 2, 3\}$ are observed.

Readers of the researcher's findings may be concerned that the linear model of the response to subsidies is misspecified. The left-hand plot in Figure 2 Panel B illustrates elements of the expanded set $\mathcal{F}^N(c)$ of distributions within a $\mu = 1$ neighborhood of the highlighted element F. For given values of c and s^*, this set contains all functions (linear or nonlinear) that pass through the point (s^*, c) and whose vector of means $E_{\tilde{F}}[Y]$ satisfies $\left\|E_F[Y] - E_{\tilde{F}}[Y]\right\|_{\Omega^{-1}} \leq \mu$. The figure highlights two particular examples $\tilde{F}_1$ and $\tilde{F}_2$, with the elements of $E_{\tilde{F}}[Y]$ for these distributions plotted with blue triangles and green squares respectively.

The right-hand plot in Panel B of Figure 2 illustrates the bias that these forms of misspecification induce in the estimator $\hat{c}$. For each $\tilde{F}$, the estimator is the extrapolation to s^* of the best linear fit to the elements of $E_{\tilde{F}}[Y]$. The interval shown around the true value of c indicates the range of such biases that could be induced by all perturbations in the μ neighborhood of F. The bias bound b_N characterized in Proposition 1 corresponds to the supremum of the induced bias across c and $F \in \mathcal{F}^N(c)$.

Figure 2: Bias from misspecification of a linear normal model



Notes: The figure illustrates key sets of interest in the linear normal example described in Section 2.2 with $\eta = (2, 0.2)$, $\omega^2 = 0.1$, and $s^* = 4$. Panel A plots examples of $E_{\tilde{F}}[Y]$ for $\tilde{F}$ in the set $\mathcal{F}^0(c)$ of distributions consistent with the base model, and highlights $E_F[Y]$ for one particular element F. The left-hand plot in Panel B shows $E_{\tilde{F}}[Y_s]$ for elements of the expanded set of distributions $\mathcal{F}^N(c)$ under misspecification within a $\mu = 1$ neighborhood of the highlighted element F. The right-hand plot in Panel B shows the bias that these forms of misspecification induce in the estimator $\hat{c}$. The left-hand plot in Panel C shows $E_{\tilde{F}}[Y_s]$ for elements of the restricted set of distributions $\mathcal{F}^R(c)$ in a $\mu = 1$ neighborhood of the highlighted element F for which $\gamma(\tilde{F}) = \gamma(F)$ when $\hat{\gamma} = (Y_2, Y_3)$. The right-hand plot in Panel C shows the bias that these forms of misspecification induce in the estimator $\hat{c}$.

Suppose, now, that the researcher wishes to make more transparent the mapping from the data moments Y_0, Y_2, and Y_3 to the estimator $\hat{c}$, allowing readers to gauge the bias that might be induced by different violations of the identifying assumptions. The informativeness of $\hat{\gamma} = Y_0$ is $\Delta = 1/6 \approx 0.17$; the informativeness of $\hat{\gamma} = Y_2$ is $\Delta = 1/6 \approx 0.17$; and the informativeness of $\hat{\gamma} = Y_3$ is $\Delta = 2/3 \approx 0.67$. Intuitively, the extrapolation to $s^* = 4$ is particularly sensitive to the observed mean at $s = 3$ and substantially less sensitive to those at $s = 0$ and $s = 2$.

As a specific example of how these results inform concerns about bias, imagine that a reader is concerned that the effect of incentivizing school attendance might be discontinuous at zero, with average attendance for incentive s equal to[7]

$$E_F[Y_s] = \tilde{\eta}_0 + 1\{s > 0\}\tilde{\eta}_1 + s\tilde{\eta}_2. \tag{6}$$

This would change the relationship between the control group mean $E_F[Y_0]$ and the quantity of interest c, but would preserve the relationship between the treatment group means $(E_F[Y_2], E_F[Y_3])'$ and the quantity of interest c. Such a reader might be comforted to learn that, according to its informativeness, Y_0 plays a relatively small role in driving the estimator relative to Y_2 and Y_3. Noting that Δ for $\hat{\gamma} = (Y_2, Y_3)$ is $1/6 + 2/3 = 5/6 \approx 0.83$,[8] we can apply Proposition 1 to conclude that believing Y_2 and Y_3 are well described by the model—e.g., being willing to accept equation (6)—would reduce the maximal bias to $\sqrt{1 - 5/6} \approx 0.41$ times its unconstrained value. The resulting bias bound b_R for $\mu = 1$ is $1 \times 0.41 = 0.41$ standard errors.

Panel C of Figure 2 illustrates our analysis of this case, taking $\hat{\gamma} = (Y_2, Y_3)$. The left-hand plot shows elements of the restricted set $\mathcal{F}^R(c)$ of distributions in a $\mu = 1$ neighborhood of the highlighted element F for which $\gamma\left(\tilde{F}\right) = \gamma(F)$. The functions in this set must pass through the points (s^*, c), $(2, E_F(Y_2))$, and $(3, E_F(Y_3))$, but they can deviate from F at $s = 0$. The right-hand plot shows the interval of resulting biases. As the plot makes clear, allowing misspecification to affect Y_0 but not Y_2 or Y_3 substantially reduces the scope for bias relative to the case without this restriction.

We can consider how these results would change under alternative hypotheses. Entertaining misspecification in Y_2 rather than Y_0 and so setting $\hat{\gamma} = (Y_0, Y_3)$ would yield the same informativeness of $\Delta = 5/6 \approx 0.83$. Entertaining misspecification in Y_3 and setting $\hat{\gamma} = (Y_0, Y_2)$, however,

[7] To see how this possibility fits into the nesting model, let

$$\zeta = (\tilde{\eta}_0, \tilde{\eta}_0 + \tilde{\eta}_1 + 2\tilde{\eta}_2, \tilde{\eta}_0 + \tilde{\eta}_1 + 3\tilde{\eta}_2)' - X\eta.$$

[8] Because we have assumed that Y_0, Y_2, and Y_3 are statistically independent, the informativeness of any combination of these—e.g., $\hat{\gamma} = (Y_2, Y_3)$—is given by the sum of the informativeness of the individual elements. This does not hold in general.

would imply much lower informativeness of $\Delta = 1/6+1/6 = 1/3 \approx 0.33$, and thus much more scope for bias, with almost double the maximal bias, equal to $\sqrt{1-1/3} \approx 0.82$ times the unconstrained value.

As a final thought experiment, we can consider how the results would change if the researcher were to extrapolate further from the observed data, say to $s^* = 8$ rather than $s^* = 4$. Intuitively, this should make the slope across the three moments more important relative to their level, and so increase the informativeness of the control mean Y_0. Indeed, in this case, the informativeness of Y_3 falls to approximately 0.51 and the informativeness of Y_0 increases to approximately 0.42. Thus, the possibility of misspecification in the control group along the lines of equation (6) would be much more consequential in this case. Interestingly, the informativeness of Y_2 is quite low, at approximately 0.07, reflecting the fact that changing Y_2 while holding Y_0 and Y_3 constant has a limited impact on the estimated slope.

3 Informativeness Under Local Misspecification

This section translates our results on finite-sample bias in the normal model to results on asymptotic bias in nonlinear models with local misspecification. We first introduce our asymptotic setting, then state regularity conditions, prove our main result under local misspecification, develop intuition for the local misspecification neighborhoods we consider, and finally discuss a version of our analysis based on probability limits rather than asymptotic bias.

We assume that a researcher observes an i.i.d. sample $D_i \in \mathcal{D}$ for $i = 1, ..., n$. The researcher considers a base model which implies that $D_i \sim F(\eta)$, for $\eta \in H$ a potentially infinite-dimensional parameter. The implied joint distribution for the sample is $\times_{i=1}^{n} F(\eta)$. The key quantity of interest $c = c(\eta)$ is a scalar that may be an element of η, a counterfactual prediction, or some other function of the model's parameters. The researcher computes (i) a scalar estimate $\hat{c}$ of c and (ii) a $p_\gamma \times 1$ vector of descriptive statistics $\hat{\gamma}$.

To allow the possibility of misspecification, as in the last section we assume that the true data generating process may not correspond to the base model. Specifically, we embed the base model $\{F(\eta) : \eta \in H\}$ in a larger nesting model $\{F(\eta, \zeta) : \eta \in H, \zeta \in Z\}$, where $F(\eta, 0) = F(\eta)$ for all $\eta \in H$. The parameter of interest remains $c(\eta)$, and the nesting model parameter ζ again allows us to index a richer set of (c, F) pairs than permitted by the base model.

Under the nesting model, the joint distribution for the sample is $\times_{i=1}^{n} F(\eta, \zeta)$ for some $(\eta, \zeta) \in H \times Z$. We may have $F(\eta, \zeta) \notin \{F(\eta) : \eta \in H\}$, so the distribution is inconsistent with the base model, or $F(\eta, \zeta) = F(\eta')$ for $\eta' \neq \eta$, so the nesting model distribution "mimics" another value of

η under the base model. We allow the possibility that $c(\eta) \neq c(\eta')$ while $F(\eta, \zeta) = F(\eta')$, so that under misspecification the quantity of interest c cannot be expressed as a functional of F.

It is straightforward to form analogues of the sets $\mathcal{F}^0(c)$, $\mathcal{F}^N(c)$, and $\mathcal{F}^R(c)$ by collecting appropriate values of $(\eta, \zeta) \in H \times Z$. However, we are not aware of tractable expressions for the ratio b^R/b^N of finite-sample biases of common estimators $\hat{c}$ in the variety of settings that we wish to consider. We therefore approximate b^R/b^N by characterizing the first-order asymptotic bias of the estimator $\hat{c}$ under sequences of data generating processes in which (η, ζ) approaches a base value $(\eta_0, 0) \in H \times Z$ at a root-n rate.

Formally, define $\mathcal{H}$ and $\mathcal{Z}$ as sets of values such that for any $h \in \mathcal{H}$ and $z \in \mathcal{Z}$, we have $\eta_0 + th \in H$ and $tz \in Z$ for $t \in \mathbb{R}$ sufficiently close to zero.[9] For $F_{h,z}(t_h, t_z) = F(\eta_0 + t_h h, t_z z)$, we consider behavior under sequences of data generating processes

$$S(h, z) = \left\{ \times_{i=1}^{n} F_{h,z}\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right) \right\}_{n=1}^{\infty}.$$

The statement that (η, ζ) approaches $(\eta_0, 0)$ at a root-n rate should not be taken literally to imply that the data generating process depends on the sample size, but rather as an approach to approximating the finite-sample behavior of estimators in situations in which the influence of misspecification is on the same order as sampling uncertainty.[10] Section 3.4 instead considers fixed misspecification and develops results on probability limits rather than asymptotic bias.

Throughout our analysis, we state assumptions in terms of the base distribution $F_0 = F(\eta_0)$. If these assumptions hold for all $\eta_0 \in H$ then our local asymptotic approximations are valid local to any point in the base model, though many of the asymptotic quantities we consider will depend on the value of η_0. Section 5 discusses consistent estimators of these quantities that do not require a priori knowledge of η_0.

3.1 Regularity Conditions

We next discuss a set of regularity conditions used in our asymptotic results. Our first assumption requires that $\hat{c}$ and $\hat{\gamma}$ behave, asymptotically, like sample averages.

[9] For $\eta_0 + th \notin H$ or $tz \notin Z$ we may define distributions arbitrarily.

[10] The order $\frac{1}{\sqrt{n}}$ perturbation to the base-model parameter η is a common asymptotic device to analyze the local behavior of estimators (see for example Chapters 7-9 of van der Vaart, 1998). Setting the degree of misspecification proportional to $\frac{1}{\sqrt{n}}$ is likewise a common device for modeling local misspecification (see e.g. Newey (1985), Andrews et al. (2017), Armstrong and Kolesar (2019)).

Assumption 1. *Under* $S(0,0)$,

$$\sqrt{n}\left(\hat{c}-c_0,\hat{\gamma}-\gamma_0\right)=\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}\phi_c\left(D_i\right),\sum_{i=1}^{n}\phi_\gamma\left(D_i\right)\right)+o_p\left(1\right), \tag{7}$$

for functions $\phi_c(D_i)$ *and* $\phi_\gamma(D_i)$*, where* $E_{F_0}[\phi_c(D_i)]=0$*,* $E_{F_0}[\phi_\gamma(D_i)]=0$*, and for*

$$\Sigma=\begin{pmatrix}\sigma_c^2 & \Sigma_{c\gamma}\\ \Sigma_{\gamma c} & \Sigma_{\gamma\gamma}\end{pmatrix}=\begin{pmatrix}E_{F_0}\left[\phi_c\left(D_i\right)^2\right] & E_{F_0}\left[\phi_c\left(D_i\right)\phi_\gamma\left(D_i\right)'\right]\\ E_{F_0}\left[\phi_\gamma\left(D_i\right)\phi_c\left(D_i\right)\right] & E_{F_0}\left[\phi_\gamma\left(D_i\right)\phi_\gamma\left(D_i\right)'\right]\end{pmatrix},$$

Σ *is finite,* $\sigma_c^2>0$*, and* $\Sigma_{\gamma\gamma}$ *is positive-definite.*

The functions $\phi_c(D_i)$ and $\phi_\gamma(D_i)$ are called the influence functions for the estimators $\hat{c}$ and $\hat{\gamma}$, respectively. Asymptotic linearity of the form in (7) holds for a wide range of estimators (see e.g. Ichimura and Newey 2015), though it can fail for James-Stein, LASSO, and other shrinkage estimators (e.g. Hansen 2016). Asymptotic linearity immediately implies that $\hat{c}$ and $\hat{\gamma}$ are jointly asymptotically normal under $S(0,0)$.

We next strengthen asymptotic normality of $(\hat{c},\hat{\gamma})$ to hold local to η_0 under the base model. We impose the following.

Assumption 2. *Let* $\gamma(\eta)$ *denote the probability limit of* $\hat{\gamma}$ *under* $\times_{i=1}^{n}F(\eta)$*, and assume that for all* $h\in\mathcal{H}$*,* $\gamma(\eta_0+th)$ *exists for* t *sufficiently close to zero. For any* $h\in\mathcal{H}$*,* $c_n(h)=c\left(\eta_0+\frac{1}{\sqrt{n}}h\right)$*, and* $\gamma_n(h)=\gamma\left(\eta_0+\frac{1}{\sqrt{n}}h\right)$*, under* $S(h,0)$ *we have*

$$\sqrt{n}\begin{pmatrix}c_n(h)-c(\eta_0)\\ \gamma_n(h)-\gamma(\eta_0)\end{pmatrix}\to\begin{pmatrix}c^\star(h)\\ \gamma^\star(h)\end{pmatrix},$$

and moreover

$$\sqrt{n}\begin{pmatrix}\hat{c}-c(\eta_0)\\ \hat{\gamma}-\gamma(\eta_0)\end{pmatrix}\to_d N\left(\begin{pmatrix}c^\star(h)\\ \gamma^\star(h)\end{pmatrix},\Sigma\right).$$

The first part of Assumption 2 requires that $c_n(h)$ and $\gamma_n(h)$ be asymptotically well-behaved, in the sense that with appropriate recentering and scaling they converge to limits that can be written as functions of h.[11] Under this assumption, we can interpret $c^\star(h)$ as the local parameter of interest, playing the same role in our local asymptotic analysis as the parameter c does in the normal model.

[11] In conjunction with our other assumptions, Assumption 2 implies that $(c^\star(h),\gamma^\star(h))$ are linear in h.

The second part of Assumption 2 requires that $(\hat{c}, \hat{\gamma})$ be a regular estimator of $(c(\eta), \gamma(\eta))$ at η_0 under the base model (see e.g., Newey 1994), and is again satisfied under mild primitive conditions in a wide range of settings. In particular, this assumption requires that $\hat{c}$ be asymptotically unbiased for our local parameter of interest under $S(h, 0)$, in the sense that $\sqrt{n}(\hat{c} - c(\eta_0)) \to_d N\left(c^\star(h), \sigma_c^2\right)$.

We next assume the distributions $F(\eta, \zeta)$ have densities $f(d; \eta, \zeta)$ with respect to a common dominating measure ν. For $(t_h, t_z) \in \mathbb{R}^2$, if we consider the perturbed distributions $F_{h,z}(t_h, t_z)$ with densities $f_{h,z}(d; t_h, t_z)$ then the information matrix for (t_h, t_z), treating (h, z) as known, is

$$I(t_h, t_z) = E_{F_{h,z}(t_h,t_z)}\left[\begin{array}{cc} \left(\frac{\frac{\partial}{\partial t_h} f_{h,z}(D_i;t_h,t_z)}{f_{h,z}(D_i;t_h,t_z)}\right)^2 & \frac{\frac{\partial}{\partial t_h} f_{h,z}(D_i;t_h,t_z)}{f_{h,z}(D_i;t_h,t_z)} \frac{\frac{\partial}{\partial t_z} f_{h,z}(D_i;t_h,t_z)}{f_{h,z}(D_i;t_h,t_z)} \\ \frac{\frac{\partial}{\partial t_h} f_{h,z}(D_i;t_h,t_z)}{f_{h,z}(D_i;t_h,t_z)} \frac{\frac{\partial}{\partial t_z} f_{h,z}(D_i;t_h,t_z)}{f_{h,z}(D_i;t_h,t_z)} & \left(\frac{\frac{\partial}{\partial t_z} f_{h,z}(D_i;t_h,t_z)}{f_{h,z}(D_i;t_h,t_z)}\right)^2 \end{array}\right].$$

We consider the two-dimensional submodels obtained by fixing (h, z), $\left\{F_{h,z}(t_h, t_z) : (t_h, t_z) \in \mathbb{R}^2\right\}$, and impose a sufficient condition for these models to be differentiable in quadratic mean at zero.

Assumption 3. *For all $h \in \mathcal{H}$, $z \in \mathcal{Z}$, there exists an open neighborhood of zero such that for (t_h, t_z) in this neighborhood, (i) $\sqrt{f_{h,z}(d; t_h, t_z)}$ is continuously differentiable with respect to (t_h, t_z) for all $d \in \mathcal{D}$ and (ii) $I(t_h, t_z)$ is finite and continuous in (t_h, t_z).*

Assumption 3 imposes standard conditions used in deriving asymptotic results, and holds in a wide variety of settings; see Chapter 7.2 in van der Vaart (1998) for further discussion.

Finally, we require that the forms of misspecification we consider be sufficiently rich. To state this assumption, let us define $s_z(d) = \frac{\partial}{\partial t_z} \log(f_{h,z}(d; 0, 0))$ as the score function corresponding to z.

Assumption 4. *The set of score functions $s_z(\cdot)$ includes all those consistent with Assumption 3, in the sense that for any $s(\cdot)$ with $E_{F_0}[s(D_i)] = 0$ and $E_{F_0}\left[s(D_i)^2\right] < \infty$ there exists $z \in \mathcal{Z}$ with $E_{F_0}\left[(s(D_i) - s_z(D_i))^2\right] = 0$.*

Assumption 4 requires that the set of score functions $s_z(D_i)$ implied by $z \in \mathcal{Z}$ include all those consistent with Assumption 3.[12] Intuitively, this means that the set of nesting model distributions holding η fixed at η_0, $\{F(\eta_0, \zeta) : \zeta \in Z\}$, looks (locally) like the set of all distributions. Consequently, the nesting model allows forms of misspecification against which all specification tests that control size have trivial local asymptotic power.[13] If this assumption fails, the local asymptotic bias bounds we derive below continue to hold, but need not be sharp.

[12] That the score function $s_z(D_i)$ has mean zero and finite variance under Assumption 3 follows from Lemma 7.6 and Theorem 7.2 in van der Vaart (1998).

[13] In particular, for $s_h(d) = \frac{\partial}{\partial t_h} \log(f_{h,z}(d; 0, 0))$ the score function corresponding to h, provided $\{s_h(D_i) : h \in \mathcal{H}\}$ contains at least one nonzero element, say $s_{\tilde{h}}(D_i)$, Assumption 4 implies that there exists $z \in \mathcal{Z}$ such that

3.2 Main Result Under Local Misspecification

We can now derive the analogue of Proposition 1 in our local asymptotic framework. As a first step, we note that under our assumptions, $\sqrt{n}\left(\hat{c}-c\left(\eta_0\right),\hat{\gamma}-\gamma\left(\eta_0\right)\right)$ is asymptotically normal with variance Σ.

Lemma 1. *If Assumptions 1 and 3 hold, then under $S\left(h,z\right)$ for any $\left(h,z\right)\in\mathcal{H}\times\mathcal{Z}$,*

$$\sqrt{n}\left(\begin{array}{c}\hat{c}-c\left(\eta_0\right)\\ \hat{\gamma}-\gamma\left(\eta_0\right)\end{array}\right)\rightarrow_d N\left(\left(\begin{array}{c}\bar{c}\left(S\left(h,z\right)\right)\\ \bar{\gamma}\left(S\left(h,z\right)\right)\end{array}\right),\Sigma\right).$$

Recall that $c^{\star}\left(h\right)$ is the parameter of interest in our local asymptotic analysis. We can thus interpret $\bar{c}\left(S\left(h,z\right)\right)-c^{\star}\left(h\right)$ as the first-order asymptotic bias of $\hat{c}$ under $S\left(h,z\right)$, analogous to $E_F\left[\hat{c}-c\right]$ under the normal model.

As in the normal model we restrict the degree of misspecification. We first consider the case of correct specification. Let

$$\mathcal{S}^0\left(c^{\star}\right)=\left\{S\left(h,0\right):h\in\mathcal{H},c^{\star}\left(h\right)=c^{\star}\right\}$$

denote the set of sequences in the base model such that local parameter of interest takes value $c^{\star}$. Limiting attention to sequences $S\in\mathcal{S}^0\left(c^{\star}\right)$ imposes correct specification, and is analogous to limiting attention to $\mathcal{F}^0\left(c\right)$ in the normal model.

To relax the assumption of correct specification, next suppose we bound the degree of local misspecification by $\mu>0$. For $S\in\mathcal{S}^0\left(\cdot\right)=\cup_{c^{\star}}\mathcal{S}^0\left(c^{\star}\right)$, let us define the neighborhood

$$\mathcal{N}\left(S\right)=\left\{S\left(h,z\right):h\in\mathcal{H},S\left(h,0\right)=S,z\in\mathcal{Z},E_{F_0}\left[s_z\left(D_i\right)^2\right]^{\frac{1}{2}}\leq\mu\right\}.$$

This is a sequence-space analogue to the neighborhood $\mathcal{N}\left(F\right)$ considered in the normal model. We note in Section 3.3 that, analogous to the normal model, the bound μ on $E_{F_0}\left[s_z\left(D_i\right)^2\right]^{\frac{1}{2}}$ corresponds to a bound on the asymptotic difference between $S\left(h,z\right)$ and S as measured by any Cressie-Read divergence, and also to a bound on the power of any size-α test to distinguish $S\left(h,z\right)$ and S. Taking a union over $\mathcal{N}\left(S\right)$ for $S\in\mathcal{S}^0\left(c^{\star}\right)$ yields the expanded model

$$\mathcal{S}^N\left(c^{\star}\right)=\bigcup_{S\in\mathcal{S}^0\left(c^{\star}\right)}\left\{\tilde{S}\in\mathcal{N}\left(S\right)\right\},$$

$E_{F_0}\left[\left(s_{\tilde{h}}\left(D_i\right)-s_z\left(D_i\right)\right)^2\right]=0$. Arguments along the same lines as e.g. Chen and Santos (2018) then imply that $S\left(\tilde{h},0\right)$ and $S\left(0,z\right)$ are asymptotically indistinguishable, and thus that no specification test which controls size under $S\left(\tilde{h},0\right)$ has nontrivial power against $S\left(0,z\right)$.

which we can interpret as the sequence-space analogue of $\mathcal{F}^N(c)$ in the normal model.

Finally, to capture the restriction that $\hat{\gamma}$ is correctly described by the model, let us define a restricted set of sequences as

$$\mathcal{S}^R(c^\star) = \bigcup_{S \in \mathcal{S}^0(c^\star)} \left\{ \tilde{S} \in \mathcal{N}(S) : \bar{\gamma}\left(\tilde{S}\right) = \bar{\gamma}(S) \right\}.$$

Limiting attention to sequences $S \in \mathcal{S}^R(c^\star)$ is analogous to limiting attention to the set $\mathcal{F}^R(c)$ in the normal model, and restricts attention to forms of misspecification that do not affect the asymptotic behavior of the descriptive statistics $\hat{\gamma}$.

Let $b_N^\star$ and $b_R^\star$ denote the worst-case first–order asymptotic bias under $\mathcal{S}^N(\cdot)$ and $\mathcal{S}^R(\cdot)$, respectively:

$$b_N^\star = \sup_{c^\star} \sup_{S \in \mathcal{S}^N(c^\star)} |\bar{c}(S) - c^\star| \tag{8}$$

$$b_R^\star = \sup_{c^\star} \sup_{S \in \mathcal{S}^R(c^\star)} |\bar{c}(S) - c^\star|. \tag{9}$$

Our main result under local misspecification is analogous to Proposition 1 under the normal model.

Proposition 2. *Under Assumptions 1-4, the set of first-order asymptotic biases for $\hat{c}$ under $S \in \mathcal{S}^N(\cdot)$ is*

$$\left\{ \bar{c}(S) - c^\star : S \in \mathcal{S}^N(c^\star) \right\} = [-\mu\sigma_c, \mu\sigma_c],$$

for any $c^\star$ such that $\mathcal{S}^N(c^\star)$ is nonempty, while the set of first-order asymptotic biases under $S \in \mathcal{S}^R(\cdot)$ is

$$\left\{ \bar{c}(S) - c^\star : S \in \mathcal{S}^R(c^\star) \right\} = \left[-\mu\sigma_c\sqrt{1-\Delta}, \mu\sigma_c\sqrt{1-\Delta} \right],$$

for any $c^\star$ such that $\mathcal{S}^R(c^\star)$ is nonempty. Hence,

$$\frac{b_R^\star}{b_N^\star} = \sqrt{1-\Delta}.$$

To develop intuition for this result and highlight the similarity to the normal model, we again sketch the proof in the special case where Σ has full rank.

Sketch of Proof Under all of the sets of sequences we consider, $\sqrt{n}\left(\hat{c} - c(\eta_0), \hat{\gamma} - \gamma(\eta_0)\right)$ converges in distribution to a normal with variance Σ and mean $(\overline{c}, \overline{\gamma}) \in \mathbb{R}^{p_\gamma + 1}$. The sets of sequences differ in the restrictions they impose on the possible values of $(\overline{c}, \overline{\gamma})$.

If the quantity of interest is $c^\star \in \mathbb{R}$ then under correct specification we have that $S \in \mathcal{S}^0(c^\star)$, and the set of possible values of $(\overline{c}, \overline{\gamma})$ consists of all those that satisfy $(\overline{c} - c^\star, \overline{\gamma} - \gamma^\star(h)) = (0,0)$ for some $h \in \mathcal{H}$ such that $c^\star(h) = c^\star$.

Under the expanded set of sequences that allows misspecification we have that $S \in \mathcal{S}^N(c^\star)$, and the set of possible values of $(\overline{c}, \overline{\gamma})$ expands to all those that satisfy $\|(\overline{c} - c^\star, \overline{\gamma} - \gamma^\star(h))\|_{\Sigma^{-1}} \leq \mu$ for some $h \in \mathcal{H}$ such that $c^\star(h) = c^\star$. This again corresponds to taking a $\|\cdot\|_{\Sigma^{-1}}$-ball of radius μ around the set of means in the correctly specified case.

Finally, under the restricted set of sequences that imposes correct specification of $\hat{\gamma}$, we have that $S \in \mathcal{S}^R(c^\star)$, and the set of possible values of $(\overline{c}, \overline{\gamma})$ consists of all those that satisfy $\|(\overline{c} - c^\star, 0)\|_{\Sigma^{-1}} \leq \mu$ for some $h \in \mathcal{H}$ such that $c^\star(h) = c^\star$ and $\gamma^\star(h) = \overline{\gamma}$.

Observe that the sets of limiting distributions for $\sqrt{n}(\hat{c} - c(\eta_0), \hat{\gamma} - \gamma(\eta_0))$ have the same structure as the sets of finite-sample distributions for $(\hat{c}, \hat{\gamma})$ in the normal case. The result then follows by the same arguments used in the normal model.

3.3 Scaling of Perturbations

This subsection notes that the bound $E_{F_0}\left[s_z(D_i)^2\right] \leq \mu$ in the definition of $\mathcal{N}(S)$ can be interpreted both as a bound on the asymptotic power of tests to distinguish elements of $\mathcal{N}(S)$ from S, and, under regularity conditions, as a bound on the asymptotic Cressie-Read divergence of $F_{h,z}\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)$ from $F_{h,z}\left(\frac{1}{\sqrt{n}}, 0\right)$.

Asymptotic Distinguishability

A natural measure for the difference between $S(h,z)$ and $S(h,0)$ is the asymptotic power of tests to distinguish the two sequences.

Proposition 3. *Under Assumption 3, the most powerful level α test of the null hypothesis*

$$H_0 : (D_1, ..., D_n) \sim \times_{i=1}^n F_{h,z}\left(\frac{1}{\sqrt{n}}, 0\right)$$

against

$$H_1 : (D_1, ..., D_n) \sim \times_{i=1}^n F_{h,z}\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)$$

has power converging to $1 - F_{N(0,1)}\left(v_\alpha - E_{F_0}\left[s_z(D_i)^2\right]^{\frac{1}{2}}\right)$ *for* v_α *the* $1-\alpha$ *quantile of the standard normal distribution.*

The proof of Proposition 3 shows that the most powerful test corresponds asymptotically to a

z-test, where the z-statistic has mean $E_{F_0}\left[s_z(D_i)^2\right]^{\frac{1}{2}}$ under H_1. Hence, $\mu = 1$ corresponds to a one-standard-error shift in this asymptotic z-statistic.

The optimal test derived in the proof of Proposition 3 relies on knowledge of the sequence of base-model distributions $S(h,0)$. By contrast, in our local asymptotic setup the assumption of correct specification imposes only that $S \in \{S(h,0) : h \in \mathcal{H}\}$, while the assumption of correct specification along with value $c^\star$ for the local parameter of interest imposes $S \in \{S(h,0) : h \in \mathcal{H}, c^\star(h) = c^\star\}$. Testing either $\{S(h,0) : h \in \mathcal{H}\}$ or $\{S(h,0) : h \in \mathcal{H}, c^\star(h) = c^\star\}$ against $S(h',z)$ is more difficult than testing $S(h',0)$ against $S(h',z)$, so Proposition 3 gives an infeasible upper bound on the power of specification tests or tests for the parameter of interest together with correct model specification.

The infeasibility of the power bound in Proposition 3 highlights an important aspect of our local analysis. A possible justification for studying local misspecification (see, e.g., Huber and Ronchetti 2009, p. 294, as quoted in Bonhomme and Weidner 2018) is that specification tests eventually detect non-local misspecification with high probability, so conditional on non-rejection it is reasonable to focus on local misspecification. By contrast, we allow some forms of local misspecification z that are statistically undetectable absent knowledge of the local base-model parameter h, but nonetheless imply bias in $\hat{c}$.[14] The misspecification bound μ should therefore be interpreted as an a-priori restriction on the set of models we are willing to contemplate.

Asymptotic Divergence

Divergence measures provide an alternative quantification for the difference between $F_{h,z}\left(\frac{1}{\sqrt{n}}, 0\right)$ and $F_{h,z}\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)$, and hence for the degree of misspecification. We consider divergences of the form

$$r\left(F_{h,z}(t_h,0), F_{h,z}(t_h,t_z)\right) = E_{F_{h,z}(t_h,0)}\left[\psi\left(\frac{f_{h,z}(D_i;t_h,t_z)}{f_{h,z}(D_i;t_h,0)}\right)\right] \tag{10}$$

for $\psi(\cdot)$ a twice continuously differentiable function with $\psi(1) = 0$ and $\psi''(1) = 2$. A leading class of such divergences is the Cressie-Read (1984) family, which takes

$$\psi(x) = \frac{2}{\lambda(\lambda+1)}\left(x^{-\lambda} - 1\right).$$

Many well-known measures for the difference between distributions, including Kullback-Leibler divergence and Hellinger distance, can be expressed as monotone transformations of Cressie-Read

[14] In particular, while for some base models the data can reject the hypothesis that $E_{F_0}\left[s_z(D_i)^2\right]^{\frac{1}{2}} \leq \mu$, except in special cases the data cannot reject that $E_{F_0}\left[s_z(D_i)^2\right]^{\frac{1}{2}} > \mu$. See Armstrong and Kolesar (2019) for related discussion.

(1984) divergences for appropriate λ.

Appendix B.1 shows that under regularity conditions

$$\lim_{n\to\infty} n \cdot r\left(F_{h,z}\left(\frac{1}{\sqrt{n}},0\right),F_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)\right) = E_{F_0}\left[s_z\left(D_i\right)^2\right]. \tag{11}$$

Hence, Cressie-Read (1984) divergences yield the same asymptotic ranking over values of z, and therefore over sequences $S(h,z)$, as that implied by $E_{F_0}\left[s_z\left(D_i\right)^2\right]$.[15]

3.4 Non-Local Misspecification

To clarify the role of local misspecification in our results it is helpful to consider the analogue of Δ under non-local misspecification. Suppose now that the data follow $\times_{i=1}^{n} F$, where F does not change with the sample size, and we may again have $F \notin \{F(\eta) : \eta \in H\}$. Let us denote the probability limits of $\hat{c}$ and $\hat{\gamma}$ under F by $\tilde{c}(F)$ and $\tilde{\gamma}(F)$, respectively. We assume for ease of exposition that these probability limits exist.

To simplify the analysis, let us fix a value η_0 of the base model parameter, so the true value of the parameter of interest is $c(\eta_0)$, and again write $F_0 = F(\eta_0)$. Suppose that for a divergence r defined as in Section 3.3 we are willing to assume that $r(F_0, F) \leq \bar{r}$ for $\bar{r}$ a known scalar. Let $\mathcal{F}(\eta_0) = \{F(\eta_0, \zeta) : \zeta \in Z\}$ denote the set of distributions for the data under misspecification, again holding η_0 fixed. Given the bound $\bar{r}$ on the divergence, the probability limit of $|\hat{c} - c(\eta_0)|$ under $F \in \mathcal{F}(\eta_0)$ is no larger than

$$\tilde{b}_N(\bar{r}) = \sup\{|\tilde{c}(F) - c(\eta_0)| : F \in \mathcal{F}(\eta_0), r(F_0, F) \leq \bar{r}\}.$$

This is a non-local analogue of the bias bound $b_N^\star$, fixing $\eta = \eta_0$. We can likewise bound the probability limit of $|\hat{c} - c(\eta_0)|$ under forms of misspecification that do not affect the probability limit of $\hat{\gamma}$,

$$\tilde{b}_R(\bar{r}) = \sup\{|\tilde{c}(F) - c(\eta_0)| : F \in \mathcal{F}(\eta_0), r(F_0, F) \leq \bar{r}, \tilde{\gamma}(F) - \tilde{\gamma}(F_0) = 0\}.$$

This is a non-local analogue of bias bound $b_R^\star$, again fixing $\eta = \eta_0$.

Provided that $\tilde{b}_N(\bar{r})$ and $\tilde{b}_R(\bar{r})$ are both finite and non-zero, we can define a non-local analogue

[15] In equation (11) we scale by n to obtain a nontrivial limit, as the divergence between $F_{h,z}\left(\frac{1}{\sqrt{n}},0\right)$ and $F_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)$ tends to zero as $n \to \infty$.

$\tilde{\Delta}(\bar{r})$ of informativeness Δ by

$$\sqrt{1-\tilde{\Delta}(\bar{r})}=\frac{\tilde{b}_N(\bar{r})}{\tilde{b}_R(\bar{r})}.$$

Intuitively, $\tilde{\Delta}(\bar{r})$ measures the extent to which, fixing a value η_0 and a neighborhood size $\bar{r}$, limiting attention to forms of misspecification that do not affect $\hat{\gamma}$ limits the scope for inconsistency of the estimator $\hat{c}$. Observe that, whereas under our assumptions Δ does not depend on the local misspecification bound μ, $\tilde{\Delta}(\bar{r})$ will typically depend on $\bar{r}$.

Appendix B.2 shows that, under regularity conditions, an analogue of $\tilde{\Delta}(\bar{r})$ based on finite collections of ζ values converges to Δ as $\bar{r}\to 0$. This provides a sense in which Δ approximates $\tilde{\Delta}(\bar{r})$ when the degree of non-local misspecification is small.

4 Sensitivity and Informativeness

Proposition 2 considers the effect of limiting attention to forms of misspecification that do not affect $\hat{\gamma}$. In some cases, however, researchers may be interested in forms of misspecification with a non-zero, but known, effect on $\hat{\gamma}$. In such cases, our assumptions again imply a relationship between the biases in $\hat{c}$ and $\hat{\gamma}$.

This relationship depends on the sensitivity of $\hat{c}$ to $\hat{\gamma}$. This is the natural extension of the sensitivity measure proposed in Andrews et al. (2017) to the current setting.

Definition. *The* **sensitivity** *of $\hat{c}$ with respect to $\hat{\gamma}$ is*

$$\Lambda=\Sigma_{c\gamma}\Sigma_{\gamma\gamma}^{-1}.$$

To build intuition, note that sensitivity characterizes the relationship between $\hat{c}$ and $\hat{\gamma}$ in the asymptotic distribution under the base model. If we assume, as in Section 2, that $\hat{c}$ and $\hat{\gamma}$ are normally distributed in finite samples, then Λ is simply the vector of coefficients from the population regression of $\hat{c}$ on $\hat{\gamma}$. In this case, element Λ_j of Λ is the effect of changing the realization of a particular $\hat{\gamma}_j$ on the expected value of $\hat{c}$, holding the other elements of $\hat{\gamma}$ constant.

Andrews et al. (2017) show that for $\hat{c}=c(\hat{\eta})$, $\hat{\eta}$ a minimum distance estimator based on moments $\hat{g}(\eta)$, and $\hat{\gamma}=\hat{g}(\eta_0)$ the estimation moments evaluated at the true parameter value, under regularity conditions sensitivity translates the effect of misspecification on $\hat{\gamma}$ to the effect on $\hat{c}$, in the sense that

$$\bar{c}(S(h,z))-\bar{c}(S(h,0))=\Lambda\left(\bar{\gamma}(S(h,z))-\bar{\gamma}(S(h,0))\right).$$

Hence, Andrews et al. (2017) show that the asymptotic bias in $\hat{c}$ is equal to Λ times the asymptotic bias in $\hat{\gamma}$. Our next proposition extends this result.

Proposition 4. *Suppose that Assumptions 1-4 hold, and let*

$$\mathcal{S}^R(c^\star, \bar{\gamma}) = \cup_{S \in \mathcal{S}^0(c^\star)} \left\{ \tilde{S} \in \mathcal{N}(S) : \bar{\gamma}\left(\tilde{S}\right) - \bar{\gamma}(S) = \bar{\gamma} \right\}.$$

Provided $\mu(\bar{\gamma})^2 = \mu^2 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma} \geq 0$*, the set of possible biases under* $S \in \mathcal{S}^R(\cdot, \bar{\gamma})$ *is*

$$\left\{ \bar{c}(S) - c^\star : S \in \mathcal{S}^R(c^\star, \bar{\gamma}) \right\} = \left[\Lambda\bar{\gamma} - \mu(\bar{\gamma})\sigma_c\sqrt{1-\Delta}, \Lambda\bar{\gamma} + \mu(\bar{\gamma})\sigma_c\sqrt{1-\Delta} \right],$$

for any $c^\star$ *such that* $\mathcal{S}^R(c^\star, \bar{\gamma})$ *is nonempty.*

Proposition 4 extends the results of Andrews et al. (2017) to the case where $\hat{\gamma}$ need not be a vector of estimation moments, and thus we may have $\Delta < 1$. It likewise extends Proposition 2 to restricted misspecification with a non-zero effect on $\hat{\gamma}$. The resulting set of first-order asymptotic biases for $\hat{c}$ is centered at $\Lambda\bar{\gamma}$ with width proportional to $\sqrt{1-\Delta}$.

Unlike in Proposition 2, the degree of misspecification now enters the width through $\mu(\bar{\gamma}) = \sqrt{\mu^2 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}}$. Intuitively, $\mu(\bar{\gamma})$ measures the degree of excess misspecification beyond $\sqrt{\bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}}$, which is the minimum necessary to allow $\bar{\gamma}\left(\tilde{S}\right) - \bar{\gamma}(S) = \bar{\gamma}$. If the degree of excess misspecification is small then the first-order asymptotic bias of $\hat{c}$ is close to $\Lambda\bar{\gamma}$, while if the degree of excess misspecification is large then a wider range of biases is possible.

5 Implementation

In a wide range of applications, convenient estimates $\hat{\Sigma}$ of Σ are available from standard asymptotic results (e.g., Newey and McFadden 1994). Given such an estimate one can construct plug-in estimates

$$\hat{\Delta} = \frac{\hat{\Sigma}_{c\gamma}\hat{\Sigma}_{\gamma\gamma}^{-1}\hat{\Sigma}_{c\gamma}'}{\hat{\sigma}_c^2}, \quad \hat{\Lambda} = \hat{\Sigma}_{c\gamma}\hat{\Sigma}_{\gamma\gamma}^{-1}. \tag{12}$$

Provided $\hat{\Sigma}$ is consistent under $S(0,0)$, consistency of $\hat{\Sigma}$, $\hat{\Delta}$, and $\hat{\Lambda}$ under the sequences we study follows immediately under our maintained assumptions that $\sigma_c^2 > 0$ and $\Sigma_{\gamma\gamma}$ has full rank.

Assumption 5. $\hat{\Sigma} \xrightarrow{p} \Sigma$ *under* $S(0,0)$.

Proposition 5. *Under Assumptions 3 and 5,* $\hat{\Sigma} \xrightarrow{p} \Sigma$, $\hat{\Delta} \xrightarrow{p} \Delta$, *and* $\hat{\Lambda} \xrightarrow{p} \Lambda$ *under* $S(h,z)$ *for any* $h \in \mathcal{H}$, $z \in \mathcal{Z}$.

Mukhin (2018) provides alternative sufficient conditions for consistent estimation of sensitivity and informativeness. Mukhin (2018) also derives results applicable to GMM models with non-local misspecification.

5.1 Implementation with Minimum Distance Estimators

We have so far imposed only high-level assumptions (specifically Assumptions 1 and 5) on $\hat{c}$, $\hat{\gamma}$, and $\hat{\Sigma}$. While these high-level assumptions hold in a wide range of settings, minimum distance estimation is an important special case that encompasses a large number of applications. To facilitate application of our results, in this section we discuss estimation of Σ in cases where $c(\eta)$ can be written as a function of a finite-dimensional vector of parameters that are estimated by GMM or another minimum distance approach (Newey and McFadden 1994), and $\hat{\gamma}$ is likewise estimated via minimum distance.

Formally, suppose that we can decompose $\eta = (\theta, \omega)$ where θ is finite-dimensional and $c(\eta)$ depends on η only through θ. In a slight abuse of notation, we write $c(\eta) = c(\theta(\eta))$. We assume that $c(\theta)$ is continuously differentiable in θ.

The researcher forms an estimate $\hat{c} = c\left(\hat{\theta}\right)$ where $\hat{\theta}$ solves

$$\min_{\theta} \hat{g}(\theta)' \hat{W} \hat{g}(\theta) \tag{13}$$

for $\hat{g}(\theta)$ a k_g-dimensional vector of moments and $\hat{W}$ a $k_g \times k_g$-dimensional weighting matrix. The researcher likewise computes $\hat{\gamma}$ by solving

$$\min_{\gamma} \hat{m}(\gamma)' \hat{U} \hat{m}(\gamma), \tag{14}$$

for $\hat{m}(\gamma)$ a k_m-dimensional vector of moments and $\hat{U}$ a $k_m \times k_m$-dimensional weighting matrix.

Provided $\hat{W}$ and $\hat{U}$ converge in probability to limits W and U, while $\sqrt{n}\hat{g}(\theta(\eta_0))$ and $\sqrt{n}\hat{m}(\gamma(\eta_0))$ are jointly asymptotically normal under $S(0,0)$,

$$\begin{pmatrix} \hat{g}(\theta(\eta_0)) \\ \hat{m}(\gamma(\eta_0)) \end{pmatrix} \to_d N\left(0, \begin{pmatrix} \Sigma_{gg} & \Sigma_{gm} \\ \Sigma_{mg} & \Sigma_{mm} \end{pmatrix}\right),$$

existing results (see for example Theorem 3.2 in Newey and McFadden, 1994) imply that under $S(0,0)$ and standard regularity conditions,

$$\sqrt{n}\begin{pmatrix} c\left(\hat{\theta}\right) - c(\theta(\eta_0)) \\ \hat{\gamma} - \gamma(\eta_0) \end{pmatrix} \to_d N(0, \Sigma), \ \Sigma = \begin{pmatrix} \Lambda_{cg} & 0 \\ 0 & \Lambda_{\gamma m} \end{pmatrix} \begin{pmatrix} \Sigma_{gg} & \Sigma_{gm} \\ \Sigma_{mg} & \Sigma_{mm} \end{pmatrix} \begin{pmatrix} \Lambda_{cg} & 0 \\ 0 & \Lambda_{\gamma m} \end{pmatrix}'.$$

Here $\Lambda_{cg} = -C\left(G'WG\right)^{-1}G'W$ and $\Lambda_{\gamma m} = -\left(M'UM\right)^{-1}M'U$ are the sensitivities of $\hat{c}$ with respect to $\hat{g}\left(\theta\left(\eta_0\right)\right)$ and of $\hat{\gamma}$ with respect to $\hat{m}\left(\gamma\left(\eta_0\right)\right)$ as defined in Andrews et al. (2017), and $C = \frac{\partial}{\partial\theta}c\left(\theta\left(\eta_0\right)\right)$.

We can consistently estimate C by $\hat{C} = \frac{\partial}{\partial\theta}c\left(\hat{\theta}\right)$. If $\hat{g}\left(\theta\right)$ and $\hat{m}\left(\gamma\right)$ are continuously differentiable then under regularity conditions (see Theorem 4.3 in Newey and McFadden, 1994) we can likewise consistently estimate G by $\hat{G} = \frac{\partial}{\partial\theta}\hat{g}\left(\hat{\theta}\right)$ and M by $\hat{M} = \frac{\partial}{\partial\gamma}\hat{m}\left(\hat{\gamma}\right)$.[16] Hence, given consistent estimators $\hat{\Sigma}_{gg}$, $\hat{\Sigma}_{gm}$, and $\hat{\Sigma}_{mm}$ we can estimate Σ by

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Lambda}_{cg} & 0 \\ 0 & \hat{\Lambda}_{\gamma m} \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_{gg} & \hat{\Sigma}_{gm} \\ \hat{\Sigma}_{mg} & \hat{\Sigma}_{mm} \end{pmatrix} \begin{pmatrix} \hat{\Lambda}_{cg} & 0 \\ 0 & \hat{\Lambda}_{\gamma m} \end{pmatrix}'$$

for $\hat{\Lambda}_{cg} = -\hat{C}\left(\hat{G}'\hat{W}\hat{G}\right)^{-1}\hat{G}'\hat{W}$ and $\hat{\Lambda}_{\gamma m} = -\left(\hat{M}'\hat{U}\hat{M}\right)^{-1}\hat{M}'\hat{U}$.

What remains is to construct estimators $\left(\hat{\Sigma}_{gg}, \hat{\Sigma}_{gm}, \hat{\Sigma}_{mm}\right)$. When $\hat{\theta}$ and $\hat{\gamma}$ are GMM or ML estimators, we can write

$$\hat{g}\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n}\phi_g\left(D_i;\theta\right), \;\; \hat{m}\left(\gamma\right) = \frac{1}{n}\sum_{i=1}^{n}\phi_m\left(D_i;\gamma\right),$$

for $\left(\phi_g\left(D_i;\theta\right), \phi_m\left(D_i;\gamma\right)\right)$ the moment functions for GMM or the score functions for ML. We can then estimate Σ by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \left[\hat{\phi}_c\left(D_i\right)^2\right] & \left[\hat{\phi}_c\left(D_i\right)\hat{\phi}_\gamma\left(D_i\right)'\right] \\ \left[\hat{\phi}_\gamma\left(D_i\right)\hat{\phi}_c\left(D_i\right)\right] & \left[\hat{\phi}_\gamma\left(D_i\right)\hat{\phi}_\gamma\left(D_i\right)'\right] \end{pmatrix}, \tag{15}$$

for

$$\hat{\phi}_c\left(D_i\right) = \hat{\Lambda}_{cg}\phi_g\left(D_i;\hat{\theta}\right) = -\hat{C}\left(\hat{G}'\hat{W}\hat{G}\right)^{-1}\hat{G}'\hat{W}\phi_g\left(D_i;\hat{\theta}\right)$$

and

$$\hat{\phi}_\gamma\left(D_i\right) = \hat{\Lambda}_{\gamma m}\phi_m\left(D_i;\hat{\gamma}\right) = -\left(\hat{M}'\hat{U}\hat{M}\right)^{-1}\hat{M}'\hat{U}\phi_m\left(D_i;\hat{\gamma}\right).$$

In the GMM case, $\phi_g\left(D_i;\hat{\theta}\right)$ and $\phi_m\left(D_i;\hat{\gamma}\right)$ are available immediately from the computation of the final objective of the solver for (13) and (14), respectively. In the case of MLE, the score is likewise often computed as part of the numerical gradient for the likelihood. The elements of $\hat{\Lambda}_{cg}$ and $\hat{\Lambda}_{\gamma m}$ are likewise commonly precomputed. The weights $\hat{W}$ and $\hat{U}$ are directly involved in the

[16] If $\hat{g}\left(\theta\right)$ and $\hat{m}\left(\gamma\right)$ are not continuously differentiable, as sometimes occurs for simulation-based estimators, we can estimate G and M in other ways. For example, we can estimate the jth column of G by the finite difference $\left(\hat{g}\left(\theta + e_j\varepsilon_n\right) - \hat{g}\left(\theta - e_j\varepsilon_n\right)\right)/2\varepsilon_n$ for e_j the jth standard basis vector, where $\varepsilon_n \to 0$ and $\varepsilon_n\sqrt{n} \to \infty$ as $n \to \infty$. See Section 7.3 of Newey and McFadden 1994 for details on this approach and sufficient conditions for its validity.

calculation of the objectives in (13) and (14), respectively. When $\hat{g}(\theta)$ and $\hat{m}(\gamma)$ are differentiable, $\hat{G}$ and $\hat{M}$ are used in standard formulae for asymptotic inference on θ and γ, and the gradient $\hat{C}$ is used in delta-method calculations for asymptotic inference on c.[17]

In this sense, in many applications estimation of Σ will involve only manipulation of vectors and matrices already computed as part of estimation of, and inference on, the parameters θ, γ, and c.

Recipe. (GMM/MLE With Differentiable Moments)

1. Estimate $\hat{\theta}$ and $\hat{\gamma}$ following (13) and (14), respectively, and compute $\hat{c} = c\left(\hat{\theta}\right)$.
2. Collect $\left\{\phi_g\left(D_i;\hat{\theta}\right)\right\}_{i=1}^n$ and $\{\phi_m(D_i;\hat{\gamma})\}_{i=1}^n$ from the calculation of the objective functions in (13) and (14), respectively.
3. Collect the numerical gradients $\hat{G} = \frac{\partial}{\partial\theta}\hat{g}\left(\hat{\theta}\right)$, $\hat{M} = \frac{\partial}{\partial\gamma}\hat{m}(\hat{\gamma})$, and $\hat{C} = \frac{\partial}{\partial\theta}c\left(\hat{\theta}\right)$ from the calculation of asymptotic standard errors for $\hat{\theta}$, $\hat{\gamma}$, and $\hat{c}$.
4. Compute $\hat{\Lambda}_{cg} = -\hat{C}\left(\hat{G}'\hat{W}\hat{G}\right)^{-1}\hat{G}'\hat{W}$ and $\hat{\Lambda}_{\gamma m} = -\left(\hat{M}'\hat{U}\hat{M}\right)^{-1}\hat{M}'\hat{U}$ using the weights $\hat{W}$ and $\hat{U}$ from the objective functions in (13) and (14), respectively.
5. Compute $\hat{\phi}_c(D_i) = \hat{\Lambda}_{cg}\phi_g\left(D_i;\hat{\theta}\right)$ and $\hat{\phi}_\gamma(D_i) = \hat{\Lambda}_{\gamma m}\phi_m(D_i;\hat{\gamma})$ for each i.
6. Compute $\hat{\Sigma}$ as in (15).
7. Compute $\hat{\Delta}$ and $\hat{\Lambda}$ as in (12).

6 Applications

6.1 The Effects of PROGRESA

Attanasio et al. (2012a) use survey data from Mexico to study the effect of PROGRESA, a randomized social experiment involving a conditional cash transfer aimed in part at increasing persistence in school. The paper estimates a parametric model via maximum likelihood. The paper uses the estimated model to predict the effect of a counterfactual intervention in which total school enrollment is increased via a budget-neutral reallocation of program funds.

The estimate of interest $\hat{c}$ is the partial-equilibrium effect of the counterfactual rebudgeting on the school enrollment of eligible children, accumulated across age groups (Attanasio et al. 2012a,

[17] Note that in cases where the function $c(\theta)$ depends on features of the data beyond θ, for example on the distribution of covariates, our formulation implicitly treats those features as fixed at their sample values for the purposes of estimating Δ and Λ. Appendix B.3 discusses how to account for such additional dependence on the data, and presents corresponding calculations for some of our applications.

sum of ordinates for the line labeled "fixed wages" in Figure 2, minus sum of ordinates for the line labeled "fixed wages" in the left-hand panel of Figure 1).

Attanasio et al. (2012a) discuss the "exogeneous variability in [their] data that drives [their] results" as follows (p. 53):

> The comparison between treatment and control villages and between eligible and ineligible households within these villages can only identify the effect of the existence of the grant. However, the amount of the grant varies by the grade of the child. The fact that children of different ages attend the same grade offers a source of variation of the amount that can be used to identify the effect of the size of the grant. Given the demographic variables included in our model and given our treatment for initial conditions, this variation can be taken as exogenous. Moreover, the way that the grant amount changes with grade varies in a non-linear way, which also helps identify the effect.
>
> Thus, the effect of the grant is identified by comparing across treatment and control villages, by comparing across eligible and ineligible households (having controlled for being "non-poor"), and by comparing across different ages within and between grades. (p. 53)

Motivated by this discussion, we define three vectors $\hat{\gamma}$ of descriptive statistics, which correspond to sample treatment-control differences from the experimental data. The first vector ("impact on eligibles") consists of the age-grade-specific treatment-control differences for eligible children (interacting elements of Attanasio et al. 2012a, Table 2, single-age rows of the column labeled "Impact on Poor 97," with the child's grade). The second vector ("impact on ineligibles") consists of the age-grade-specific treatment-control differences for ineligible children (interacting elements of Attanasio et al. 2012a, Table 2, single-age rows of the column labeled "Impact on non-eligible," with the child's grade). The third vector consists of both of these groups of statistics.

We estimate the informativeness of each vector $\hat{\gamma}$ for the estimate $\hat{c}$ following the recipe in Section 5.1. Because model estimation is via maximum likelihood and $\hat{\gamma}$ can be represented as GMM, the recipe applies directly.

Table 1 reports the estimated informativeness of each vector of descriptive statistics. The estimated informativeness for the combined vector is 0.28. This is largely accounted for by the age-grade-specific treatment-control differences for eligible children.

Assuming that $\hat{\gamma}$ is correctly specified by the researchers' model reduces the worst-case bias in $\hat{c}$ by an estimated factor of $1 - \sqrt{1 - 0.28} \approx 0.15$ in the sense of Proposition 2. Further reduction

in the worst-case bias requires accepting aspects of the researchers' model that relate c to features of the data orthogonal to $\hat{\gamma}$.

To illustrate the distinction between informativeness and identification highlighted in Section 2.1.1, now let c be the partial-equilibrium effect of the actual program on the school enrollment of eligible children, accumulated across age groups. The parameter c is nonparametrically identified, and can be nonparametrically estimated by comparing the school enrollment of eligible children in treatment and control villages (as in Attanasio et al. 2012a, Table 2, column labeled "Impact on Poor 97"). The parameter c can also be estimated parametrically using the researcher's estimated model (as in Attanasio et al. 2012a, sum of ordinates for the line labeled "fixed wages" in the left-hand panel of Figure 1). The descriptive statistics $\hat{\gamma}$ have an informativeness of 1 for a natural nonparametric estimator, and an estimated informativeness of 0.31 for the parametric estimator, indicating that assumptions beyond those required for nonparametric identification are necessary to guarantee that the parametric estimator is unbiased in the sense of Proposition 2.

6.2 Newspaper Demand

Gentzkow (2007a) uses survey data from a cross-section of individuals to estimate demand for print and online newspapers in Washington DC. The paper estimates a parametric model via maximum likelihood. A central goal of Gentzkow's (2007a) paper is to estimate the extent to which online editions of papers crowd out readership of the associated print editions, which in turn depends on a key parameter governing the extent of print-online substitutability.

The estimate of interest $\hat{c}$ is the change in readership of the *Washington Post* print edition that would occur if the *Post* online edition were removed from the choice set (Gentzkow 2007a, Table 10, row labeled "Change in *Post* readership").

Gentzkow (2007a) discusses two features of the data that can help to distinguish correlated tastes from true substitutability: (i) a set of instruments—such as a measure of Internet access at work—that plausibly shift the utility of online papers but do not otherwise affect the utility of print papers; and (ii) a coarse form of panel data—separate measures of consumption in the last day and last five weekdays—that make it possible to relate changes in consumption of the print edition to changes in consumption of the online edition over time for the same individual (p. 730).

Motivated by Gentzkow's (2007a) discussion, we define three vectors $\hat{\gamma}$ of descriptive statistics. The first vector ("IV coefficient") is the coefficient from a 2SLS regression of last-five-weekday print readership on last-five-weekday online readership, instrumenting for the latter with the set of instruments (Gentzkow 2007a, Table 4, Column 2, first row). The second vector ("panel coefficient") is the coefficient from an OLS regression of last-one-day print readership on last-one-day online

readership controlling for the full set of interactions between indicators for print readership and indicators for online readership in the last five weekdays. Each of these regressions includes the standard set of demographic controls from Gentzkow (2007a, Table 5). The third vector $\hat{\gamma}$ consists of both the IV coefficient and the panel coefficient. Thus, the first two vectors have dimension 1, and the third has dimension 2.

We estimate the informativeness of each vector $\hat{\gamma}$ for the estimate $\hat{c}$ following the recipe in Section 5.1. Because model estimation is via maximum likelihood and $\hat{\gamma}$ can be represented as GMM, the recipe applies directly.

Table 2 reports the estimated informativeness of each vector of descriptive statistics. The estimated informativeness of the combined vector is 0.51. This is accounted for almost entirely by the panel coefficient, which alone has an estimated informativeness of 0.50. The IV coefficient, by contrast, has an estimated informativeness of only 0.01.

Gentzkow's (2007a) discussion of identification highlights both the exclusion restrictions underlying the IV coefficient and the panel variation underlying the panel coefficient as potential sources of identification, and if anything places more emphasis on the former. Based on Gentzkow's (2007a) discussion, and the large literature showing that exclusion restrictions can be used to establish nonparametric identification in closely related models (Matzkin 2007), it is tempting to conclude that accepting the assumptions linking the counterfactual c to the IV coefficient under Gentzkow's (2007a) model would greatly limit the scope for bias in Gentzkow's (2007a) estimator $\hat{c}$.

Our findings suggest otherwise. Assuming that the IV coefficient is correctly specified by the researcher's model reduces the worst-case bias in $\hat{c}$ by an estimated factor of only $1 - \sqrt{1 - 0.01} < 0.01$ in the sense of Proposition 2. By contrast, assuming that the panel coefficient is correctly specified by the researcher's model reduces the worst-case bias in $\hat{c}$ by an estimated factor of $1 - \sqrt{1 - 0.50} \approx 0.29$. A reader interested in evaluating the scope for bias in the researcher's estimator may therefore wish to focus more attention on assumptions related to the panel coefficient, such as those concerning the time structure of preference shocks, than on assumptions related to the IV coefficient, such as exclusion restrictions.

6.3 Long-term Care Insurance

Hendren (2013a) uses data on insurance eligibility and self-reported beliefs about the likelihood of different types of "loss" events (e.g., becoming disabled) to recover the distribution of underlying beliefs and rationalize why some groups are routinely denied insurance coverage. The paper estimates a parametric model via maximum likelihood. We focus here on Hendren's (2013a) model of

the market for long-term care (LTC) insurance.

The estimate of interest $\hat{c}$ is the *minimum pooled price ratio* among rejectees (Hendren 2013a, Table V, row labeled "Reject," column labeled "LTC"). The minimum pooled price ratio determines the range of preferences for which insurance markets cannot exist (Hendren 2013a, Corollary 2 to Theorem 1). This ratio is a key output of the analysis, as it provides an economic rationale for the insurance denials that are the paper's focus.

Hendren (2013a) explains that the parameters that determine the minimum pooled price ratio are identified from the relationship between elicited beliefs and the eventual realization of loss events such as long term care (pp. 1751-2).

Motivated by Hendren's (2013a) discussion, we define four vectors $\hat{\gamma}$ of descriptive statistics. The first vector ("fractions in focal-point groups") consists of the fraction of respondents who report exactly 0, the fraction who report exactly 0.5, and the fraction who report exactly 1. The second vector ("fractions in non-focal-point groups") consists of the fractions of respondents whose reports are in each of the intervals (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5), (0.5, 0.6], (0.6, 0.7], (0.7, 0.8], (0.8, 0.9], and (0.9, 1). The third vector ("fraction in each group needing LTC") consists of the fraction of respondents giving each of the preceding reports who eventually need long-term care. The fourth vector $\hat{\gamma}$ consists of all three of the other vectors.

Hendren's (2013a) discussion suggests that the third vector will be especially informative for the minimum pooled price ratio.

We estimate the informativeness of each vector $\hat{\gamma}$ for the estimate $\hat{c}$ following the recipe in Section 5.1. Because model estimation is via maximum likelihood and $\hat{\gamma}$ can be represented as GMM, the recipe applies directly.

Table 3 reports the estimated informativeness of each vector of descriptive statistics. The estimated informativeness of the combined vector is 0.70. The estimated informativeness is 0.01 with respect to the fractions in focal point groups, 0.02 with respect to the fractions in non-focal-point groups, and 0.68 with respect to the fraction in each group needing LTC. Assuming that the fractions in each group needing LTC are correctly specified by the researcher's model reduces the worst-case bias in $\hat{c}$ by an estimated factor of $1 - \sqrt{1 - 0.68} \approx 0.43$. This finding seems consistent with the author's discussion.

7 Conclusions

Descriptive analysis has become an important complement to structural estimation. Researchers often highlight key descriptive features that they argue will play an important role in driving their

structural conclusions. A reader who finds the economic assumptions linking the reduced-form moments to the key structural quantities credible might then be more confident in the researcher's conclusions.

We propose one way to make this informal argument precise. We define a measure Δ of the informativeness of descriptive statistics $\hat{\gamma}$ for a structural estimate $\hat{c}$. Informativeness captures the share of variation in $\hat{c}$ that is explained by $\hat{\gamma}$ under their joint asymptotic distribution. We show that, under some conditions, informativeness also governs the extent to which knowing the model correctly describes the descriptive statistics (and hence their relationship to the structural quantity of interest) limits the scope for bias in the structural estimates. In this sense, descriptive analysis based on statistics with high informativeness can indeed increase confidence in structural estimates.

Informativeness can be computed at negligible cost even for complex models, and we provide a convenient recipe for computing it. We show in the context of our applications that reporting informativeness can sharpen the interpretation of descriptive analysis in important economic settings. We recommend that researchers report estimated informativeness alongside their descriptive analyses.

References

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, Ririn Purnamasari, and Matthew Wai-Poi. 2016. Self-targeting: Evidence from a field experiment in Indonesia. *Journal of Political Economy* 124(2): 371–427.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro. 2017. Measuring the sensitivity of parameter estimates to estimation moments. *Quarterly Journal of Economics* 132(4): 1553–1592.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of Economic Perspectives* 24(2): 3-30.

Armstrong, Timothy and Michal Kolesár. 2019. Sensitivity analysis using approximate moment condition models. *Cowles Foundation Discussion Paper No. 2158R.*
SSRN: https://ssrn.com/abstract=3337748.

Attanasio, Orazio P., Costas Meghir, and Ana Santiago. 2012a. Education choices in Mexico: Using a structural model and a randomized experiment to evaluate PROGRESA. *Review of Economic Studies* 79(1): 37–66.

Attanasio, Orazio P., Costas Meghir, and Ana Santiago. 2012b. Supplementary data for Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA. Accessed at <https://academic.oup.com/restud/article/79/1/37/1562110#supple mentary-data> in October 2017.

Berkowitz, Daniel, Mehmet Caner, and Ying Fang. 2008. Are "nearly exogenous instruments" reliable? *Economic Letters* 101(1): 20–23.

Bonhomme, Stéphane and Martin Weidner. 2018. Minimizing sensitivity to model misspecification. arXiv:1807.02161v2 [econ.EM].

Chen, Xiaohong and Andres Santos. 2018. Overidentification in regular models. *Econometrica* 86(5): 1771-1817.

Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi. 2012. Plausibly exogenous. *Review of Economics and Statistics* 94(1): 260–272.

Cressie, Noel, and Timothy RC Read. 1984. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B* 46(3): 440–464.

Dridi, Ramdan, Alain Guay, and Eric Renault. 2007. Indirect inference and calibration of dynamic stochastic general equilibrium models. *Journal of Econometrics* 136(2): 397-430.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. Incentives work: Getting teachers to come to school. *American Economic Review* 102(4): 1241–1278.

Einav, Liran, Amy Finkelstein, Stephen P. Ryan, Paul Schrimpf, and Mark R. Cullen. 2013. Selection on moral hazard in health insurance. *American Economic Review* 103(1): 178–219.

Fetter, Daniel K. and Lee M. Lockwood. 2018. Government old-age support and labor supply: Evidence from the Old Age Assistance Program. *American Economic Review* 108(8): 2174-2211.

Gentzkow, Matthew. 2007a. Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review* 97(3): 713-744.

Gentzkow, Matthew. 2007b. Supplementary data for Valuing new goods in a model with complementarity: Online newspapers. Accessed at <https://www.aeaweb.org/aer/data/june07/20050374_data.zip> in February 2016.

Gentzkow, Matthew and Jesse M. Shapiro. 2015. Measuring the sensitivity of parameter estimates to sample statistics. NBER Working Paper No. 20673.

Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson. 2014. Competition and ideological diversity: Historical evidence from US newspapers. *American Economic Review* 104(10): 3073–3114.

Guggenberger, Patrik. 2012. On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption. *Econometric Theory* 28(2): 387–421.

Hansen, Bruce E. 2016. Efficient shrinkage in parametric models. *Journal of Econometrics* 190(1): 115-132.

Hansen, Lars P., and Thomas J. Sargent. 2001. Acknowledging misspecification in macroeconomic theory. *Review of Economic Dynamics* 4(3): 519-35.

Hansen, Lars P., and Thomas J. Sargent. 2005. Robust estimation and control under commitment. *Journal of Economic Theory* 124(2): 258-301.

Hansen, Lars P., and Thomas J. Sargent. 2016. Sets of models and prices of uncertainty. NBER Working Paper No. 22000.

Hansen, Lars P., Thomas J. Sargent, Gauhar Turmuhambetova, and Noah Williams. 2006. Robust control and model misspecification. *Journal of Economic Theory* 128(1): 45-90.

Heckman, James J. 2010. Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature* 48(2): 356-98.

Hendren, Nathaniel. 2013a. Private information and insurance rejections. *Econometrica* 81(5): 1713–1762.

Hendren, Nathaniel. 2013b. Supplementary data for Private information and insurance rejections. Accessed at <https://www.econometricsociety.org/content/supplement-private-information-and-insurance-rejections-0> in March 2014.

Huber, Peter J. and Elvezio M. Ronchetti. 2009. *Robust Statistics* (2nd ed). Hoboken, NJ: John Wiley & Sons.

Ichimura, Hidehiko and Whitney K. Newey. 2015. The influence function of semiparametric estimators. arXiv:1508.01378v1 [stat.ME].

Keane, Michael P. 2010. Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics* 156(1): 3–20.

Kitamura, Yuichi, Taisuke Otsu, and Kirill Evdokimov. 2013. Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica* 81(3): 1185-1201.

Lehmann, Erich L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses.* New York: Springer.

Morten, Melanie. 2019. Temporary migration and endogenous risk sharing in village India. *Journal of Political Economy* 127(1): 1-46.

Matzkin, Rosa L. 2007. Nonparametric identification. In James J. Heckman and Edward E. Leamer, Eds., *Handbook of Econometrics,* Vol. 6(B), Ch. 73: 5307-5368. Amsterdam: North-Holland.

Mukhin, Yaroslav. 2018. Sensitivity of regular estimators. arXiv:1805.08883v1 [econ.EM].

Nakamura, Emi and Jón Steinsson. 2018. Identification in macroeconomics. *Journal of Economic Perspectives* 32(3): 59-86.

Newey, Whitney K. 1985. Generalized method of moments specification testing. *Journal of Econometrics* 29(3): 229–256.

Newey, Whitney K. 1994. The asymptotic variance of semiparametric estimators. *Econometrica* 62(6): 1349-1382.

Newey, Whitney K., and Daniel McFadden. 1994. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden, Eds., *Handbook of Econometrics,* Vol. 4, Ch. 36: 2111-2245. Amsterdam: North-Holland.

Pakes, Ariel. 2014. Behavioral and descriptive forms of choice models. *International Economic Review* 55(3): 603-624.

Rieder, Helmut. 1994. *Robust Asymptotic Statistics.* New York: Springer.

Spenkuch, Jörg L., B. Pablo Montagnes, and Daniel B. Magleby. 2018. Backward induction in the wild? Evidence from sequential voting in the US Senate. *American Economics Review* 108(7): 1971-2013.

van der Vaart, Aad W. 1998. *Asymptotic Statistics.* Cambridge, UK: Cambridge University Press.

A Proofs

A.1 Proof of Proposition 1

First consider $F \in \mathcal{F}^N(c)$. By the definition of $\mathcal{F}^N(\cdot)$ there exist $\eta \in \mathbb{R}^p$, $\zeta \in \mathbb{R}^k$ such that $F = F(\eta, \zeta)$ and $c = c(\eta)$. Note, moreover, that since $c(\eta) = L'\eta$ while $E_F[\hat{c}] = L'\eta + C'\zeta$, $E_F[\hat{c} - c] = C'\zeta$. Thus, our task reduces to showing that

$$\left\{C'\zeta : \zeta \in \mathbb{R}^k, \|\zeta\|_{\Omega^{-1}} \leq \mu\right\} = [-\mu\sigma_c, \mu\sigma_c].$$

Note, however, that $C'\zeta = C'\Omega^{\frac{1}{2}}\Omega^{-\frac{1}{2}}\zeta$, so by the Cauchy-Schwarz inequality, $|C'\zeta| \leq \sigma_c \|\zeta\|_{\Omega^{-1}}$. Hence, to prove the result we need only show that any bias $\bar{c}$ with $|\bar{c}| \leq \mu\sigma_c$ can be achieved. To this end, pick such a $|\bar{c}| \leq \mu\sigma_c$. Consider $\zeta = \frac{\bar{c}}{\sigma_c^2}\Omega C$ and note that $C'\zeta = \bar{c}$, $\|\zeta\|_{\Omega^{-1}} = \frac{\bar{c}}{\sigma_c} \leq \mu$, as desired.

Next consider $F \in \mathcal{F}^R(c)$. By the definition of $\mathcal{F}^R(\cdot)$ there exist $\eta \in \mathbb{R}^p$, $\zeta \in \mathbb{R}^k$ such that $F = F(\eta, \zeta)$, $c = c(\eta)$, and $\Gamma'(X\eta + \zeta) = \Gamma' X\eta$. Thus, our task reduces to showing that

$$\left\{C'\zeta : \zeta \in \mathbb{R}^k, \|\zeta\|_{\Omega^{-1}} \leq \mu, \Gamma'\zeta = 0\right\} = \left[-\mu\sigma_c\sqrt{1-\Delta}, \mu\sigma_c\sqrt{1-\Delta}\right].$$

Let us first show that for any ζ with $\|\zeta\|_{\Omega^{-1}} \leq \mu$ and $\Gamma'\zeta = 0$, $C'\zeta$ satisfies these bounds. To this end, let us define $\tilde{C} = C - \Gamma\Lambda'$ for $\Lambda = \Sigma_{c\gamma}\Sigma_{\gamma\gamma}^{-1}$, and note that for any ζ with $\Gamma'\zeta = 0$, $\tilde{C}'\zeta = C'\zeta$. Note, next, that by the Cauchy-Schwarz inequality, $\left|\tilde{C}'\zeta\right| \leq \sqrt{\tilde{C}'\Omega\tilde{C}}\|\zeta\|_{\Omega^{-1}}$, and

$$\tilde{C}'\Omega\tilde{C} = \sigma_c^2 - \Sigma_{c\gamma}\Sigma_{\gamma\gamma}^{-1}\Sigma_{\gamma c} = \sigma_c^2(1-\Delta),$$

from which the result follows. We next want to show that for any $\bar{c}$ with $|\bar{c}| \leq \mu\sigma_c\sqrt{1-\Delta}$ there exists ζ with $\|\zeta\|_{\Omega^{-1}} \leq \mu$ and $\Gamma'\zeta = 0$ such that $C'\zeta = \bar{c}$. This result is trivial if $\Delta = 1$, so let us suppose that $\Delta < 1$. In this case, let us consider $\bar{c}$ with $|\bar{c}| \leq \mu\sigma_c\sqrt{1-\Delta}$. Define $\zeta = \frac{\bar{c}}{\sigma_c^2(1-\Delta)}\Omega\tilde{C}$ and note that $\Gamma'\zeta = 0$, $C'\zeta = \tilde{C}'\zeta = \bar{c}$, while

$$\|\zeta\|_{\Omega^{-1}}^2 = \frac{\bar{c}^2}{\sigma_c^4}\left(\sigma_c^2 - \Delta\right) = \frac{\bar{c}^2}{\sigma_c^2(1-\Delta)},$$

which is bounded above by μ^2.

A.2 Proof of Lemma 1

By Lemma 7.6 of van der Vaart (1998), Assumption 3 implies that $\sqrt{f_{h,z}(D_i; t_h, t_z)}$ is differentiable in quadratic mean in the sense that for all $(h, z) \in \mathcal{H} \times \mathcal{Z}$,

$$\int \left(\sqrt{f_{h,z}(D_i; t_h, t_z)} - \sqrt{f_{h,z}(D_i; 0, 0)} - \frac{1}{2}\left(t_h s_h(d) + t_z s_z(d)\right)\sqrt{f_{h,z}(D_i; 0, 0)}\right)^2 d\nu(d) = o\left(\left\|(t_h, t_z)'\right\|^2\right)$$

as $(t_h, t_z) \to 0$. Hence, Theorem 7.2 of van der Vaart (1998) implies that under $S(0,0)$, defining $F^n = \times_{i=1}^n F$ and again taking $s_z(d) = \frac{\partial}{\partial t_z} \log(f_{h,z}(d;0,0))$ and $s_h(d) = \frac{\partial}{\partial t_h} \log(f_{h,z}(d;0,0))$,

$$\log\left(\frac{dF_{h,z}^n\left(\frac{t_h}{\sqrt{n}}, \frac{t_z}{\sqrt{n}}\right)}{dF_0^n}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(t_h s_h(D_i) + t_z s_z(D_i)\right) - \frac{1}{2} \begin{pmatrix} t_h \\ t_z \end{pmatrix}' I_{h,z}(0,0) \begin{pmatrix} t_h \\ t_z \end{pmatrix} + o_p(1)$$

and that $E_{F_0}[s_h(D_i)] = E_{F_0}[s_z(D_i)] = 0$, Since $E_{F_0}\left[s_h(D_i)^2\right]$ and $E_{F_0}\left[s_z(D_i)^2\right]$ are finite, Assumption 1, the Central Limit Theorem, and Slutsky's Lemma imply that under $S(0,0)$, for $g(D_i; h, z) = s_h(D_i) + s_z(D_i)$,

$$\begin{pmatrix} \log\left(\frac{dF_{h,z}^n\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)}{dF_0^n}\right) & \frac{1}{\sqrt{n}} \sum \phi_c(D_i) & \frac{1}{\sqrt{n}} \sum \phi_\gamma(D_i)' \end{pmatrix}'$$

$$\to_d N\left(\begin{pmatrix} -\frac{1}{2} E_{F_0}\left[g(D_i; h, z)^2\right] \\ 0 \\ 0 \end{pmatrix}, \Sigma^*\right),$$

for

$$\Sigma^* = \begin{pmatrix} E_{F_0}\left[g(D_i; h, z)^2\right] & E_{F_0}\left[g(D_i; h, z)\, \phi_c(D_i)\right] & E_{F_0}\left[g(D_i; h, z)\, \phi_\gamma(D_i)'\right] \\ E_{F_0}\left[g(D_i; h, z)\, \phi_c(D_i)\right] & E_{F_0}\left[\phi_c(D_i)^2\right] & E_{F_0}\left[\phi_c(D_i)\, \phi_\gamma(D_i)'\right] \\ E_{F_0}\left[g(D_i; h, z)\, \phi_\gamma(D_i)\right] & E_{F_0}\left[\phi_\gamma(D_i)\, \phi_c(D_i)\right] & E_{F_0}\left[\phi_\gamma(D_i)\, \phi_\gamma(D_i)'\right] \end{pmatrix}.$$

By Le Cam's first lemma (see Example 6.5 of van der Vaart 1998) the convergence in distribution of $\log\left(dF_{h,z}^n\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right) / dF_0^n\right)$ to a normal with mean equal to $-\frac{1}{2}$ of its variance implies that the sequences $\times_{i=1}^n F_0$ and $\times_{i=1}^n dF_{h,z}^n\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)$ are mutually contiguous. Le Cam's third lemma (see Example 6.7 of van der Vaart 1998) then implies that under $S(h,z)$,

$$\begin{pmatrix} \log\left(\frac{dF_{h,z}^n\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)}{dF_0^n}\right) & \frac{1}{\sqrt{n}} \sum \phi_c(D_i) & \frac{1}{\sqrt{n}} \sum \phi_\gamma(D_i)' \end{pmatrix}'$$

$$\to_d N\left(\begin{pmatrix} \frac{1}{2} E_{F_0}\left[g(D_i; h, z)^2\right] \\ E_{F_0}\left[\phi_c(D_i)\, g(D_i; h, z)\right] \\ E_{F_0}\left[\phi_\gamma(D_i)\, g(D_i; h, z)\right] \end{pmatrix}, \Sigma^*\right).$$

Together with contiguity, Assumption 1 implies that

$$\sqrt{n}\left(\hat{c} - c(\eta_0), \hat{\gamma}' - \gamma(\eta_0)'\right) - \frac{1}{\sqrt{n}}\left(\sum \phi_c(D_i), \sum \phi_\gamma(D_i)'\right) = o_p(1),$$

from which the result is immediate for

$$\left(\begin{array}{c}\bar{c}\left(S\left(h,z\right)\right)\\ \bar{\gamma}\left(S\left(h,z\right)\right)\end{array}\right)=\left(\begin{array}{c}E_{F_0}\left[\phi_c\left(D_i\right)g\left(D_i;h,z\right)\right]\\ E_{F_0}\left[\phi_\gamma\left(D_i\right)g\left(D_i;h,z\right)\right]\end{array}\right)=\left(\begin{array}{c}E_{F_0}\left[\phi_c\left(D_i\right)\left(s_h\left(D_i\right)+s_z\left(D_i\right)\right)\right]\\ E_{F_0}\left[\phi_\gamma\left(D_i\right)\left(s_h\left(D_i\right)+s_z\left(D_i\right)\right)\right]\end{array}\right).$$

A.3 Proof of Proposition 2

Let us first consider the case with $S\in\mathcal{S}^N\left(c^\star\right)$, with $\mathcal{S}^N\left(c^\star\right)$ nonempty. By Assumption 2 and Lemma 1,

$$c^\star\left(h\right)=\bar{c}\left(S\left(h,0\right)\right)=E_{F_0}\left[\phi_c\left(D_i\right)s_h\left(D_i\right)\right].$$

By the definition of $\mathcal{S}^N\left(c^\star\right)$ and Lemma 1, for any $S\in\mathcal{S}^N\left(c^\star\right)$ there exist $\left(h,z\right)\in\mathcal{H}\times\mathcal{Z}$ with $S=S\left(h,z\right)$ and $c^\star\left(h\right)=c^\star$. For this $\left(h,z\right)$ we can write

$$\bar{c}\left(S\right)-c^\star=E_{F_0}\left[\phi_c\left(D_i\right)\left(s_h\left(D_i\right)+s_z\left(D_i\right)\right)\right]-E_{F_0}\left[\phi_c\left(D_i\right)s_h\left(D_i\right)\right]=E_{F_0}\left[\phi_c\left(D_i\right)s_z\left(D_i\right)\right].$$

Writing $\bar{c}_z=E_{F_0}\left[\phi_c\left(D_i\right)s_z\left(D_i\right)\right]$ for brevity, our task thus reduces to showing

$$\left\{\bar{c}_z:z\in\mathcal{Z},E_{F_0}\left[s_z\left(D_i\right)^2\right]\leq\mu^2\right\}=\left[-\mu\sigma_c,\mu\sigma_c\right].$$

Note, however, that by the Cauchy-Schwarz inequality

$$\left|\bar{c}_z\right|\leq\sqrt{E_{F_0}\left[\phi_c\left(D_i\right)^2\right]}\sqrt{E_{F_0}\left[s_z\left(D_i\right)^2\right]}\leq\mu\sigma_c.$$

Hence, for any $z\in\mathcal{Z}$ with $E_{F_0}\left[s_z\left(D_i\right)^2\right]\leq\mu^2$, $\bar{c}_z$ necessarily satisfies the bounds. Going the other direction, for any $\bar{c}$ with $\left|\bar{c}\right|\leq\mu\sigma_c$, if we take $s^*\left(D_i\right)=\frac{\bar{c}}{\sigma_c^2}\phi_c\left(D_i\right)$, we have $E_{F_0}\left[s^*\left(D_i\right)\phi_c\left(D_i\right)\right]=\bar{c}$, while $E_{F_0}\left[s^*\left(D_i\right)^2\right]=\bar{c}^2/\sigma_c^2\leq\mu^2$. By Assumption 4, however, there exists $z\in\mathcal{Z}$ with $E_{F_0}\left[\left(s^*\left(D_i\right)-s_z\left(D_i\right)\right)^2\right]=0$, so $\bar{c}_z=\bar{c}$ and $E_{F_0}\left[s_z\left(D_i\right)^2\right]\leq\mu^2$, as desired.

For the case with $S\in\mathcal{S}^R\left(c^\star\right)$, the result is a special case of Proposition 4 proved below, setting $\bar{\gamma}=0$.

A.4 Proof of Proposition 3

By the Neyman-Pearson Lemma (see Theorem 3.2.1 in Lehmann and Romano 2005), the most powerful level-α test of $H_0:\left(D_1,...,D_n\right)\sim\times_{i=1}^nF_{h,z}\left(\frac{1}{\sqrt{n}},0\right)$ against $H_1:\left(D_1,...,D_n\right)\sim\times_{i=1}^nF_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)$ rejects when the log likelihood ratio $\log\left(dF_{h,z}^n\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)/dF_{h,z}^n\left(\frac{1}{\sqrt{n}},0\right)\right)$ exceeds a critical value $v_{\alpha,n}$ chosen to ensure rejection probability α under H_0 (and may randomize when the log likelihood ratio exactly equals the critical value). Here we again abbreviate $\times_{i=1}^nF=F^n$.

From the quadratic expansion of the likelihood in the proof of Lemma 1, however, we see that under $S\left(0,0\right)$, for $g\left(D_i;h,z\right)=s_h\left(D_i\right)+s_z\left(D_i\right)$, since $s_z\left(d\right)=s_h\left(d\right)=0$ when $h=z=0$,

$$\left(\begin{array}{cc} \log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},0\right)}{dF^n_0}\right) & \log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)}{dF^n_0}\right) \end{array} \right)' \rightarrow_d$$

$$N\left(\left(\begin{array}{c} -\frac{1}{2}E_{F_0}\left[g\left(D_i;h,0\right)^2\right] \\ -\frac{1}{2}E_{F_0}\left[g\left(D_i;h,z\right)^2\right] \end{array} \right), \tilde{\Sigma}\right)$$

for

$$\tilde{\Sigma} = \left(\begin{array}{cc} E_{F_0}\left[g\left(D_i;h,0\right)^2\right] & E_{F_0}\left[g\left(D_i;h,0\right)g\left(D_i;h,z\right)\right] \\ E_{F_0}\left[g\left(D_i;h,0\right)g\left(D_i;h,z\right)\right] & E_{F_0}\left[g\left(D_i;h,z\right)^2\right] \end{array} \right).$$

Le Cam's third lemma thus implies that under $S\left(h,0\right)$,

$$\left(\begin{array}{cc} \log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},0\right)}{dF^n_0}\right) & \log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)}{dF^n_0}\right) \end{array} \right)' \rightarrow_d$$

$$N\left(\left(\begin{array}{c} \frac{1}{2}E_{F_0}\left[g\left(D_i;h,0\right)^2\right] \\ -\frac{1}{2}E_{F_0}\left[g\left(D_i;h,z\right)^2\right] + E_{F_0}\left[g\left(D_i;h,0\right)g\left(D_i;h,z\right)\right] \end{array} \right), \tilde{\Sigma}\right),$$

while under $S\left(h,z\right)$,

$$\left(\begin{array}{cc} \log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},0\right)}{dF^n_0}\right) & \log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)}{dF^n_0}\right) \end{array} \right)' \rightarrow_d$$

$$N\left(\left(\begin{array}{c} -\frac{1}{2}E_{F_0}\left[g\left(D_i;h,0\right)^2\right] + E_{F_0}\left[g\left(D_i;h,0\right)g\left(D_i;h,z\right)\right] \\ \frac{1}{2}E_{F_0}\left[g\left(D_i;h,z\right)^2\right] \end{array} \right), \tilde{\Sigma}\right).$$

Since

$$\log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)}{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},0\right)}\right) = \log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)}{dF^n_0}\right) - \log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},0\right)}{dF^n_0}\right),$$

while $g\left(D_i;h,0\right) - g\left(D_i;h,z\right) = -g\left(D_i;0,z\right)$ and $E_{F_0}\left[g\left(D_i;0,z\right)^2\right] = E_{F_0}\left[s_z\left(D_i\right)^2\right]$, we see that

$$\log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)}{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},0\right)}\right) \rightarrow_d$$

$$\begin{cases} N\left(-\frac{1}{2}E_{F_0}\left[g\left(D_i;0,z\right)^2\right], E_{F_0}\left[g\left(D_i;0,z\right)^2\right]\right) & \text{Under } S\left(h,0\right) \\ N\left(\frac{1}{2}E_{F_0}\left[g\left(D_i;0,z\right)^2\right], E_{F_0}\left[g\left(D_i;0,z\right)^2\right]\right) & \text{Under } S\left(h,z\right). \end{cases}$$

Hence, since $v_{\alpha,n}$ corresponds to the $1-\alpha$ quantile of the log likelihood ratio under the null, we see that it converges to the $1-\alpha$ quantile of a $N\left(-\frac{1}{2}E_{F_0}\left[s_z\left(D_i\right)^2\right], E_{F_0}\left[s_z\left(D_i\right)^2\right]\right)$ distribution.

Thus,

$$\frac{\log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)}{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},0\right)}\right)-v_{\alpha,n}}{\sqrt{E_{F_0}\left[s_z\left(D_i\right)^2\right]}}\to_d\begin{cases}N\left(-v_\alpha,1\right) & \text{under } S\left(h,0\right)\\ N\left(\sqrt{E_{F_0}\left[s_z\left(D_i\right)^2\right]}-v_\alpha,1\right) & \text{under } S\left(h,z\right)\end{cases}$$

for v_α the $1-\alpha$ quantile of a standard normal distribution, from which the result follows.

A.5 Proof of Proposition 4

By Assumption 2 and Lemma 1 we again have

$$c^\star\left(h\right)=\bar{c}\left(S\left(h,0\right)\right)=E_{F_0}\left[\phi_c\left(D_i\right)s_h\left(D_i\right)\right].$$

Note, next, that by the definition of $\mathcal{S}^R\left(c^\star,\bar{\gamma}\right)$ and Lemma 1, for any $S\in\mathcal{S}^R\left(c^\star,\bar{\gamma}\right)$ there exist $(h,z)\in\mathcal{H}\times\mathcal{Z}$ with $S=S\left(h,z\right)$, $c^\star\left(h\right)=c^\star$, and

$$E_{F_0}\left[\phi_\gamma\left(D_i\right)\left(s_h\left(D_i\right)+s_z\left(D_i\right)\right)\right]-E_{F_0}\left[\phi_\gamma\left(D_i\right)s_h\left(D_i\right)\right]=E_{F_0}\left[\phi_\gamma\left(D_i\right)s_z\left(D_i\right)\right]=\bar{\gamma}.$$

Thus, writing $\bar{\gamma}_z=E_{F_0}\left[\phi_\gamma\left(D_i\right)s_z\left(D_i\right)\right]$ and $\bar{c}_z=E_{F_0}\left[\phi_c\left(D_i\right)s_z\left(D_i\right)\right]$ for brevity, our task reduces to showing that

$$\left\{\bar{c}_z:z\in\mathcal{Z},\bar{\gamma}_z=\bar{\gamma},E_{F_0}\left[s_z\left(D_i\right)^2\right]\leq\mu^2\right\}=\left[\Lambda\bar{\gamma}-\mu\left(\bar{\gamma}\right)\sigma_c\sqrt{1-\Delta},\Lambda\bar{\gamma}+\mu\left(\bar{\gamma}\right)\sigma_c\sqrt{1-\Delta}\right].$$

Define $s\left(D_i;\bar{\gamma}\right)=\phi_\gamma\left(D_i\right)'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}$, and

$$\varepsilon_z\left(D_i\right)=s_z\left(D_i\right)-s\left(D_i;\bar{\gamma}_z\right).$$

Note that $E_{F_0}\left[\phi_\gamma\left(D_i\right)\varepsilon_z\left(D_i\right)\right]=0$ and $E_{F_0}\left[s\left(D_i;\bar{\gamma}_z\right)\varepsilon_z\left(D_i\right)\right]=0$ by construction. We can write

$$\bar{c}_z=E_{F_0}\left[\phi_c\left(D_i\right)s_z\left(D_i\right)\right]=E_{F_0}\left[\phi_c\left(D_i\right)\phi_\gamma\left(D_i\right)'\right]\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}_z+E_{F_0}\left[\phi_c\left(D_i\right)\varepsilon_z\left(D_i\right)\right]$$

$$=\Lambda\bar{\gamma}_z+E_{F_0}\left[\phi_c\left(D_i\right)\varepsilon_z\left(D_i\right)\right].$$

Next, define

$$\tilde{\phi}_c\left(D_i\right)=\phi_c\left(D_i\right)-\Lambda\phi_\gamma\left(D_i\right)$$

and note that

$$E_{F_0}\left[\phi_c\left(D_i\right)\varepsilon_z\left(D_i\right)\right]=E_{F_0}\left[\tilde{\phi}_c\left(D_i\right)\varepsilon_z\left(D_i\right)\right].$$

The Cauchy-Schwarz inequality then implies that

$$\left|E_{F_0}\left[\tilde{\phi}_c\left(D_i\right)\varepsilon_z\left(D_i\right)\right]\right|\leq\sqrt{E_{F_0}\left[\tilde{\phi}_c\left(D_i\right)^2\right]}\sqrt{E_{F_0}\left[\varepsilon_z\left(D_i\right)^2\right]}$$

$$= \sqrt{\sigma_c^2 - \Lambda\Sigma_{\gamma\gamma}\Lambda'}\sqrt{E_{F_0}\left[\varepsilon_z\left(D_i\right)^2\right]} = \sigma_c\sqrt{1-\Delta}\sqrt{E_{F_0}\left[s_z\left(D_i\right)^2\right] - \bar{\gamma}_z'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}_z}.$$

Combining these results we see that for z such that $\bar{\gamma}_z = \bar{\gamma}$ and $E_{F_0}\left[s_z\left(D_i\right)^2\right] \leq \mu^2$,

$$\bar{c}_z \in \left[\Lambda\bar{\gamma} - \sigma_c\sqrt{1-\Delta}\sqrt{\mu^2 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}}, \Lambda\bar{\gamma} + \sigma_c\sqrt{1-\Delta}\sqrt{\mu^2 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}}\right],$$

which are the bounds stated in the proposition. In particular,

$$0 \leq E_{F_0}\left[\varepsilon_z\left(D_i\right)^2\right] \leq \mu^2 - \bar{\gamma}_z'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}_z,$$

so if $\bar{\gamma}_z = \bar{\gamma}$ we must have $\bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma} \leq \mu^2$ in order that $E_{F_0}\left[s_z\left(D_i\right)^2\right] \leq \mu^2$. Hence, if $\mu^2 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma} < 0$ there exists no z with $\bar{\gamma}_z = \bar{\gamma}$ and $E_{F_0}\left[s_z\left(D_i\right)^2\right] \leq \mu^2$.

To complete the proof it remains to show that these bounds are tight, so that for any $(\bar{c}, \bar{\gamma}, \mu)$ with

$$\bar{c} \in \left[\Lambda\bar{\gamma} - \sigma_c\sqrt{1-\Delta}\sqrt{\mu^2 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}}, \Lambda\bar{\gamma} + \sigma_c\sqrt{1-\Delta}\sqrt{\mu^2 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}}\right] \tag{16}$$

there exists $z \in \mathcal{Z}$ with $\bar{c}_z = \bar{c}$, $\bar{\gamma}_z = \bar{\gamma}$, and $E_{F_0}\left[s_z\left(D_i\right)^2\right] \leq \mu^2$. To prove that this is the case, let us assume without loss of generality that $\mu = 1$ (the result in other cases corresponds to a rescaling of this case). If $\Delta < 1$, define

$$s^*\left(D_i; \bar{c}, \bar{\gamma}\right) = s\left(D_i; \bar{\gamma}\right) + \tilde{\phi}_c\left(D_i\right)\frac{\bar{c} - \Lambda\bar{\gamma}}{\sigma_c^2\left(1-\Delta\right)}.$$

Note that

$$E_{F_0}\left[\phi_\gamma\left(D_i\right)s^*\left(D_i; \bar{c}, \bar{\gamma}\right)\right] = \bar{\gamma}$$

while

$$E_{F_0}\left[\phi_c\left(D_i\right)s^*\left(D_i; \bar{c}, \bar{\gamma}\right)\right] = \Lambda\bar{\gamma} + E_{F_0}\left[\tilde{\phi}_c\left(D_i\right)^2\right]\frac{\bar{c} - \Lambda\bar{\gamma}}{\sigma_c^2\left(1-\Delta\right)} = \bar{c}.$$

Moreover,

$$E_{F_0}\left[s^*\left(D_i; \bar{c}, \bar{\gamma}\right)^2\right] = E_{F_0}\left[s\left(D_i; \bar{\gamma}\right)^2\right] + E_{F_0}\left[\tilde{\phi}_c\left(D_i\right)^2\right]\frac{\left(\bar{c} - \Lambda\bar{\gamma}\right)^2}{\sigma_c^4\left(1-\Delta\right)^2}$$

$$= \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma} + \frac{\left(\bar{c} - \Lambda\bar{\gamma}\right)^2}{\sigma_c^2\left(1-\Delta\right)}.$$

However, by (16) we know that

$$\left|\bar{c} - \Lambda\bar{\gamma}\right| \leq \sigma_c\sqrt{1-\Delta}\sqrt{1 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}}$$

and thus that

$$\frac{(\bar{c}-\Lambda\bar{\gamma})^2}{\sigma_c^2(1-\Delta)} \leq \left(1-\bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}\right)$$

so $E_{F_0}\left[s^*(D_i;\bar{c},\bar{\gamma})^2\right] \leq 1$. By Assumption 4, however, there exists $z \in \mathcal{Z}$ with

$$E_{F_0}\left[\left(s_z(D_i) - s^*(D_i;\bar{c},\bar{\gamma})\right)^2\right] = 0,$$

and thus z yields $\bar{c}_z = \bar{c}$, $\bar{\gamma}_z = \bar{\gamma}$, and $E_{F_0}\left[s_z(D_i)^2\right] \leq 1$ as desired. In cases with $\Delta = 1$, on the other hand, we can use $s^*(D_i;\bar{c},\bar{\gamma}) = s(D_i;\bar{\gamma})$.

A.6 Proof of Proposition 5

The proof of Lemma 1 shows that the distribution of the data under $S(h,z)$ is mutually contiguous with that under $S(0,0)$. Hence, to establish convergence in probability under all local perturbations, it suffices to establish convergence in probability under $S(0,0)$. Consistency of $\hat{\Delta}$ and $\hat{\Lambda}$ under $S(0,0)$ is implied by the Continuous Mapping Theorem (see e.g. Theorem 2.3 of van der Vaart 1998) and the maintained assumptions that $\sigma_c^2 > 0$ and $\Sigma_{\gamma\gamma}$ has full rank.

B Extensions

B.1 Asymptotic Divergence

This section studies the asymptotic behavior of the divergence

$$r\left(F_{h,z}\left(\frac{1}{\sqrt{n}},0\right),F_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)\right) = E_{F_{h,z}(t_h,0)}\left[\psi\left(\frac{f_{h,z}\left(D_i;\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)}{f_{h,z}\left(D_i;\frac{1}{\sqrt{n}},0\right)}\right)\right] \tag{17}$$

as $n \to \infty$, where as in the main text we assume that $\psi(1) = 0$ and $\psi''(1) = 2$. To derive our results we impose the following assumption.

Assumption 6. *For $t = (t_h, t_z) \in \mathbb{R}^2$ and $f_{h,z}(D_i;t) = f_{h,z}(D_i;t_h,t_z)$, $f_{h,z}(D_i;t)$ is twice continuously differentiable in t at 0, and there exists an open neighborhood $\mathcal{B}$ of zero such that*

$$E_{F_0}\left[\sup_{t\in\mathcal{B}}\left(\left|\frac{\partial}{\partial t_z}f_{h,z}(D_i;t)\right| + \left|\frac{\partial^2}{\partial t_z^2}f_{h,z}(D_i;t)\right| + \left|\frac{f_{h,z}(D_i;t_h,0)}{f_{h,z}(D_i;0)}\psi'\left(\frac{f_{h,z}(D_i;t)}{f_{h,z}(D_i;t_h,t)}\right)\frac{\frac{\partial}{\partial t_z}f_{h,z}(D_i;t)}{f_{h,z}(D_i;t)}\right|\right)\right],$$

$$E_{F_0}\left[\sup_{(t,\tilde{t})\in\mathcal{B}^2}\frac{f_{h,z}(D_i;t_h,0)}{f_{h,z}(D_i;0)}\left|\psi'\left(\frac{f_{h,z}(D_i;\tilde{t})}{f_{h,z}(D_i;t)}\right)\frac{\frac{\partial^2}{\partial t_z^2}f_{h,z}(D_i;\tilde{t})}{f_{h,z}(D_i;t)}\right|\right],$$

and

$$E_{F_0}\left[\sup_{(t,\tilde{t})\in\mathcal{B}^2}\left|\frac{f_{h,z}(D_i;t_h,0)}{f_{h,z}(D_i;0)}\psi''\left(\frac{f_{h,z}(D_i;\tilde{t})}{f_{h,z}(D_i;t)}\right)\left(\frac{\frac{\partial}{\partial t_z}f_{h,z}(D_i;\tilde{t})}{f_{h,z}(D_i;t)}\right)^2\right|\right]$$

are finite.

Under this assumption, we obtain the asymptotic approximation to divergence discussed in the main text.

Proposition 6. *Under Assumptions 3 and 6,*

$$\lim_{n\to\infty} n \cdot r\left(F_{h,z}\left(\frac{1}{\sqrt{n}},0\right),F_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)\right) = E_{F_0}\left[s_z\left(D_i\right)^2\right].$$

Proof of Proposition 6 Recall than $r\left(F_{h,z}\left(\frac{1}{\sqrt{n}},0\right),F_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)\right)$ can be written as is (17). Assumption 6 and Leibniz's rule implies for n sufficiently large we can exchange integration and differentiation twice, so by Taylor's Theorem with a mean-value residual,[18]

$$n \cdot r\left(F_{h,z}\left(\frac{1}{\sqrt{n}},0\right),F_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)\right) =$$

$$n \cdot E_{F_0}\left[\begin{array}{c} \frac{f_{h,z}(D_i;t_n)}{f_{h,z}(D_i;0)}\left(\psi\left(\frac{f_{h,z}(D_i;t_n)}{f_{h,z}(D_i;t_n)}\right)+\psi'\left(\frac{f_{h,z}(D_i;t_n)}{f_{h,z}(D_i;t_n)}\right)\frac{\frac{\partial}{\partial t_z}f_{h,z}(D_i;t_n)}{f_{h,z}(D_i;t_n)}\frac{1}{\sqrt{n}}+\right. \\ \left.\frac{1}{2}\left(\psi'\left(\frac{f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}(D_i;t_n)}\right)\frac{\frac{\partial^2}{\partial t_z^2}f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}(D_i;t_n)}+\psi''\left(\frac{f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}(D_i;t_n)}\right)\left(\frac{\frac{\partial}{\partial t_z}f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}(D_i;t_n)}\right)^2\right)\frac{1}{n}\right)\end{array}\right]$$

for $t_n = \left(\frac{1}{\sqrt{n}},0\right)$, $\tilde{t}_n = \left(\frac{1}{\sqrt{n}},\tilde{t}_{z,n}\right)$ and $\tilde{t}_{z,n} \in \left[0,\frac{1}{\sqrt{n}}\right]$. Thus, since $\psi(1)=0$ by assumption,

$$n \cdot r\left(F_{h,z}\left(\frac{1}{\sqrt{n}},0\right),F_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)\right) =$$

$$E_{F_0}\left[\begin{array}{c} \sqrt{n}\psi'(1)\frac{\frac{\partial}{\partial t_z}f_{h,z}(D_i;t_n)}{f_{h,z}(D_i;0)}+ \\ \frac{1}{2}\frac{f_{h,z}(D_i;t_n)}{f_{h,z}(D_i;0)}\left(\psi'\left(\frac{f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}(D_i;t_n)}\right)\frac{\frac{\partial^2}{\partial t_z^2}f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}(D_i;t_n)}+\psi''\left(\frac{f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}(D_i;t_n)}\right)\left(\frac{\frac{\partial}{\partial t_z}f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}(D_i;t_n)}\right)^2\right)\end{array}\right].$$

Assumption 6 and Leibniz's rule imply that for n sufficiently large,

$$E_{F_0}\left[\frac{\frac{\partial}{\partial t_z}f_{h,z}\left(D_i;t_n\right)}{f_{h,z}\left(D_i;0\right)}\right] = \int \frac{\partial}{\partial t_z}f_{h,z}\left(d;\frac{1}{\sqrt{n}},0\right)d\nu(d) = \frac{\partial}{\partial t_z}\int f_{h,z}\left(d;\frac{1}{\sqrt{n}},0\right)d\nu(d) = 0.$$

Hence, we see that

$$n \cdot r\left(F_{h,z}\left(\frac{1}{\sqrt{n}},0\right),F_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)\right) =$$

[18] Specifically, note that for

$$q(t_h,t_z) = r\left(F_{h,z}\left(t_h,0\right),F_{h,z}\left(t_h,t_z\right)\right)$$

we can write

$$q(t_h,t_z) = q(t_h,0)+\frac{\partial}{\partial t_z}q(t_h,0)t_z+\frac{1}{2}\frac{\partial^2}{\partial t_z^2}q(t_h,\tilde{t}_z)t_z^2$$

with $\tilde{t}_z \in [0,t_z]$.

$$E_{F_0}\left[\frac{1}{2}\frac{f_{h,z}\left(D_i;t_n\right)}{f_{h,z}\left(D_i;0\right)}\left(\psi'\left(\frac{f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}\left(D_i;t_n\right)}\right)\frac{\frac{\partial^2}{\partial t_z^2}f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}\left(D_i;t_n\right)}+\psi''\left(\frac{f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}\left(D_i;t_n\right)}\right)\left(\frac{\frac{\partial}{\partial t_z}f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}\left(D_i;t_n\right)}\right)^2\right)\right].$$

Since $\psi''(1)=2$, the Dominated Convergence Theorem and Assumption 6 imply that

$$E_{F_0}\left[\frac{1}{2}\frac{f_{h,z}\left(D_i;t_n\right)}{f_{h,z}\left(D_i;0\right)}\left(\psi'\left(\frac{f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}\left(D_i;t_n\right)}\right)\frac{\frac{\partial^2}{\partial t_z^2}f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}\left(D_i;t_n\right)}+\psi''\left(\frac{f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}\left(D_i;t_n\right)}\right)\left(\frac{\frac{\partial}{\partial t_z}f_{h,z}\left(D_i;\tilde{t}_n\right)}{f_{h,z}\left(D_i;t_n\right)}\right)^2\right)\right]$$

$$\to\frac{1}{2}E_{F_0}\left[\psi'(1)\frac{\frac{\partial^2}{\partial t_z^2}f_{h,z}\left(D_i;0\right)}{f_{h,z}\left(D_i;0\right)}+\psi''(1)\left(\frac{\frac{\partial}{\partial t_z}f_{h,z}\left(D_i;0\right)}{f_{h,z}\left(D_i;0\right)}\right)^2\right]$$

$$=E_{F_0}\left[\frac{1}{2}\psi'(1)\frac{\frac{\partial^2}{\partial t_z^2}f_{h,z}\left(D_i;0\right)}{f_{h,z}\left(D_i;0\right)}+s_z\left(D_i\right)^2\right].$$

However, Assumption 6 and Leibniz's rule imply that

$$E_{F_0}\left[\frac{\frac{\partial^2}{\partial t_z^2}f_{h,z}\left(D_i;0\right)}{f_{h,z}\left(D_i;0\right)}\right]=\int\frac{\partial^2}{\partial t_z^2}f_{h,z}\left(d;0\right)d\nu\left(d\right)=\frac{\partial^2}{\partial t_z^2}\int f_{h,z}\left(d;0\right)d\nu\left(d\right)=0,$$

so

$$\lim_{n\to\infty}n\cdot r\left(F_{h,z}\left(\frac{1}{\sqrt{n}},0\right),F_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)\right)=E_{F_0}\left[s_z\left(D_i\right)^2\right],$$

as we wanted to show.

B.2 Non-Local Misspecification

This section develops our informativeness measure based on probability limits, rather than first-order asymptotic bias.

Under Assumptions 1, 3, and 4 provided the estimators $\hat{c}$ and $\hat{\gamma}$ are regular in the sense discussed in Newey (1994), Theorem 2.1 of Newey (1994) implies that the probability limits $\tilde{c}(\cdot)$ and $\tilde{\gamma}(\cdot)$ are asymptotically linear functionals, in the sense that

$$\begin{array}{c}\lim_{t_z\to0}\left\|\tilde{c}\left(F_{0,z}\left(0,t_z\right)\right)-c\left(\eta_0\right)-t_zE_{F_0}\left[s_z\left(D_i\right)\phi_c\left(D_i\right)\right]\right\|/t_z=0\text{ for all }z\in\mathcal{Z}\\ \lim_{t_z\to0}\left\|\tilde{\gamma}\left(F_{0,z}\left(0,t_z\right)\right)-\gamma\left(\eta_0\right)-t_zE_{F_0}\left[s_z\left(D_i\right)\phi_\gamma\left(D_i\right)\right]\right\|/t_z=0\text{ for all }z\in\mathcal{Z}.\end{array}\tag{18}$$

Assumption 2 would be implied by an assumption that $(\hat{c},\hat{\gamma})$ are regular in the base model, so the assumption of regularity of $(\hat{c},\hat{\gamma})$ in the nesting model can be understood as a strengthening of this assumption. See Newey (1994) and Rieder (1994) for discussion. Since (18) only restricts behavior as $t_z\to0$ for fixed z, rather than studying $\tilde{\Delta}(\bar{r})$ as defined in the main text let us instead consider an analogue defined using finite collections of paths. Specifically, for each $z\in\mathcal{Z}$ let

$$\bar{t}\left(z,\bar{r}\right)=\inf\left\{t_z\in\mathbb{R}_+:r\left(F_0,F_{0,z}\left(0,t_z\right)\right)\geq\bar{r}\right\}$$

denote the largest value of t such that $r\left(F_0, F_{0,z}\left(0, t_z\right)\right) < \bar{r}$ for all $t_z < \bar{t}\left(z, \bar{r}\right)$. Let $\mathcal{Z}_+ \subset \mathcal{Z}$ denote the set of z with $E_{F_0}\left[s_z\left(D_i\right)^2\right] > 0$.

Let $Q \subset \mathcal{Z}_+$ denote a finite subset of $\mathcal{Z}_+$, and let $\mathcal{Q}$ denote the set of all such finite subsets. Finally, let

$$\tilde{b}_N\left(\bar{r}, Q\right) = \sup\left\{\left|\tilde{c}\left(F_{0,z}\left(0, t_z\right)\right) - c\left(\eta_0\right)\right| : z \in Q, t_z < \bar{t}\left(z, \bar{r}\right)\right\}$$

denote the analogue of $\tilde{b}_N\left(\bar{r}\right)$ based on the finite set of paths Q, and for $\varepsilon > 0$ let

$$\tilde{b}_{R,\varepsilon}\left(\bar{r}, Q\right) = \sup\left\{\left|\tilde{c}\left(F_{0,z}\left(0, t_z\right)\right) - c\left(\eta_0\right)\right| : z \in Q, t_z < \bar{t}\left(z, \bar{r}\right), \left\|\tilde{\gamma}\left(F_{0,z}\left(0, t_z\right)\right) - \gamma\left(F_0\right)\right\| \leq \varepsilon\sqrt{\bar{r}}\right\}$$

denote the analogue of $\tilde{b}_R\left(\bar{r}, Q\right)$ based on Q which allows the probability limit of $\hat{\gamma}$ to change by at most $\varepsilon\sqrt{\bar{r}}$. Because $\tilde{b}_{R,0}\left(\bar{r}, Q\right)$ may equal 0 even for large $\bar{r}$ due to the approximation error in (18), we consider limits as $\varepsilon \downarrow 0$ (i.e., as $\varepsilon \to 0$ from above). Based on these objects, we define the analogue of $\tilde{\Delta}\left(\bar{r}\right)$ as

$$\tilde{\Delta}\left(\bar{r}, \mathcal{Q}\right) = \sup_{Q_1 \in \mathcal{Q}} \inf_{Q_2 \in \mathcal{Q}} \lim_{\varepsilon \downarrow 0} \frac{\tilde{b}_{R,\varepsilon}\left(\bar{r}, Q_1\right)}{\tilde{b}_N\left(\bar{r}, Q_2\right)},$$

provided the limit exists.

Proposition 7. *Suppose Assumptions 1, 3, and 4 hold, that the estimators $\hat{c}$ and $\hat{\gamma}$ are regular, and that Assumption 6 holds for $h = 0$ and all $z \in \mathcal{Z}_+$. For $r\left(F_0, F\right) = E_{F_0}\left[\psi\left(\frac{dF}{dF_0}\right)\right]$ with $\psi\left(\cdot\right)$ twice continuously differentiable and $\psi\left(1\right) = 0$, $\psi''\left(1\right) = 2$,*

$$\sup_{Q_1 \in \mathcal{Q}} \inf_{Q_2 \in \mathcal{Q}} \lim_{\varepsilon \downarrow 0} \lim_{\bar{r} \downarrow 0} \frac{\tilde{b}_{R,\varepsilon}\left(\bar{r}, Q_1\right)}{\tilde{b}_N\left(\bar{r}, Q_2\right)} = \sqrt{1 - \Delta}.$$

It is important that we take the limit as $\bar{r} \downarrow 0$ inside the limit as $\varepsilon \downarrow 0$ and the sup and inf, since this order of limits allows us to take advantage of the approximation result (18).

Proof of Proposition 7 Note, first, that our Assumptions 1-4 imply the conditions of Theorem 2.1 of Newey (1994) other than regularity of $(\hat{c}, \hat{\gamma})$. Specifically, our Assumption 3 implies that our paths are what Newey (1994) terms "regular." Condition (i) of Theorem 2.1 in Newey (1994) is then immediate from our Assumption 4. Condition (ii) likewise follows from Assumption 4. Condition (iii) is implied by our Assumption 1. Regularity of $(\hat{c}, \hat{\gamma})$ is assumed, so Theorem 2.1 of Newey (1994) implies (18).

Note, next, that for any $z \in \mathcal{Z}_+$, the proof of Proposition 6 implies that

$$\lim_{t_z \downarrow 0} r\left(F_0, F_{0,z}\left(0, t_z\right)\right) / t_z^2 = E_{F_0}\left[s_z\left(D_i\right)^2\right].$$

Hence, as $\bar{r} \downarrow 0$, $\bar{t}(z,\bar{r})/\sqrt{\bar{r}} \to E\left[s_z(D_i)^2\right]^{-\frac{1}{2}}$. For all $z \in \mathcal{Z}_+$, (18) implies that

$$\begin{aligned}\lim_{\bar{r}\downarrow 0} \sup_{t_z \leq \bar{t}(z,\bar{r})} \left\|\tilde{c}\left(F_{0,z}(0,t_z)\right) - c(\eta_0) - t_z E_{F_0}\left[s_z(D_i)\,\phi_c(D_i)\right]\right\| / t_z = 0 \\ \lim_{\bar{r}\downarrow 0} \sup_{t_z \leq \bar{t}(z,\bar{r})} \left\|\tilde{\gamma}\left(F_{0,z}(0,t_z)\right) - \gamma(\eta_0) - t_z E_{F_0}\left[s_z(D_i)\,\phi_\gamma(D_i)\right]\right\| / t_z = 0,\end{aligned}$$

and thus that

$$\left\{\frac{1}{\sqrt{\bar{r}}}\left(\tilde{c}\left(F_{0,z}(0,t_z)\right) - c(\eta_0), \tilde{\gamma}\left(F_{0,z}(0,t_z)\right) - \gamma(\eta_0)\right) : t_z \leq \bar{t}(z,\bar{r})\right\}$$

$$\to \left\{\tilde{t}_z\left(E_{F_0}\left[s_z(D_i)\,\phi_c(D_i)\right], E_{F_0}\left[s_z(D_i)\,\phi_\gamma(D_i)\right]\right) : \tilde{t}_z \leq E_{F_0}\left[s_z(D_i)^2\right]^{-\frac{1}{2}}\right\}$$

in the Hausdorff sense as $\bar{r} \downarrow 0$. Correspondingly, for any $Q \in \mathcal{Q}$,

$$\left\{\frac{1}{\sqrt{\bar{r}}}\tilde{c}\left(F_{0,z}(0,t_z) - c(\eta_0), \tilde{\gamma}\left(F_{0,z}(0,t_z)\right) - \gamma(\eta_0)\right) : z \in Q, t_z \leq \bar{t}(z,\bar{r})\right\}$$

$$\to \left\{\tilde{t}_z\left(E_{F_0}\left[s_z(D_i)\,\phi_c(D_i)\right], E_{F_0}\left[s_z(D_i)\,\phi_\gamma(D_i)\right]\right) : z \in Q, \tilde{t}_z \leq E_{F_0}\left[s_z(D_i)^2\right]^{-\frac{1}{2}}\right\}.$$

Hence, for any nonempty $Q \in \mathcal{Q}$

$$\frac{1}{\sqrt{\bar{r}}}\tilde{b}_N(\bar{r},Q) \to \max\left\{\frac{\left|E_{F_0}\left[s_z(D_i)\,\phi_c(D_i)\right]\right|}{E_{F_0}\left[s_z(D_i)^2\right]^{\frac{1}{2}}} : z \in Q\right\} \text{ as } \bar{r} \downarrow 0.$$

Matters are somewhat more delicate for $\tilde{b}_{R,\varepsilon}(\bar{r},Q)$. Note, in particular, that for $\varepsilon > 0$, as $\bar{r} \downarrow 0$ we have

$$\frac{1}{\sqrt{\bar{r}}}\tilde{b}_{R,\varepsilon}(\bar{r},Q) \to$$

$$\sup\left\{\tilde{t}_z E_{F_0}\left[s_z(D_i)\,\phi_c(D_i)\right] : z \in Q, \tilde{t}_z \leq E_{F_0}\left[s_z(D_i)^2\right]^{-\frac{1}{2}}, \tilde{t}_z \left\|E_{F_0}\left[s_z(D_i)\,\phi_\gamma(D_i)\right]\right\| \leq \varepsilon\right\}$$

$$= \sup\left\{\tilde{t}_z E_{F_0}\left[s_z(D_i)\,\phi_c(D_i)\right] : z \in Q, \tilde{t}_z \leq \min\left\{E_{F_0}\left[s_z(D_i)^2\right]^{-\frac{1}{2}}, \frac{\varepsilon}{\left\|E_{F_0}\left[s_z(D_i)\,\phi_\gamma(D_i)\right]\right\|}\right\}\right\},$$

where we define $\varepsilon/0 = \infty$ for $\varepsilon > 0$. Consequently,

$$\frac{1}{\sqrt{\bar{r}}}\tilde{b}_{R,\varepsilon}(\bar{r},Q) \to$$

$$\sup\left\{\tilde{t}_z\left|E_{F_0}\left[s_z(D_i)\,\phi_c(D_i)\right]\right| : z \in Q, \tilde{t}_z \leq \min\left\{E_{F_0}\left[s_z(D_i)^2\right]^{-\frac{1}{2}}, \frac{\varepsilon}{\left\|E_{F_0}\left[s_z(D_i)\,\phi_\gamma(D_i)\right]\right\|}\right\}\right\}.$$

Note, however, that by the Cauchy-Schwarz inequality and $E_{F_0}\left[s_z(D_i)^2\right] < \infty$, $E_{F_0}\left[s_z(D_i)\,\phi_c(D_i)\right]$

is finite for all $z \in \mathcal{Z}$, for any z with $E_{F_0}\left[s_z\left(D_i\right)\phi_\gamma\left(D_i\right)\right] \neq 0$,

$$\frac{\varepsilon}{\left\|E_{F_0}\left[s_z\left(D_i\right)\phi_\gamma\left(D_i\right)\right]\right\|} E_{F_0}\left[s_z\left(D_i\right)\phi_c\left(D_i\right)\right] \to 0$$

as $\varepsilon \downarrow 0$. Hence, as $\varepsilon \downarrow 0$,

$$\sup\left\{\tilde{t}_z\left|E_{F_0}\left[s_z\left(D_i\right)\phi_c\left(D_i\right)\right]\right| : z \in Q, \tilde{t}_z \leq \min\left\{E_{F_0}\left[s_z\left(D_i\right)^2\right]^{-\frac{1}{2}}, \frac{\varepsilon}{\left\|E_{F_0}\left[s_z\left(D_i\right)\phi_\gamma\left(D_i\right)\right]\right\|}\right\}\right\}$$

$$\to \max\left\{\frac{\left|E_{F_0}\left[s_z\left(D_i\right)\phi_c\left(D_i\right)\right]\right|}{E_{F_0}\left[s_z\left(D_i\right)^2\right]^{\frac{1}{2}}} : z \in Q_0\right\}$$

for $Q_0 = \left\{z \in Q : E_{F_0}\left[s_z\left(D_i\right)\phi_\gamma\left(D_i\right)\right] = 0\right\}$, where we define this max to be zero if Q_0 is empty.

This immediately implies that

$$\lim_{\varepsilon\downarrow 0}\lim_{\bar{r}\downarrow 0}\frac{\tilde{b}_{R,\varepsilon}\left(\bar{r}, Q_1\right)}{\tilde{b}_N\left(\bar{r}, Q_2\right)} = \frac{\max\left\{\left|E_{F_0}\left[s_z\left(D_i\right)\phi_c\left(D_i\right)\right]\right| / E_{F_0}\left[s_z\left(D_i\right)^2\right]^{\frac{1}{2}} : z \in Q_{1,0}\right\}}{\max\left\{\left|E_{F_0}\left[s_z\left(D_i\right)\phi_c\left(D_i\right)\right]\right| / E_{F_0}\left[s_z\left(D_i\right)^2\right]^{\frac{1}{2}} : z \in Q_2\right\}}$$

for $Q_{1,0} = \left\{z \in Q_1 : E_{F_0}\left[s_z\left(D_i\right)\phi_\gamma\left(D_i\right)\right] = 0\right\}$, provided the denominator on the right hand side is non-zero.[19]

To complete the proof, note that for $\mathcal{Q}_0$ the set of possible Q_0,

$$\sup_{Q_1\in\mathcal{Q}}\inf_{Q_2\in\mathcal{Q}}\lim_{\varepsilon\downarrow 0}\lim_{\bar{r}\downarrow 0}\frac{\tilde{b}_{R,\varepsilon}\left(\bar{r}, Q_1\right)}{\tilde{b}_N\left(\bar{r}, Q_2\right)} = \frac{\sup_{Q_0\in\mathcal{Q}_0}\max\left\{\left|E_{F_0}\left[s_z\left(D_i\right)\phi_c\left(D_i\right)\right]\right| / E_{F_0}\left[s_z\left(D_i\right)^2\right]^{\frac{1}{2}} : z \in Q_0\right\}}{\sup_{Q\in\mathcal{Q}}\max\left\{\left|E_{F_0}\left[s_z\left(D_i\right)\phi_c\left(D_i\right)\right]\right| / E_{F_0}\left[s_z\left(D_i\right)^2\right]^{\frac{1}{2}} : z \in Q\right\}}.$$

The proofs of Propositions 2 and 4 show, however, that

$$\max_{z\in\mathcal{Z}_+}\left|E_{F_0}\left[s_z\left(D_i\right)\phi_c\left(D_i\right)\right]\right| / E_{F_0}\left[s_z\left(D_i\right)^2\right]^{\frac{1}{2}} = \sigma_c$$

and

$$\max_{z\in\mathcal{Z}_+ : E_{F_0}[s_z(D_i)\phi_\gamma(D_i)]=0}\left|E_{F_0}\left[s_z\left(D_i\right)\phi_c\left(D_i\right)\right]\right| / E_{F_0}\left[s_z\left(D_i\right)^2\right]^{\frac{1}{2}} = \sigma_c\sqrt{1-\Delta}.$$

Hence,

$$\sup_{Q_1\in\mathcal{Q}}\inf_{Q_2\in\mathcal{Q}}\lim_{\varepsilon\downarrow 0}\lim_{\bar{r}\downarrow 0}\frac{\tilde{b}_{R,\varepsilon}\left(\bar{r}, Q_1\right)}{\tilde{b}_N\left(\bar{r}, Q_2\right)} = \sqrt{1-\Delta},$$

as we wanted to show.

[19] If the denominator on the right hand side is zero, we define the limit as $+\infty$.

B.3 Accounting for Richer Dependence of $\hat{c}$ on the Data

In Section 5, for cases where the function $c(\theta)$ depends on the distribution of the data other than through θ, we effectively fix the distribution of the data at the empirical distribution for the purposes of estimating Δ and Λ. Here we discuss how to allow for uncertainty about the distribution of data in a special case, and present corresponding calculations for our applications.

Suppose in particular that

$$\hat{c} = \frac{1}{n}\sum_i c\left(\hat{\theta}; D_i\right) \tag{19}$$

for some function $c(\cdot)$. In contrast to the setup in Section 5, here we allow that $\hat{c}$ depends on the data directly, and not only through the dependence of $\hat{c}$ on $\hat{\theta}$.

In this case, one can show that the recipe in Section 5 applies, with the modification that

$$\hat{\phi}_c(D_i) = c\left(\hat{\theta}; D_i\right) + \hat{\Lambda}_{cg}\phi_g\left(D_i; \hat{\theta}\right) \tag{20}$$

where $\phi_g\left(D_i; \hat{\theta}\right)$ and $\hat{\Lambda}_{cg}$ are as defined in Section 5, and $\hat{C}$ in the definition of $\hat{\Lambda}_{cg}$ is now given by the gradient of $\frac{1}{n}\sum_i c(\theta; D_i)$ with respect to θ at $\hat{\theta}$.

The proof of this result, which we omit, proceeds by noting that we can augment the GMM parameter vector as (c, θ), and correspondingly augment the moment equation as $(c(\theta; D_i) - c, \phi_g(D_i; \theta))$, following which we can derive the estimated influence function for $\hat{c}$ as we would for any element of $\hat{\theta}$.

In the cases of Attanasio et al. (2012a) and Gentzkow (2007a), we can represent the calculation of $\hat{c}$ in the form given in (19) and thus calculate $\hat{\Delta}$ using the modified estimated influence function in (20). In the case of Attanasio et al. (2012a), the estimates in Table 1 change from 0.283, 0.227, and 0.056, respectively, to 0.277, 0.221, and 0.055. In the case of Gentzkow (2007a), the estimates in Table 2 change from 0.514, 0.009, and 0.503, respectively, to 0.517, 0.008, and 0.507.

Table 1: Estimated informativeness of descriptive statistics for the effect of a counterfactual rebudgeting of PROGRESA (Attanasio et al. 2012a)

Descriptive statistics $\hat{\gamma}$	Estimated informativeness $\hat{\Delta}$
All	0.283
Impact on eligibles	0.227
Impact on ineligibles	0.056

Notes: The table shows the estimated informativeness $\hat{\Delta}$ of three vectors $\hat{\gamma}$ of descriptive statistics for the estimated partial-equilibrium effect $\hat{c}$ of the counterfactual rebudgeting on the school enrollment of eligible children, accumulated across age groups (Attanasio et al. 2012a, sum of ordinates for the line labeled "fixed wages" in Figure 2, minus sum of ordinates for the line labeled "fixed wages" in the left-hand panel of Figure 1). Vector $\hat{\gamma}$ "impact on eligibles" consists of the age-grade-specific treatment-control differences for eligible children (interacting elements of Attanasio et al. 2012a, Table 2, single-age rows of the column labeled "Impact on Poor 97," with the child's grade). Vector $\hat{\gamma}$ "impact on ineligibles" consists of the age-grade-specific treatment-control differences for ineligible children (interacting elements of Attanasio et al. 2012a Table 2, single-age rows of the column labeled "Impact on non-eligible," with the child's grade). Vector $\hat{\gamma}$ "all" consists of both of these groups of statistics. Estimated informativeness $\hat{\Delta}$ is calculated according to the recipe in Section 5.1 using the replication code and data posted by Attanasio et al. (2012b).

Table 2: Estimated informativeness of descriptive statistics for the effect of eliminating the *Post* online edition (Gentzkow 2007a)

Descriptive statistics $\hat{\gamma}$	Estimated informativeness $\hat{\Delta}$
All	0.514
IV coefficient	0.009
Panel coefficient	0.503

Notes: The table shows the estimated informativeness $\hat{\Delta}$ of three vectors $\hat{\gamma}$ of descriptive statistics for the estimated effect $\hat{c}$ on the readership of the *Post* print edition if the *Post* online edition were removed from the choice set (Gentzkow 2007a, table 10, row labeled "Change in *Post* readership"). Vector $\hat{\gamma}$ "IV coefficient" is the coefficient from a 2SLS regression of last-five-weekday print readership on last-five-weekday online readership, instrumenting for the latter with the set of excluded variables such as Internet access at work (Gentzkow 2007a, Table 4, Column 2, first row). Vector $\hat{\gamma}$ "panel coefficient" is the coefficient from an OLS regression of last-one-day print readership on last-one-day online readership controlling for the full set of interactions between indicators for print readership and for online readership in the last five weekdays. Each of these regressions includes the standard set of demographic controls from Gentzkow (2007a, Table 5). Vector $\hat{\gamma}$ "all" consists of both the IV coefficient and the panel coefficient. Estimated informativeness $\hat{\Delta}$ is calculated according to the recipe in Section 5.1 using the replication code and data posted by Gentzkow (2007b).

Table 3: Estimated informativeness of descriptive statistics for the minimum pooled price ratio (Hendren 2013a)

Descriptive statistics $\hat{\gamma}$	Estimated informativeness $\hat{\Delta}$
All	0.700
Fractions in focal point groups	0.005
Fractions in non-focal point groups	0.018
Fraction in each group needing LTC	0.676

Notes: The table shows the estimated informativeness $\hat{\Delta}$ of four vectors $\hat{\gamma}$ of descriptive statistics for the "minimum pooled price ratio" $\hat{c}$ (Hendren 2013a, Table V, row labeled "Reject," column labeled "LTC"). Vector $\hat{\gamma}$ "fractions in focal point groups" consists of the fraction of respondents who report exactly 0, the fraction who report exactly 0.5, and the fraction who report exactly 1. Vector $\hat{\gamma}$ "fractions in non-focal point groups" consists of the fractions of respondents whose reports are in each of the intervals (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5), (0.5, 0.6], (0.6, 0.7], (0.7, 0.8], (0.8, 0.9], and (0.9, 1). Vector $\hat{\gamma}$ "fraction in each group needing LTC" consists of the fractions of respondents giving each of the preceding reports who eventually need long-term care. Vector $\hat{\gamma}$ "all" consists of all three of the other vectors. Estimated informativeness $\hat{\Delta}$ is calculated according to the recipe in Section 5.1 using the replication code and data posted by Hendren (2013b), supplemented with additional calculations provided by the author.