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*

July 2020

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 exactly or asymptotically unbiased under some base model. The researcher also reports descriptive statistics $\hat{\gamma}$ that estimate features γ of the distribution of the data that are related to c under the base model. A reader entertains a less restrictive model that is local to the base model, under which the estimate $\hat{c}$ may be biased. We study the reduction in worst-case bias from a restriction that requires the reader's model to respect the relationship between c and γ specified by the base model. Our main result shows that the proportional reduction in worst-case bias 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.

keywords: local misspecification, transparency

[*]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). Another 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 estimates more interpretable. 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 is interested in a scalar quantity of interest c (say, the effect of a counterfactual policy). The researcher specifies a *base model* that relates the value of c to the distribution F of some data (say, the joint distribution of the data in a randomized experiment). The researcher reports an estimate $\hat{c}$ of c that is unbiased (either exactly or asymptotically) under the base model. A reader of the research may not accept all of the assumptions of the base model, and may therefore be concerned that $\hat{c}$ is biased.

The researcher also reports a vector $\hat{\gamma}$ of descriptive statistics (say, sample mean outcomes in different arms of the experiment). These statistics uncontroversially estimate some features $\gamma = \gamma(F)$ of the distribution F (say, population mean outcomes in different arms). Because the base model specifies the relationship between c and F, it also implicitly specifies the relationship between c and γ, which may or may not be correct.

Suppose the researcher is able to convince the reader of the relationship between c and γ specified by the base model (say, by arguing that the counterfactual policy is similar to one of the arms of the experiment). Should this lessen the reader's concern about bias in $\hat{c}$, even if the reader does not accept the base model in its entirety?

We answer this question focusing on the worst-case bias when the alternative model contem-

[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, 2020).

plated by the reader is local to the base model in an appropriate sense. To outline our approach, it will be useful to define the base model as a correspondence $\mathcal{F}^0(\cdot)$, where $\mathcal{F}^0(c)$ is the set of distributions F consistent with a given value of c under the model. The identified set for c given some F under the base model is found by taking the preimage of F under $\mathcal{F}^0(\cdot)$. We assume that c is point identified under the base model, so that for any F consistent with the base model, the identified set given F is a singleton.

The reader contemplates a model that is less restrictive than the base model. We describe the reader's model by a correspondence $\mathcal{F}^N(\cdot)$, where $\mathcal{F}^N(c) \supseteq \mathcal{F}^0(c)$ is the set of distributions F consistent with a given value of c under the reader's model. Because $\mathcal{F}^N(c) \supseteq \mathcal{F}^0(c)$ for all c, the identified set for c given some F is larger under the reader's model than under the base model, and may not be a singleton. Moreover, $\hat{c}$ may be biased under the reader's model. 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.

To formalize the idea that the reader's model is local to the base model, we suppose that 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

$$\mathcal{F}^N(c) = \cup_{F \in \mathcal{F}^0(c)} \left\{ \tilde{F} \in \mathcal{N}(F) \right\}. \tag{1}$$

We take the neighborhood $\mathcal{N}(F)$ to contain distributions $\tilde{F}$ within a given statistical distance of F.

To formalize the possibility that the researcher convinces the reader of the relationship between c and γ prescribed by the base model, we consider restricting attention to the elements $\tilde{F} \in \mathcal{N}(F)$ such that $\gamma\left(\tilde{F}\right) = \gamma(F)$. This results in the restricted correspondence $\mathcal{F}^{RN}(\cdot)$

$$\mathcal{F}^{RN}(c) = \cup_{F \in \mathcal{F}^0(c)} \left\{ \tilde{F} \in \mathcal{N}(F) : \gamma\left(\tilde{F}\right) = \gamma(F) \right\}. \tag{2}$$

If $\mathcal{F}^0(\cdot)$ implies that only certain values of γ are consistent with a given value of c, then $\mathcal{F}^{RN}(\cdot)$ preserves that implication whereas $\mathcal{F}^N(\cdot)$ may not. For this reason $\mathcal{F}^N(c) \supseteq \mathcal{F}^{RN}(c) \supseteq \mathcal{F}^0(c)$ for all c, i.e., the correspondence $\mathcal{F}^{RN}(\cdot)$ is less restrictive than the base model, but more restrictive than the reader's model. Let b^{RN} denote the largest possible absolute bias in $\hat{c}$ that can arise under $\mathcal{F}^{RN}(\cdot)$. Because $\mathcal{F}^{RN}(\cdot)$ is more restrictive than $\mathcal{F}^N(\cdot)$, we know that $b^{RN} \leq b^N$.

We focus on characterizing the ratio b^{RN}/b^N, which lies between zero and one. Section 2 shows how to derive the correspondences $\mathcal{F}^N(\cdot)$, $\mathcal{F}^{RN}(\cdot)$, and $\mathcal{F}^0(\cdot)$, and the worst-case biases b^N and b^{RN}, from explicitly parameterized economic models. Section 3 provides an exact characterization of b^{RN}/b^N in a linear model with normal errors. Section 4 provides an approximate characterization of b^{RN}/b^N in more general nonlinear models, obtained via a local asymptotic analysis. Sections 3

and 4 show that under given conditions the ratio b^{RN}/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 information in the data that determines $\hat{c}$. 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. Section 5 shows 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.

Section 6 implements 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. In the second application, to Gentzkow (2007a), the quantity c of interest is the effect of removing the online edition of the *Washington Post* 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 parameter governing the existence of insurance markets, and the descriptive statistics $\hat{\gamma}$ summarize the joint distribution of self-reported beliefs about the likelihood of loss events and the realizations of these events. In each case, we report an estimate of Δ for various definitions of $\hat{\gamma}$, and we discuss the implications for the interpretation of $\hat{c}$. 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 bias 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.

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 $\Delta = 1$).[2] In Online Appendix A, we generalize our main result to accommodate the setting of Andrews et al. (2017) and so provide a unified treatment of sensitivity and informativeness.

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 Kolesár, 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 Kolesár (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 readers of applied research to sharpen their opinions about the reliability of the researcher's conclusions, thus improving transparency in the sense of Andrews et al. (2020).

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).[3] It is common in applied research to describe

[2] The present paper draws on the analysis of "sensitivity to descriptive statistics" in Gentzkow and Shapiro (2015).

[3] See also Dridi et al. (2007) and Nakamura and Steinsson (2018) for discussion of the appropriate choice of

the data features that "primarily identify" structural parameters or "drive" estimates of those parameters.[4] As Keane (2010) and others have noted, such statements are not directly related to the formal notion of identification in econometrics (see also Andrews et al. 2020). Their intended meaning is therefore up for grabs. If researchers are prepared to reinterpret these as statements about informativeness, then our approach provides a way to sharpen and quantify these statements at low cost to researchers.

2 Setup and Key Definitions

The introduction describes our approach in terms of correspondences between the quantity of interest c and the distribution F of the data. In this section we first show how to derive these correspondences from explicitly parameterized economic models, and then use these correspondences to define the worst-case biases that we characterize in our analysis. Section 4 defines analogous objects in a local asymptotic framework.

Suppose that, under the base model considered by the researcher, both the distribution of the data F and the quantity of interest c are determined by a structural parameter $\eta \in H$. Formally, under the base model, we have that $F = F(\eta)$ and $c = c(\eta)$ so the correspondence $\mathcal{F}^0(\cdot)$ is given by

$$\mathcal{F}^0(c) = \{F(\eta) : \eta \in H, c(\eta) = c\}.$$

Because the structural parameter η determines the distribution F, it also determines $\gamma = \gamma(\eta) = \gamma(F(\eta))$.

Suppose further that, under the reader's model, the distribution of the data F is determined by η and by a misspecification parameter $\zeta \in Z$ (say, indexing economic forces omitted from the researcher's model) that is normalized to zero under the base model. Formally, under the reader's model we have that $F = F(\eta, \zeta)$, with $F(\eta, 0) = F(\eta)$ for all $\eta \in H$, and correspondingly that $\gamma = \gamma(\eta, \zeta) = \gamma(F(\eta, \zeta))$, with $\gamma(\eta, 0) = \gamma(\eta)$ for all $\eta \in H$. We focus on settings where forms of misspecification indexed by ζ are rich, in the sense that the range of $F(\eta, \zeta)$ under $\zeta \in Z$ does not depend on η. For simplicity we continue to write the quantity of interest as a function of η alone, $c = c(\eta)$. Forms of misspecification that change the mapping from η to c but yield the same set of (c, F) pairs are equivalent to those we study.[5]

moments to match when fitting macroeconomic models.

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

[5] Specifically, consider a setup where the distribution of the data is $F = F(\eta, \zeta)$ as above, while the quantity of interest is $c = \tilde{c}(\eta, \zeta)$. Our assumptions imply that the set $\{(c(\eta), F(\eta, \zeta)) : \eta \in H, \zeta \in Z\}$ is a Cartesian product, equal to $\{c(\eta) : \eta \in H\} \times \{F(\eta, \zeta) : \eta \in H, \zeta \in Z\}$. So long as $\{(\tilde{c}(\eta, \zeta), F(\eta, \zeta)) : \eta \in H, \zeta \in Z\}$ is likewise a Cartesian product (implying that c remains unidentified absent further restrictions), and $\{\tilde{c}(\eta, \zeta) : \eta \in H, \zeta \in Z\} =$

We formalize the idea that the reader's model is local to the base model as follows. Let $r(\eta, \zeta) \geq 0$ denote some Cressie-Read (1984) divergence between the distribution $F(\eta)$ and the distribution $F(\eta, \zeta)$, so that $r(\eta, 0) = 0$ for all $\eta \in H$. For any distribution $F = F(\eta)$ consistent with the base model, we define the neighborhood $\mathcal{N}(F)$ to consist of all distributions $F(\eta, \zeta)$ such that the divergence $r(\eta, \zeta)$ is less than some scalar bound $\mu \geq 0$:

$$\mathcal{N}(F) = \{F(\eta, \zeta) : \eta \in H, F(\eta) = F, \zeta \in Z, r(\eta, \zeta) \leq \mu\}.$$

We then define the reader's model $\mathcal{F}^N(\cdot)$ as in (1).[6] The neighborhood $\mathcal{N}(F)$ is increasing in μ. Hence, larger values of μ imply imply a greater relaxation of assumptions as we move from the base model $\mathcal{F}^0(\cdot)$ to the reader's model $\mathcal{F}^N(\cdot)$. We suppress the dependence of $\mathcal{N}(F)$ and $\mathcal{F}^N(c)$ on μ for brevity.

The base model specifies a relationship between c and γ in the sense that if the quantity of interest takes value c, then the feature γ must take a value $\gamma(\eta)$ for some $\eta \in H$ such that $c = c(\eta)$. The reader's model $\mathcal{F}^N(\cdot)$ need not respect the base model's specification of the relationship between c and γ. By contrast, the model $\mathcal{F}^{RN}(\cdot)$, defined in (2), respects the base model's specification of the relationship between c and γ in the sense that, for any $F \in \mathcal{F}^{RN}(c)$, there is some $\eta \in H$ such that $c = c(\eta)$, $\gamma(F) = \gamma(\eta)$, and $F \in \mathcal{N}(F(\eta))$.[7] Hence, a given (c, γ) pair is compatible with $\mathcal{F}^{RN}(\cdot)$ if and only if it is compatible with $\mathcal{F}^0(\cdot)$.

The researcher chooses an estimator $\hat{c}$ that is unbiased under the base model in the sense that $E_F[\hat{c} - c] = 0$ for any $F \in \mathcal{F}^0(c)$, where $E_F[\cdot]$ denotes the expectation when the data are distributed according to F. The estimator $\hat{c}$ may be biased under the reader's model $\mathcal{F}^N(\cdot)$, and indeed if we take μ to infinity the parameter c is completely unidentified under $\mathcal{F}^N(\cdot)$. The largest absolute bias in $\hat{c}$ that is possible under $\mathcal{F}^N(\cdot)$ is

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

Considering $\mathcal{F}^{RN}(\cdot)$ rather than $\mathcal{F}^N(\cdot)$ can reduce the worst-case bias in $\hat{c}$. The largest absolute bias in $\hat{c}$ that is possible under $\mathcal{F}^{RN}(\cdot)$ is

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

$\{c(\eta) : \eta \in H\}$ (implying that the change from $c = c(\eta)$ to $c = \tilde{c}(\eta, \zeta)$ does not change the set of possible c's), the correspondences $\mathcal{F}(\cdot)$ that we consider are the same whether constructed from the setup with $c = c(\eta)$ or that with $c = \tilde{c}(\eta, \zeta)$.

[6] Specifically, $\mathcal{F}^N(c) = \{F(\eta, \zeta) : \eta \in H, c(\eta) = c, \zeta \in Z, r(\eta, \zeta) \leq \mu\}$.

[7] To see that this is the case, note that $\mathcal{F}^{RN}(c) = \{F(\eta, \zeta) : \eta \in H, c(\eta) = c, \zeta \in Z, r(\eta, \zeta) \leq \mu, \gamma(F(\eta, \zeta)) = \gamma(\eta)\}$.

The proportional reduction in worst-case bias from limiting attention to $\mathcal{F}^{RN}(\cdot)$ is measured by the ratio b_{RN}/b_N, which is the primary focus of our analysis.

3 Informativeness in a Linear Normal Setting

To build intuition for our approach, we next specialize to a linear normal setting and provide an exact characterization of the ratio b^{RN}/b^N. We illustrate with a stylized example, and conclude the section with some further discussion of our approach and its limitations.

3.1 Characterization of Worst-Case Bias

Now suppose that $H = \mathbb{R}^p$, $Z = \mathbb{R}^k$, and that under $F(\eta, \zeta)$ the data $Y \in \mathbb{R}^k$ follow

$$Y \sim N(X\eta + \zeta, \Omega) \tag{3}$$

for X and Ω known, nonrandom matrices with full column rank.

The quantity of interest is some linear function $c(\eta) = L'\eta$ of the parameters, with $L \in \mathbb{R}^{p\times 1}$ a known, non-random vector. The researcher chooses a linear estimator $\hat{c} = C'Y$ for C a vector. The researcher ensures that $\hat{c}$ is unbiased for c under $\mathcal{F}^0(\cdot)$ by choosing $C' = L'M$ for some matrix M with $MX = I_p$.

The researcher computes the vector $\hat{\gamma} = \Gamma'Y$ of descriptive statistics, with $\Gamma \in \mathbb{R}^{k\times p_\gamma}$ a known, non-random matrix. The vector $\hat{\gamma}$ is trivially unbiased for $\gamma(\eta, \zeta) = \Gamma'(X\eta + \zeta)$.

Absent any restriction on ζ the quantity of interest c is entirely unidentified. Intuitively, without any restriction on ζ, the mean of the data Y is entirely unrestricted for any fixed η, making it impossible to learn $c = L'\eta$. The reader's model $\mathcal{F}^N(\cdot)$ limits the size of ζ. In particular, given (3), the assumption that $r(\cdot,\cdot)$ is in the Cressie-Read family implies that $r(\eta, \zeta)$ is a strictly increasing transformation of $\|\zeta\|_{\Omega^{-1}}$, for $\|V\|_A = \sqrt{V'AV}$. Thus, for this section we define $\mathcal{N}(\cdot)$ based on the restriction $\|\zeta\|_{\Omega^{-1}} \leq \mu$.[8]

Under the base model $\mathcal{F}^0(\cdot)$,

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

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

[8] That is, we let

$$\mathcal{N}(F) = \left\{F(\eta, \zeta) : \eta \in H, F(\eta) = F, \zeta \in Z, \|\zeta\|_{\Omega^{-1}} \leq \mu\right\}.$$

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. Informativeness determines the ratio of worst-case biases b_{RN}/b_N.

Proposition 1. *The set of possible biases under* $\mathcal{F}^N(\cdot)$ *is*

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

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

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

for any c*. Hence,* $b_N = \mu\sigma_c$*,* $b_{RN} = \mu\sigma_c\sqrt{1-\Delta}$*, and*

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

All proofs are collected at the end of the paper.

Importantly, the value of Δ, and hence the proportional reduction in worst-case bias from restricting from $\mathcal{F}^N(\cdot)$ to $\mathcal{F}^{RN}(\cdot)$, does not depend on μ. In addition to characterizing the worst-case biases b_{RN} and b_N, Proposition 1 characterizes the set of possible biases under $\mathcal{F}^N(\cdot)$ and $\mathcal{F}^{RN}(\cdot)$, showing in particular that any absolute bias smaller than the worst case is achievable. Imposing additional restrictions on ζ, beyond those captured by $\mathcal{F}^N(\cdot)$ or $\mathcal{F}^{RN}(\cdot)$, could further reduce the worst-case bias.

3.2 Example

To fix ideas, suppose that a researcher observes i.i.d. data from a randomized evaluation of a conditional cash transfer program. The program gives each household a payment of size s if their children attend school regularly. Households are uniformly randomized among subsidy levels $s \in \{0,1,2\}$. We can think of those receiving $s=0$ as the control group.

The data consist of the average school attendance Y_s of children assigned subsidy $s \in \{0,1,2\}$. The quantity of interest c is the expected attendance at a counterfactual subsidy level $s^* > 2$.

Under the base model the mean of Y_s is given by

$$\eta_1 + \eta_2 s \tag{4}$$

for $s \in \{0, 1, 2, s^*\}$. Average attendance Y_s is independent and homoskedastic across arms of the experiment, with standard deviation ω.[9]

Under the base model $\mathcal{F}^0(\cdot)$, c can be estimated by linear extrapolation of average attendance from two or more of the observed subsidy levels $s \in \{0, 1, 2\}$ to subsidy level s^*. We continue to assume that the researcher chooses a linear estimator $\hat{c}$ that is unbiased for c under $\mathcal{F}^0(\cdot)$.[10]

Under the reader's model $\mathcal{F}^N(\cdot)$, the estimator $\hat{c}$ may be biased. Intuitively, if $\zeta \neq 0$ then the mean of Y_s may be nonlinear in s, so that linear extrapolation to s^* may produce a biased estimate of c. The restriction to $\mathcal{F}^{RN}(\cdot)$ can lessen the scope for bias. The economic content of the restriction depends on the choice of Γ, which in turn determines the descriptive statistic $\hat{\gamma} = \Gamma' Y$ and the informativeness Δ.

As a concrete example, suppose that a reader entertains that the effect of incentivizing school attendance may be discontinuous at zero, with the mean of Y_s for $s \in \{0, 1, 2, s^*\}$ given by

$$\tilde{\eta}_0 + 1\{s > 0\}\tilde{\eta}_1 + s\tilde{\eta}_2 \tag{5}$$

for $\tilde{\eta}$ a composite of η and ζ with $\tilde{\eta}_1 \neq 0$.[11] The model $\mathcal{F}^N(\cdot)$ allows that the mean of Y_s may follow (5), as long as $\tilde{\eta}_1$ is sufficiently small.[12] The bound b^N thus reflects a worst case over scenarios that include (5).

Whether the set $\mathcal{F}^{RN}(\cdot)$ allows that the mean of Y_s may follow (5) depends on the choice of Γ. If $\Gamma = \begin{pmatrix} e_1 & e_2 \end{pmatrix}$ for e_s the basis vector corresponding to subsidy s, so that $\hat{\gamma} = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix}$, then under $\mathcal{F}^{RN}(\cdot)$ the mean of Y_s is linear in s for $s > 0$, which is consistent with (5). If instead

[9] To cast this example into the notation of Section 3.1, take

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \ \Omega = \omega^2 I_3, L = \begin{pmatrix} 0 \\ s^* \end{pmatrix}.$$

[10] For example, if we take $M = (X'X)^{-1} X'$, then $\hat{c}$ is the the ordinary least squares extrapolation to s^*, and is also the maximum likelihood estimator of c under $\mathcal{F}^0(\cdot)$.

[11] Specifically, choose $\tilde{\eta}$ so that

$$\tilde{X}\tilde{\eta} = \zeta + X\eta$$

for

$$\tilde{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

[12] In particular, to ensure that a given $\tilde{\eta}$ is consistent with $\|\zeta\|_{\Omega^{-1}} \leq \mu$, it suffices that $|\tilde{\eta}_1| \leq \omega\mu$.

$\Gamma = \begin{pmatrix} e_0 & e_1 \end{pmatrix}$, so that $\hat{\gamma} = \begin{pmatrix} Y_0 & Y_1 \end{pmatrix}$, then under $\mathcal{F}^{RN}(\cdot)$ the mean of Y_s is linear in s for $s \neq 2$, which is not consistent with (5). The bound b^{RN} thus reflects a worst case over scenarios that may or may not include (5), depending on the choice of Γ.

The informativeness Δ measures the extent to which the restriction to $\mathcal{F}^{RN}(\cdot)$ lessens the scope for bias, $b_{RN}/b_N = \sqrt{1-\Delta}$. Again imagine a reader who entertains that the mean of Y_s may follow (5). Learning that Δ is close to one for $\hat{\gamma} = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix}$ might be reassuring to this reader because the restriction from $\mathcal{F}^{N}(\cdot)$ to $\mathcal{F}^{RN}(\cdot)$ greatly lessens the scope for bias in $\hat{c}$ while still allowing for (5). Learning that Δ is close to one for $\hat{\gamma} = \begin{pmatrix} Y_0 & Y_1 \end{pmatrix}$ might not be as reassuring, because in this case the restriction from $\mathcal{F}^{N}(\cdot)$ to $\mathcal{F}^{RN}(\cdot)$ rules out (5).

3.3 Discussion

3.3.1 Relationship to Analysis of Identification

Our analysis is distinct from an analysis of identification. We focus on the behavior of a particular estimator $\hat{c}$ 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 $\mathcal{F}^{0}(\cdot)$, but differs from $\hat{c}$.[13] An analysis of identification would conclude that c is point-identified under $\mathcal{F}^{RN}(\cdot)$, whereas our analysis would conclude that the estimator $\hat{c}$ may be biased under $\mathcal{F}^{RN}(\cdot)$.

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 the identified set for c based on the distribution of $\hat{c}$ is $[\hat{c} - \mu\sigma_c, \hat{c} + \mu\sigma_c]$ under $\mathcal{F}^{N}(\cdot)$ and $\left[\hat{c} - \mu\sigma_c\sqrt{1-\Delta}, \hat{c} + \mu\sigma_c\sqrt{1-\Delta}\right]$ under $\mathcal{F}^{RN}(\cdot)$. Under this interpretation, the ratio b_{RN}/b_N measures how much the identified set shrinks when we restrict from $\mathcal{F}^{N}(\cdot)$ to $\mathcal{F}^{RN}(\cdot)$.

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

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

Second, the correspondence $\mathcal{F}^{RN}(\cdot)$ requires that the relationship between c and γ specified by the base model be correct local to each point in the base model. This is more restrictive than requiring that the pair (c, γ) be globally consistent with the base model, which yields the correspondence $\mathcal{F}^{GN}(\cdot)$ with

$$\mathcal{F}^{GN}(c) = \left(\mathcal{F}^{N}(c) \cap \left(\cup_{F^* \in \mathcal{F}^0(c)} \left\{\tilde{F} : \gamma\left(\tilde{F}\right) = \gamma\left(F^*\right)\right\}\right)\right) \tag{6}$$

for all c. If any (c, γ) pair is possible under $\mathcal{F}^0(\cdot)$, then $\mathcal{F}^{GN}(\cdot)$ is equivalent to $\mathcal{F}^{N}(\cdot)$, but $\mathcal{F}^{RN}(\cdot)$ need not be. More generally $\mathcal{F}^{N}(c) \supseteq \mathcal{F}^{GN}(c) \supseteq \mathcal{F}^{RN}(c) \supseteq \mathcal{F}^{0}(c)$, and the ratio of worst case bias under $\mathcal{F}^{GN}(\cdot)$ to worst-case bias under $\mathcal{F}^{N}(\cdot)$ is bounded below by $\sqrt{1-\Delta}$.

Third, 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 mean, 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 mean, 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, and not solely on economic objects such as the distribution of potential outcomes.

4 Informativeness Under Local Misspecification

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

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 parameter of interest remains $c(\eta)$. The researcher computes (i) a scalar estimate $\hat{c}$ of c and (ii) a $p_\gamma \times 1$ vector of descriptive statistics $\hat{\gamma}$.

As in Section 2, to allow the possibility of misspecification we suppose that under the reader's model $D_i \sim F(\eta, \zeta)$ for some $(\eta, \zeta) \in H \times Z$, where $F(\eta, 0) = F(\eta)$ for all $\eta \in H$. The joint distribution for the sample under the reader's model is $\times_{i=1}^{n} F(\eta, \zeta)$. Defining the correspondences $\mathcal{F}^N(\cdot)$ and $\mathcal{F}^{RN}(\cdot)$ as in Section 2, we are interested in the ratio of worst-case biases b^{RN}/b^N.

While Section 3 exactly characterizes b^{RN}/b^N in the linear normal model, we are not aware of similarly tractable expressions in general nonlinear settings. In this section, we therefore instead approximate b^{RN}/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.[14] 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 is instead a device to approximate the finite-sample behavior of estimators in situations where the influence of misspecification is on the same order as sampling uncertainty.[15] Section 4.4 instead considers fixed misspecification and develops results based on probability limits.

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.

4.1 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.

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

[15] The order $\frac{1}{\sqrt{n}}$ perturbation to the base-model parameter η is a common asymptotic tool 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 technique for modeling local misspecification (see e.g. Newey (1985), Andrews et al. (2017), and Armstrong and Kolesár (2019)).

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

$$\sqrt{n}\left(\hat{c}-c\left(\eta_0\right),\hat{\gamma}-\gamma\left(\eta_0\right)\right)=\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}\phi_c\left(D_i\right),\sum_{i=1}^{n}\phi_\gamma\left(D_i\right)\right)+o_p\left(1\right), \tag{7}$$

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

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

and moreover

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

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. 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.

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 implies that $\hat{c}$ is asymptotically unbiased for our local parameter of interest $c^{\star}(h)$ under $S(h,0)$.

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

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

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) $I_{h,z}(t_h,t_z)$ 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 of 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 $s_h(d)=\frac{\partial}{\partial t_h}\log\left(f_{h,z}(d;0,0)\right)$, $s_z(d)=\frac{\partial}{\partial t_z}\log\left(f_{h,z}(d;0,0)\right)$ as the score functions corresponding to h and z, respectively.

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 $E_{F_0}\left[\left(s(D_i)-s_z(D_i)\right)^2\right]=0$.*

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.[16] 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, and so is the local analogue of the richness condition discussed in Section 2. If this assumption fails, the local asymptotic bias bounds we derive below continue to hold, but need not be sharp.

Under Assumption 4, the nesting model allows forms of misspecification against which all specification tests that control size have trivial local asymptotic power.[17] This highlights an important

[16] 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).

[17] In particular, for $h\in\mathcal{H}$, Assumption 4 implies that there exists $z\in\mathcal{Z}$ such that $E_{F_0}\left[\left(s_h(D_i)-s_z(D_i)\right)^2\right]=0$. Arguments along the same lines as e.g. Chen and Santos (2018) then imply that $S(h,0)$ and $S(0,z)$ are asymptotically indistinguishable, and thus that no specification test which controls the rejection probability under $S(h,0)$ has nontrivial power against $S(0,z)$.

aspect of our local analysis. A possible justification for bounding the degree of misspecification (see, e.g., Huber and Ronchetti 2009, p. 294, as quoted in Bonhomme and Weidner 2018) is that specification tests eventually detect unbounded misspecification with arbitrarily high probability, so conditional on non-rejection it is reasonable to focus on bounded, and in particular local, misspecification. By contrast, we allow some forms of misspecification that are statistically undetectable absent knowledge of the true parameters. Hence, restrictions on the magnitude of misspecification in our setting should be understood as a-priori restrictions on the set of models considered, rather than a-posteriori restrictions based on which models survive specification tests.

4.2 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. Moreover, we obtain a simple expression for its asymptotic mean.

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

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

where

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

Moreover, $c^\star\left(h\right)=E_{F_0}\left[\phi_c\left(D_i\right)s_h\left(D_i\right)\right]$, and $\gamma^\star\left(h\right)=E_{F_0}\left[\phi_\gamma\left(D_i\right)s_h\left(D_i\right)\right]$.

Recall that $c^\star\left(h\right)$ is the parameter of interest in our local asymptotic analysis. We can thus interpret $\bar{c}\left(S\left(h,z\right)\right)-c^\star\left(h\right)=E_{F_0}\left[\phi_c\left(D_i\right)s_z\left(D_i\right)\right]$ 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 the 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 $\mathcal{F}^0\left(c\right)$.

To relax the assumption of correct specification, next suppose we bound the degree of local

misspecification by $\mu \geq 0$. For $S \in \mathcal{S}^0(\cdot) = \bigcup_{c^\star} \mathcal{S}^0(c^\star)$, let us define the neighborhood

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

For reasons elaborated in Section 4.3 below, $\mathcal{N}(S)$ is a sequence-space analogue of the neighborhood $\mathcal{N}(F)$ defined in Section 2. Taking a union over $\mathcal{N}(S)$ for $S \in \mathcal{S}^0(c^\star)$ yields

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

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

Finally, let us define a restricted set of sequences as

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

Limiting attention to sequences $S \in \mathcal{S}^{RN}(c^\star)$ is analogous to limiting attention to the set $\mathcal{F}^{RN}(c)$.

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

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

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

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 $S \in \mathcal{S}^{RN}(\cdot)$ is

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

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

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

4.3 Scaling of Perturbations

Under regularity conditions, the bound $E_{F_0}\left[s_z(D_i)^2\right] \leq \mu$ in the definition of $\mathcal{N}(S)$ can be interpreted as a bound on the asymptotic Cressie-Read divergence of $F_{h,z}\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)$ from $F_{h,z}\left(\frac{1}{\sqrt{n}}, 0\right)$.

Specifically, we consider divergences of the form

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

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 (1984) divergences for appropriate λ.

Online Appendix B shows that under regularity conditions

$$\lim_{n\to\infty} n \cdot r_{h,z}(t_h, t_z) = E_{F_0}\left[s_z(D_i)^2\right]. \tag{11}$$

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 $E_{F_0}\left[s_z(D_i)^2\right]$.[18] Online Appendix C shows that bounds on $E_{F_0}\left[s_z(D_i)^2\right]$ also correspond to bounds on the asymptotic power of tests to distinguish elements of $\mathcal{N}(S)$ from S.

4.4 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 reader believes the data follow $\times_{i=1}^{n} F(\eta, \zeta)$, where (η, ζ) do not change with the sample size. Let us denote the probability limits of $\hat{c}$ and $\hat{\gamma}$ under F by $\tilde{c}(F)$ and $\gamma(F)$, respectively. We assume for ease of exposition that these probability limits exist.

To simplify the analysis, let us further fix a value η_0 of the base model parameter, so the true value of the parameter of interest is $c(\eta_0)$. Suppose that for a divergence r of the form considered

[18] 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 $n \to \infty$.

in Section 4.3, $r(\eta,\zeta) = E_{F(\eta,0)}[\psi(f(D_i;\eta,\zeta)/f(D_i;\eta,0))]$, the reader believes that the degree of misspecification is bounded in the sense that $r(\eta_0,\zeta) \leq \mu$. Given this bound, the probability limit of $|\hat{c} - c(\eta_0)|$ is no larger than

$$\tilde{b}_N(\mu) = \sup\{|\tilde{c}(F(\eta_0,\zeta)) - c(\eta_0)| : \zeta \in Z, r(\eta_0,\zeta) \leq \mu\},$$

where we now make the dependence on μ explicit. 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 an analogue of $\mathcal{F}^{RN}(\cdot)$,

$$\tilde{b}_{RN}(\mu) = \sup\{|\tilde{c}(F(\eta_0,\zeta)) - c(\eta_0)| : \zeta \in Z, r(\eta_0,\zeta) \leq \mu, \gamma(F(\eta_0,\zeta)) - \gamma(F_0) = 0\}.$$

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

Provided that $\tilde{b}_N(\mu)$ and $\tilde{b}_{RN}(\mu)$ are both finite and non-zero, we can define a non-local analogue $\tilde{\Delta}(\mu)$ of informativeness Δ by

$$\sqrt{1 - \tilde{\Delta}(\mu)} = \frac{\tilde{b}_{RN}(\mu)}{\tilde{b}_N(\mu)}.$$

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

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) or via a bootstrap (e.g., Hall 1992). Given such an estimate one can construct a plug-in estimate

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

Provided $\hat{\Sigma}$ is consistent under $S(0,0)$, consistency of $\hat{\Sigma}$ and $\hat{\Delta}$ 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 3. *Under Assumptions 3 and 5,* $\hat{\Sigma} \xrightarrow{p} \Sigma$ *and* $\hat{\Delta} \xrightarrow{p} \Delta$ *under* $S(h,z)$ *for any* $h \in \mathcal{H}$, $z \in \mathcal{Z}$.

Mukhin (2018) provides alternative sufficient conditions for consistent estimation of informativeness, and 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, so we can write it as $c(\theta)$. 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 $S(0,0)$,

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

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

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

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 $\hat{M} = \frac{\partial}{\partial\gamma}\hat{m}\left(\hat{\gamma}\right)$.[19] 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

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

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

[19] 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.[20]

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 $\hat{\Delta}$ as in (12).

6 Applications

In this section we present and interpret estimates of Δ for three structural articles in economics, each of which estimates the parameters η of a base model via maximum likelihood. In each case, we estimate the informativeness of each vector $\hat{\gamma}$ for the estimator $\hat{c}$ following the recipe in Section 5.1. Because in each case model estimation is via maximum likelihood and $\hat{\gamma}$ can be represented as GMM, the recipe applies directly.

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

[20] 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. Online Appendix E discusses how to account for such additional dependence on the data, and presents corresponding calculations for some of our applications.

in school. The paper uses the estimated base 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, 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 "exogenous 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.

Table I 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.

Table I. 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).

Restricting from $\mathcal{F}^N(\cdot)$ to $\mathcal{F}^{RN}(\cdot)$ reduces the worst-case bias 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 would require including in $\hat{\gamma}$ descriptive statistics that are orthogonal to the treatment-control differences we consider, thus imposing that $\mathcal{F}^{RN}(\cdot)$ respects the relationship specified by the base model $\mathcal{F}^0(\cdot)$ between c and the features of the distribution of the data estimated by these orthogonal statistics.

To illustrate the distinction between informativeness and identification highlighted in Section 3.3.1, 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. 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.

Table II 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 relationship specified by the base model between the counterfactual c and the pop-

Table II. 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).

ulation value of the IV coefficient would greatly limit the scope for bias in Gentzkow's (2007a) estimator $\hat{c}$.

Our findings suggest otherwise. When $\hat{\gamma}$ consists only of the IV coefficient, restricting from $\mathcal{F}^N(\cdot)$ to $\mathcal{F}^{RN}(\cdot)$ 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, when $\hat{\gamma}$ consists only of the panel coefficient, restricting from $\mathcal{F}^N(\cdot)$ to $\mathcal{F}^{RN}(\cdot)$ reduces the worst-case bias in $\hat{c}$ by an estimated factor of $1-\sqrt{1-0.50} \approx 0.29$. Intuitively, a reader interested in evaluating the scope for bias in $\hat{c}$ may wish to focus more attention on the assumptions of the base model that relate c to the population value of the panel coefficient (e.g., restrictions on the time structure of preference shocks), than on assumptions that relate c to the population value of the IV coefficient (e.g., 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. 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

Table III. 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.

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.

Table III 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. When $\hat{\gamma}$ consists only of the fraction in each group needing LTC, restricting from $\mathcal{F}^N(\cdot)$ to $\mathcal{F}^{RN}(\cdot)$ 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. It is common for a researcher to report descriptive statistics $\hat{\gamma}$ that estimate features γ of the distribution of the data that are in turn related to the quantity c of interest under the researcher's model. A reader who accepts the relationship between the features γ and the structural quantity c specified by the researcher's model, and who believes that the statistics $\hat{\gamma}$ play an important role in "driving" the structural estimate $\hat{c}$, may then be more confident in the researcher's conclusions.

We propose one way to formalize this logic. 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 reduction in worst-case bias from accepting the relationship between γ and c specified by the researcher's model. 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 structural estimates in important economic settings. We recommend that researchers report estimated informativeness alongside their descriptive analyses.

Proofs

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 $C'\zeta = \bar{c}$ and $\|\zeta\|_{\Omega^{-1}} = \frac{\bar{c}}{\sigma_c} \leq \mu$, as desired.

Next consider $F \in \mathcal{F}^{RN}(c)$. By the definition of $\mathcal{F}^{RN}(\cdot)$ 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, 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 $\left|\tilde{C}'\zeta\right| \leq \sqrt{\tilde{C}'\Omega\tilde{C}}\,\|\zeta\|_{\Omega^{-1}}$ by the Cauchy-Schwarz inequality, and that

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

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 $\Delta < 1$, and pick some $\bar{c}$ 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 $\Gamma'\zeta = 0$ and $C'\zeta = \tilde{C}'\zeta = \bar{c}$, while

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

which is bounded above by μ^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 $(h, z) \in \mathcal{H} \times \mathcal{Z}$,

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

as $(t_h, t_z) \to 0$. Hence, Theorem 7.2 of van der Vaart (1998) implies that under $S(0,0)$, defining $F^n = \times_{i=1}^n F$,

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

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 $g(D_i; h, z) = s_h(D_i) + s_z(D_i)$,

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

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

for

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

By Le Cam's first lemma (see Example 6.5 of van der Vaart 1998) the convergence in distribution of $\log\left(dF^n_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)/dF^n_0\right)$ to a normal with mean equal to $-\frac{1}{2}$ of its variance implies that the sequences $\times^n_{i=1}F_0$ and $\times^n_{i=1}F^n_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)$ are mutually contiguous. Le Cam's third lemma (see Example 6.7 of van der Vaart 1998) then implies that under $S\left(h,z\right)$,

$$\left(\begin{array}{cc} \log\left(\frac{dF^n_{h,z}\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right)}{dF^n_0}\right) & \frac{1}{\sqrt{n}}\sum\phi_c\left(D_i\right) \quad \frac{1}{\sqrt{n}}\sum\phi_\gamma\left(D_i\right)' \end{array}\right)'$$

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

Together with contiguity, Assumption 1 implies that

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

under $S\left(h,z\right)$, from which the result is immediate for

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

Finally, note that Assumption 2 implies $c^\star\left(h\right)=\bar{c}\left(S\left(h,0\right)\right)$ and $\gamma^\star\left(h\right)=\bar{\gamma}\left(S\left(h,0\right)\right)$. Consequently, by the results above we have $c^\star\left(h\right)=E_{F_0}\left[\phi_c\left(D_i\right)s_h\left(D_i\right)\right]$, and $\gamma^\star\left(h\right)=E_{F_0}\left[\phi_\gamma\left(D_i\right)s_h\left(D_i\right)\right]$.

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 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

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

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 $|\bar{c}| \leq \mu\sigma_c$, 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 $S \in \mathcal{S}^{RN}\left(c^\star\right)$, note that by the definition of $\mathcal{S}^{RN}\left(c^\star\right)$ and Lemma 1, for any $S \in \mathcal{S}^{RN}\left(c^\star\right)$ 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

$$E_{F_0}\left[\phi_\gamma\left(D_i\right)\left(s_h\left(D_i\right) + s_z\left(D_i\right)\right)\right] - E_{F_0}\left[\phi_\gamma\left(D_i\right) s_h\left(D_i\right)\right] = E_{F_0}\left[\phi_\gamma\left(D_i\right) s_z\left(D_i\right)\right] = 0.$$

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

$$\left\{\bar{c}_z : z \in \mathcal{Z}, \bar{\gamma}_z = 0, E_{F_0}\left[s_z\left(D_i\right)^2\right] \leq \mu^2\right\} = \left[-\mu\sigma_c\sqrt{1-\Delta}, \mu\sigma_c\sqrt{1-\Delta}\right].$$

Let $\Lambda = \Sigma_{\gamma\gamma}^{-1}\Sigma_{\gamma c}$. For $\tilde{\phi}_c\left(D_i\right) = \phi_c\left(D_i\right) - \Lambda'\phi_\gamma\left(D_i\right)$, note that if $\bar{\gamma}_z = 0$ then

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

The Cauchy-Schwarz inequality then implies that

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

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

Hence, we see that for z such that $E_{F_0}\left[s_z\left(D_i\right)^2\right] \leq \mu^2$,

$$\bar{c}_z \in \left[-\mu\sigma_c\sqrt{1-\Delta}\mu, \mu\sigma_c\sqrt{1-\Delta}\right],$$

which are the bounds stated in the proposition.

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

$$\bar{c} \in \left[-\mu\sigma_c\sqrt{1-\Delta}, \mu\sigma_c\sqrt{1-\Delta}\right]$$

there exists $z \in \mathcal{Z}$ with $\bar{c}_z = \bar{c}$, $\bar{\gamma}_z = 0$, and $E_{F_0}\left[s_z(D_i)^2\right] \leq \mu^2$.This result is trivial if $\Delta = 1$, so let us suppose that $\Delta < 1$ and pick some $\bar{c}$ with $|\bar{c}| \leq \mu\sigma_c\sqrt{1-\Delta}$. Now define

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

Note that $E_{F_0}[\phi_\gamma(D_i) s^*(D_i; \bar{c})] = 0$, while

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

Moreover,

$$E_{F_0}\left[s^*(D_i; \bar{c})^2\right] = \frac{\bar{c}^2}{\sigma_c^2(1-\Delta)}.$$

However, by the definition of $\bar{c}$ we know that $|\bar{c}| \leq \mu\sigma_c\sqrt{1-\Delta}$, so $E_{F_0}\left[s^*(D_i; \bar{c})^2\right] \leq \mu^2$. By Assumption 4, however, there exists $z \in \mathcal{Z}$ with

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

and thus z yields $\bar{c}_z = \bar{c}$, $\bar{\gamma}_z = 0$, and $E_{F_0}\left[s_z(D_i)^2\right] \leq \mu^2$ as desired.

Proof of Proposition 3 As shown in the proof of Lemma 1, under Assumption 3 the log likelihood ratio $\log\left(dF^n_{h,z}\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right) / dF^n_{h,z}\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)\right)$ converges under $S(0,0)$ to a normal distribution with mean equal to $-\frac{1}{2}$ times its variance. Le Cam's First Lemma thus implies 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 $S(h,z)$, it suffices to establish convergence in probability under $S(0,0)$. Consistency of $\hat{\Delta}$ under $S(0,0)$ is implied by Assumption 5, 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.

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.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro. 2020. Transparency in structural research. NBER Working Paper No. 26631.

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#supplementary-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.openicpsr.org/openicpsr/project/116273/version/V1/view> in May 2020.

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.

Hall, Peter. 1992. *The Bootstrap and Edgeworth Expansion.* Springer Series in Statistics. New York: Springer-Verlag.

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.