

On the Informativeness of Descriptive Statistics for Structural Estimates

Isaiah Andrews, *Harvard and NBER**

Matthew Gentzkow, *Stanford University and NBER*

Jesse M. Shapiro, *Brown University and NBER*

October 2018

Abstract

Researchers often present treatment-control differences or other descriptive statistics alongside structural estimates that answer policy or counterfactual questions of interest. We ask to what extent confidence in the researcher's interpretation of the former should increase a reader's confidence in the latter. We consider a structural estimate \hat{c} that may depend on a vector of descriptive statistics $\hat{\gamma}$. We define a class of misspecified models in a neighborhood of the assumed model. We then compare the bounds on the bias of \hat{c} due to misspecification across all models in this class with the bounds across the subset of these models in which misspecification does not affect $\hat{\gamma}$. Our main result shows that the ratio of the lengths of these tight bounds depends only on a quantity we call the *informativeness* of $\hat{\gamma}$ for \hat{c} , which can be easily estimated even for complex models. We recommend that researchers report the estimated informativeness of descriptive statistics. We illustrate with applications to three recent papers.

*E-mail: iandrews@fas.harvard.edu, gentzkow@stanford.edu, jesse_shapiro_1@brown.edu. We acknowledge funding from the National Science Foundation, 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, Eric Renault, Jon Roth, Susanne Schennach, José de Souza, and participants at the Radcliffe Institute Conference on Statistics When the Model is Wrong and the HBS Conference on Economic Models of Competition and Collusion 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 treatment-control differences or other descriptive statistics alongside structural estimates that answer policy or counterfactual questions of interest. One leading case is where the structural model is estimated on data from a randomized experiment, and the descriptive statistics are treatment-control differences (e.g., Attanasio et al. 2012a; Duflo et al. 2012; Alatas et al. 2016). A second leading case is where the structural model is estimated on observational data, and the descriptive statistics are regression coefficients or correlations that capture important relationships (e.g., Gentzkow 2007a; Einav et al. 2013; Gentzkow et al. 2014; Morten forthcoming).

An appeal of this approach is that transparent, convincing descriptive statistics may lend credibility to the related structural estimates. Focusing on the case of randomized experiments, Attanasio et al. (2012a) write, “One can think of using experimental variation... so as to estimate more credibly structural models capable of richer policy analysis.... Experimental variation can help identify economic effects under more general conditions than the observational data, while the structural model can help provide an interpretation of the experimental results and broaden the usefulness of the experiment” (p. 39). In the case of observational data, reporting descriptive statistics like regression coefficients can help readers evaluate the causal interpretation of key relationships in the data, which in turn may affect readers’ views of the credibility of structural estimates.

While this logic has intuitive appeal, the formal sense in which it holds is typically not made precise. Let $\hat{\gamma}$ denote a vector of descriptive statistics, and let \hat{c} denote a structural estimate of interest, which may depend on $\hat{\gamma}$ as well as other features of the data. Under the model assumed by the researcher, \hat{c} and $\hat{\gamma}$ are consistent for values (c_0, γ_0) . A reader is concerned about the possibility that the model may be misspecified. Should the knowledge that $\hat{\gamma}$ is indeed a valid estimator of γ_0 —for example, because it comes from a randomized experiment and the economic assumptions linking γ_0 to c_0 are correct—increase the reader’s confidence in the validity of \hat{c} ? If so, to what extent?

In this paper, we introduce a framework for answering these questions. We consider a reader who entertains a class of potential true models in a neighborhood of the researcher’s assumed model. We follow a large literature in focusing on local misspecification that shrinks with the sample size, so the effect of misspecification remains on the same order as sampling variation, and

thus may lead to asymptotic bias in the parameter estimates. We show that the space of such *local perturbations* can be indexed so that each element is defined by a direction φ (capturing the nature of the misspecification) and a magnitude μ (capturing the degree of misspecification). Our main object of interest is the comparison of (i) the maximal asymptotic bias across all local perturbations of magnitude μ and (ii) the maximal asymptotic bias across the subset of these perturbations under which the descriptive statistics $\hat{\gamma}$ are asymptotically unbiased estimators of γ_0 . The ratio of (ii) to (i) captures the extent to which the reader’s confidence in the structural estimate \hat{c} is improved by knowing that the statistics $\hat{\gamma}$ are unbiased for γ_0 .

Our main result shows that this ratio depends only on a scalar Δ , 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. The ratio of (ii) to (i) is given by $\sqrt{1 - \Delta}$. Intuitively, when informativeness is high, $\hat{\gamma}$ captures most of the information in the data that drives \hat{c} , so knowing that the former is correctly described by the model significantly reduces the scope for bias in the latter. When informativeness is low, \hat{c} is mainly driven by features of the data orthogonal to $\hat{\gamma}$, so lack of bias in $\hat{\gamma}$ does not meaningfully reduce the scope for bias in \hat{c} . We recommend that researchers report an estimate of informativeness whenever they present descriptive evidence as support for structural estimates.

Informativeness can be estimated at low cost even for computationally challenging models. We show that a consistent estimator of Δ can be obtained from manipulation of the estimated influence functions of \hat{c} and $\hat{\gamma}$. In the large range of settings in which estimated influence functions are available from the calculations used to obtain \hat{c} and $\hat{\gamma}$, the additional computation required to estimate Δ is trivial. This tractability is a benefit of our focus on local perturbations.

Importantly, the value of Δ , and the consistency of the proposed estimator for Δ , are preserved under all local perturbations. Moreover, we show that the ranking of local perturbations implied by the magnitude μ coincides asymptotically with that implied by all divergences in the Cressie-Read (1984) family, including the Kullback-Leibler divergence. Hence, μ has a natural scale, and a reader need not take a strong stand on the perturbation of interest, or on the notion of magnitude, in order to interpret and estimate Δ . For readers interested in using our results to construct quantitative bounds on asymptotic bias, we provide an interpretation of the magnitude μ in terms of the power of the most powerful test of the researcher’s base model against the perturbation. We also show a sense in which Δ approximates its non-local analogue, when the degree of misspecification is

small.

The set of local perturbations of a given magnitude μ defines a class of potential true models from a purely statistical or probabilistic point of view, without reference to any particular economic setting. We argue that this class provides a useful and tractable default for the analysis of misspecification that can be readily applied in a wide range of settings of interest. We note, however, that a researcher who is prepared to impose restrictions (say, from economic theory) on the likely form of misspecification may be able to improve on the bounds we characterize. We also stress that the economic content of the restriction that $\hat{\gamma}$ is asymptotically unbiased for γ_0 depends on the setting. Informativeness Δ helps a reader to use her prior information about the plausibility of this restriction to assess the credibility of an estimator, but high or low values of informativeness do not on their own imply high or low credibility.

As a more concrete illustration of our proposed approach, consider an example based on Atanasio et al. (2012a), where a parametric structural model is estimated by maximum likelihood using individual-level data from a randomized experiment. The randomized treatment is an incentive offered to parents to keep their children in school, where the amount of the incentive varies according to characteristics such as the children's age and grade. The descriptive statistics $\hat{\gamma}$ are treatment-control differences in school attendance for different age-grade cells. The structural estimate \hat{c} is the impact of a counterfactual policy that would change the allocation of the incentives. This estimate \hat{c} is related to $\hat{\gamma}$, but since it is estimated by maximum likelihood, it may depend on other features of the data as well. How should believing that the experimental treatment effect estimates $\hat{\gamma}$ are well-specified affect a reader's confidence in the counterfactual?

Our informativeness measure offers a precise answer. At one extreme, we could imagine that the structural estimate \hat{c} depends only on $\hat{\gamma}$, as it would if it were estimated by indirect inference based on $\hat{\gamma}$ alone. In this case, \hat{c} is equal to some function $\tilde{c}(\hat{\gamma})$, and informativeness is $\Delta = 1$. So long as the reader believes that the randomization was carried out correctly and the assumptions mapping γ to c through the function $\tilde{c}(\cdot)$ are correct, she can be confident in the structural estimate. At the other extreme, we could imagine that the structural estimates are completely unrelated to the sample treatment-control differences. In this case, informativeness would be $\Delta = 0$, and confidence in the validity of the experiment and the way the model relates γ to c would not translate into confidence in the structural estimate.

This example highlights an important subtlety. Informativeness captures the extent to which knowing that $\hat{\gamma}$ is correctly described by the model limits the potential bias due to misspecification.

Here, correctly described means that $\hat{\gamma}$ is an asymptotically unbiased estimator of the value γ_0 predicted by the model at base parameter values consistent with c_0 . That $\hat{\gamma}$ is an estimated treatment effect from a randomized experiment, or from a valid natural experiment, is not on its own sufficient for this condition to hold. It must also be the case that the population value of this treatment effect aligns with the prediction γ_0 of the model. In the example above, this means that the function $\tilde{c}(\cdot)$ is correctly specified. The need for this requirement should be clear. If we use an economic model to map from observed data to predictions of counterfactuals that are never observed, statistical information alone cannot confirm or challenge those predictions. Thus, knowledge of Δ does not eliminate the need to think about the validity of structural assumptions. It does, however, allow the reader to judge whether assessing a subset of those assumptions—those that determine the behavior of $\hat{\gamma}$ —is sufficient to form a good sense of the credibility of the estimate \hat{c} .

Our results are related to Andrews et al. (2017). In that paper, we propose a measure Λ of the *sensitivity* of a parameter estimate \hat{c} to a vector of statistics $\hat{\gamma}$, focusing on the case where $\hat{\gamma}$ are estimation moments that fully determine the estimator \hat{c} (and so $\Delta = 1$). Here, we propose a complementary measure of the extent to which a vector of descriptive statistics $\hat{\gamma}$ determines the value of \hat{c} .¹ In an extension, we generalize our main result to allow $\hat{\gamma}$ to have possibly nonzero asymptotic bias $\bar{\gamma}$. In this case, the set of possible asymptotic biases in \hat{c} is an interval centered at $\Lambda\bar{\gamma}$, with width proportional to $\sqrt{1 - \Delta}$. This extension thus provides a unified treatment of sensitivity and informativeness.

We implement our proposal for three recent papers in economics, each of which reports or discusses descriptive statistics alongside structural estimates. In the first application, to Attanasio et al. (2012a), the parameter estimate \hat{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 parameter estimate \hat{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 parameter estimates \hat{c} of interest are estimates of a model parameter and of a key quantity governing the existence of insurance markets, and the descriptive statistics $\hat{\gamma}$ summarize the joint distribution of self-reported probabilities of loss events and the realizations of these events. In each case, our choice of $\hat{\gamma}$ is guided by the authors' discussion. In each case, we report an estimate of Δ for various combinations of \hat{c} and $\hat{\gamma}$, and we

¹Our work here draws on the analysis of “sensitivity to descriptive statistics” in Gentzkow and Shapiro (2015).

discuss the implications for a reader’s confidence in the conclusions. These applications illustrate how estimates of Δ can be presented and discussed in applied research.

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

Our work is also closely related to the large literature on local misspecification (e.g., Newey 1985; Conley et al. 2012; Andrews et al. 2017). Much of this literature focuses on testing and confidence set construction (e.g. Berkowitz et al. 2008; Guggenberger 2012; Armstrong and Kolesar, 2018) 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 the same form of local misspecification that we consider here. Bonhomme and Weidner (2018) derive minimax robust estimators and accompanying confidence intervals for economic parameters of interest under a form of local misspecification closely related to the one we study. Armstrong and Kolesar (2018) 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 little scope to improve on their procedures by “estimating” the degree of misspecification, motivating a sensitivity analysis. In contrast to this literature, we focus on characterizing the relationship between a set of descriptive statistics and a given structural estimator, with the goal of allowing consumers of research to sharpen their opinions about the reliability of the researcher’s conclusions.

Finally, our work relates to discussions about the appropriate role of descriptive statistics in structural econometric analysis (e.g., Pakes 2014).² It is common in applied research to describe the data features that “primarily identify” structural parameters or “drive” estimates of those parameters.³ As Keane (2010) and others have noted, such statements are not directly related to the formal notion of identification in econometrics. Their intended meaning is therefore up for grabs. If researchers are prepared to reinterpret these as statements linking correct specification of descriptive statistics to confidence in related structural estimates, then our approach provides a way to sharpen and quantify these statements at low cost to researchers.

²See also Dridi et al. (2007) and Nakamura and Steinsson (2018) for discussion of the appropriate choice of moments to match when fitting macroeconomic models.

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

2 Setup and Main Result

2.1 Setup

A researcher observes an i.i.d. sample $D_i \in \mathcal{D}$ for $i = 1, \dots, n$. The researcher's model implies that $D_i \sim F(\eta)$, for $\eta \in H$ a potentially infinite-dimensional parameter. The implied distribution for the sample is $F^n(\eta) = \times_n F(\eta)$. We fix a *base distribution* $F_0 = F(\eta_0)$ consistent with the model, and let η_0 denote the base value of η .

The key quantity of interest $c = c(\eta)$ is a scalar that may be an element of η , a counterfactual prediction, or any other function of the model's parameters. Because we are interested in model misspecification, we assume that the true value $c_0 = c(\eta_0)$ of c is defined outside the model—for example, from some ideal experiment—and thus that we can meaningfully discuss the true value of c even when the model may be incorrect. Note that we do not exclude the possibility that $c_0 = c(\eta')$ for some $\eta' \neq \eta_0$. The researcher computes (i) an estimate \hat{c} of c and (ii) a $p_\gamma \times 1$ vector of descriptive statistics $\hat{\gamma}$.

We assume that under F_0^n , the estimator \hat{c} and the descriptive statistics $\hat{\gamma}$ are jointly asymptotically normal,

$$(1) \quad \sqrt{n} \begin{pmatrix} \hat{c} - c_0 \\ \hat{\gamma} - \gamma_0 \end{pmatrix} \rightarrow_d N(0, \Sigma),$$

where γ_0 is the probability limit of $\hat{\gamma}$ under F_0^n and the asymptotic variance Σ is finite. We assume throughout that the asymptotic variances σ_c^2 and $\Sigma_{\gamma\gamma}$ corresponding to \hat{c} and $\hat{\gamma}$ are positive and full-rank, respectively. We let $\Sigma_{c\gamma}$ denote the off-diagonal block of Σ corresponding to the asymptotic covariance of \hat{c} and $\hat{\gamma}$.

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

$$\Delta = \frac{\Sigma_{c\gamma} \Sigma_{\gamma\gamma}^{-1} \Sigma_{c\gamma}'}{\sigma_c^2}.$$

Informativeness measures the extent to which variation in \hat{c} is explained by variation in $\hat{\gamma}$. It corresponds to the R^2 from the population regression of \hat{c} on $\hat{\gamma}$ in their joint asymptotic distribution under F_0^n . It is immediate that $\Delta \in [0, 1]$.

To introduce the possibility that the model may be misspecified, we follow a standard approach

in the literature (e.g., Newey 1985; Andrews et al. 2017) and consider misspecification that is on the same order as sampling uncertainty. We consider sequences of alternative models that approach the base distribution at a \sqrt{n} rate. We call such a sequence a *local perturbation*.⁴ We index each local perturbation by its direction $\varphi \in \Phi$ and magnitude $\mu \in \mathbb{R}$. The magnitude μ is defined so that a perturbation of magnitude $\mu = 0$ coincides with the base distribution, and so that the ranking of perturbations implied by μ coincides asymptotically with the ranking based on Kullback-Liebler divergence. In Section 3.2 we show that this ranking also coincides asymptotically with that implied by a large class of divergences, and that μ has an interpretation in terms of the most powerful test of the base distribution against a perturbation.

More precisely, for each direction φ in the possibly infinite-dimensional set Φ , we define a family of distributions $F_\varphi(\vartheta)$ indexed by a scalar parameter $\vartheta \in \mathbb{R}_+$ such that $F_\varphi(0) = F_0$. Each $F_\varphi(\cdot)$ defines a one-dimensional family of distributions, which one can think of as a path passing through the base distribution. The local perturbation with direction φ and magnitude μ is then defined to be the sequence of joint distributions $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right) = \times_n F_\varphi\left(\frac{\mu}{\sqrt{n}}\right)$. When $\mu > 0$, the model may be misspecified in the sense that $F_\varphi\left(\frac{\mu}{\sqrt{n}}\right) \notin \{F(\eta) : c(\eta) = c_0, \eta \in H\}$, or even $F_\varphi\left(\frac{\mu}{\sqrt{n}}\right) \notin \{F(\eta) : \eta \in H\}$. Thus the true distribution of the data may be inconsistent with the restrictions implied by c_0 under our assumed model, or inconsistent with the model altogether.

The interpretation of the values c_0 , η_0 , and γ_0 in the presence of misspecification is important in what follows, and it is worth pausing to reiterate. The true value c_0 of c is assumed to be an economic quantity defined outside the model, and so it is meaningful to ask how well \hat{c} estimates c_0 under misspecification. By contrast, η need not be defined outside the model—for example, it might include coefficients whose values are only defined relative to the distribution of various error terms—and so we do not interpret η_0 as a true value. Instead, our local misspecification approach assumes that the true data generating process is close to some (unknown) base distribution $F_0 = F(\eta_0)$ consistent with the assumed model and with the true value c_0 (in the sense that $c_0 = c(\eta_0)$). Likewise, γ_0 denotes the probability limit of $\hat{\gamma}$ under the base distribution, while γ may or may not be defined outside the model. We hold c_0 , η_0 , and γ_0 fixed for simplicity, but show in Appendix B.1 that our results are robust to instead considering $(\eta_{0,n}, c_{0,n}, \gamma_{0,n})$ that converge to (η_0, c_0, γ_0) .

Example. Say that η is finite-dimensional and that $\hat{\gamma}$ is a vector of regression coefficients estimated on the sample, such as treatment-control differences from a randomized experiment. The quantity

⁴Rieder (1994) defines a subset of these perturbations called “simple perturbations” and derives results for this class.

c describes an unobserved counterfactual and is estimated with $\hat{c} = c(\hat{\eta})$ where $\hat{\eta}$ is an estimator of η that is jointly asymptotically normal with \hat{c} and $\hat{\gamma}$. We can now consider three cases:

1. (Minimum distance) The estimator $\hat{\eta}$ can be written as $\hat{\eta} = \eta(\hat{\gamma}) + o_p(1/\sqrt{n})$ where η is a nontrivial, smooth function. This is the case when, for example, $\hat{\eta}$ is estimated via classical minimum distance, or simulated analogues such as indirect inference, using the descriptive statistics $\hat{\gamma}$ as the target moments. In this case, under sufficient regularity we will have $\Delta = 1$.
2. (MLE) The estimator $\hat{\eta}$ is the maximum likelihood estimator of the model. In many cases this estimator cannot be expressed asymptotically as a function only of $\hat{\gamma}$, and in such cases $\Delta < 1$.
3. (Irrelevant descriptive statistics) The estimator $\hat{\eta}$ can be written as a function of a set of moments that are statistically independent of $\hat{\gamma}$, say because $\hat{\eta}$ is estimated using data from one population (e.g., males) and $\hat{\gamma}$ is estimated using data from another (e.g., females). In this case $\Delta = 0$.

2.2 Main Result

We introduce two conditions that are sufficient for our main result. In Section 3, we revisit these conditions and give primitive assumptions under which they are guaranteed to hold.

Condition 1. Under a local perturbation with direction φ and magnitude μ ,

$$(2) \quad \sqrt{n} \begin{pmatrix} \hat{c} - c_0 \\ \hat{\gamma} - \gamma_0 \end{pmatrix} \rightarrow_d N \left(\begin{pmatrix} \mu \bar{c}_\varphi \\ \mu \bar{\gamma}_\varphi \end{pmatrix}, \Sigma \right),$$

for some $\bar{c}_\varphi \in \mathbb{R}$ and $\bar{\gamma}_\varphi \in \mathbb{R}^{p_\gamma}$, and for Σ the same as in (1).

The term $\mu \bar{c}_\varphi$ is the *first-order asymptotic bias* (“asymptotic bias,” for short) of the estimator \hat{c} under a given local perturbation. Because Σ is unaffected by the perturbation, the asymptotic bias is the only asymptotic effect of misspecification on the researcher’s conclusions. Our main goal in this section is to bound the size of this bias. Condition 1 implies that this bias is the product of the magnitude μ of the perturbation and a term \bar{c}_φ that depends only on the perturbation’s direction.

Condition 2. If $\Delta < 1$, the set of values $(\bar{c}_\varphi, \bar{\gamma}_\varphi)$ associated with the set of all directions $\varphi \in \Phi$ is

$$(3) \quad \mathcal{B} = \left\{ (\bar{c}, \bar{\gamma}) \in \mathbb{R} \times \mathbb{R}^{p_\gamma} : \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} \leq 1 \right\}.$$

If $\Delta = 1$, this set of values is

$$(4) \quad \mathcal{B} = \left\{ (\bar{c}, \bar{\gamma}) \in \mathbb{R} \times \mathbb{R}^{p_\gamma} : \bar{c} = \Lambda \bar{\gamma}, \bar{\gamma}' \Sigma_{\gamma\gamma}^{-1} \bar{\gamma} \leq 1 \right\},$$

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

The set defined by (3) is an ellipsoid. Figure 1 shows an example for the case in which $p_\gamma = 1$. As this figure illustrates, the range of possible values of \bar{c}_φ grows tighter the more we can restrict the value of $\bar{\gamma}_\varphi$. Our main question of interest is to what extent the bounds on \bar{c}_φ are tightened by assuming $\bar{\gamma}_\varphi = 0$. This is illustrated in the figure by the comparison of the length of the inner interval (“ $\bar{\gamma}_\varphi = 0$ ”) to the length of the outer interval (“ $\bar{\gamma}_\varphi$ unconstrained”). When $\Delta = 1$, this ellipsoid collapses and the value of \bar{c}_φ is fully determined by $\bar{\gamma}_\varphi$, with $\bar{c}_\varphi = \Lambda \bar{\gamma}_\varphi$ for a vector of coefficients Λ . We discuss the role of Λ at length in Section 4.

Our main result shows that in the general case, the analogues of the two intervals in Figure 1 are simple functions of informativeness Δ , the magnitude of the perturbation μ , and the standard error σ_c , and that the ratio of the widths of the intervals depends only on Δ . Proofs for this, and other results, are in Appendix A.

Proposition 1. *Suppose Conditions 1 and 2 hold, and consider the set of local perturbations with magnitude μ . The asymptotic biases in \hat{c} associated with perturbations in this set are*

$$\mathcal{B}^\mu = [-\sigma_c \mu, \sigma_c \mu].$$

The asymptotic biases in \hat{c} associated with the subset of these perturbations for which $\bar{\gamma}_\varphi = 0$ are

$$\mathcal{B}_0^\mu = \left[-\sigma_c \mu \sqrt{1 - \Delta}, \sigma_c \mu \sqrt{1 - \Delta} \right].$$

The ratio of the widths of these intervals is

$$\frac{|\mathcal{B}_0^\mu|}{|\mathcal{B}^\mu|} = \sqrt{1 - \Delta}.$$

Thus, the extent to which knowing that $\bar{\gamma}_\phi = 0$ constrains the potential bias due to misspecification depends on the informativeness of $\hat{\gamma}$ for the outcome of interest. When $\Delta = 0$, knowing that $\hat{\gamma}$ is asymptotically unbiased for γ_0 tells us nothing about potential bias in \hat{c} . When $\Delta = 1$, knowing that $\hat{\gamma}$ is asymptotically unbiased for γ_0 implies that there is no bias in \hat{c} under *any* local perturbation, regardless of its direction or magnitude.

Figure 2 illustrates the conclusions of Proposition 1 by plotting the limits of the intervals \mathcal{B}^μ and \mathcal{B}_0^μ (normalized with respect to the standard error σ_c of \hat{c}) as a function of μ . The figure is drawn for the case $\Delta = 0.75$. When $\mu = 2$, the unconstrained range of asymptotic biases is plus or minus two standard errors. Knowing that $\hat{\gamma}$ is asymptotically unbiased for γ_0 reduces the width of this interval to plus or minus $2\sqrt{1-0.75} = 1$ standard error. We return to the interpretation of μ in Section 3.2 below.

We propose that researchers interested in linking a structural estimate \hat{c} to a vector of descriptive statistics $\hat{\gamma}$ report an estimate of informativeness Δ . We show below that it can be trivial to estimate Δ even when \hat{c} is costly to compute. Having an estimate of informativeness permits a reader to gauge the extent to which confidence in the model’s predictions about $\hat{\gamma}$ translate into confidence in the researcher’s inferences about c_0 .⁵

Example. (continued) Recall the three cases from the earlier example:

1. (Minimum distance) Here, $\Delta = 1$. Therefore by Proposition 1, if $\hat{\gamma}$ is correctly specified ($\bar{\gamma}_\phi = 0$), then there is no asymptotic bias in \hat{c} ($\bar{c}_\phi = 0$). Intuitively, this is because $\bar{\gamma}_\phi = 0$ implies both that $\hat{\gamma}$ is asymptotically unbiased for γ_0 and that the relation $c_0 = c(\eta(\gamma_0))$ is correctly specified.
2. (MLE) Here, $\Delta < 1$. Therefore, by Proposition 1, $\frac{|\mathcal{B}_0^\mu|}{|\mathcal{B}^\mu|} = \sqrt{1-\Delta} > 0$. Knowing that $\bar{\gamma}_\phi = 0$ still leaves open the possibility that \hat{c} is asymptotically biased. Intuitively, this is because \hat{c} depends on information in the data that is not fully captured in $\hat{\gamma}$. Note that this is true even if there exists some relation $c_0 = c(\eta(\gamma_0))$ that is correctly specified as in the preceding case, and hence estimation could in principle have been based only on $\hat{\gamma}$.
3. (Irrelevant descriptive statistics) Here, $\Delta = 0$. Therefore, by Proposition 1, knowing that $\bar{\gamma}_\phi = 0$ does not restrict \bar{c}_ϕ at all. Intuitively, this is because \hat{c} depends entirely on features

⁵If we impose an upper bound on μ , we can use the bounds in Proposition 1 to directly adjust our confidence intervals for misspecification, with or without the assumption that $\bar{\gamma}_\phi = 0$; see Armstrong and Kolesar (2018) for a discussion of such adjustments.

of the data unrelated to $\hat{\gamma}$.

3 Regularity Conditions and Interpretation of Perturbations

3.1 Regularity Conditions

This subsection develops asymptotic results that provide the foundation for our analysis. In particular, we introduce assumptions that imply Conditions 1 and 2.

Perturbations

We assume the distributions $F_\varphi(\vartheta)$ have densities $f_\varphi(\vartheta)$ with respect to a dominating measure ν . That is, sets that are assigned zero mass by ν are likewise assigned probability zero under all $F_\varphi(\vartheta)$. Note that ν need not be Lebesgue measure, so we do not require that the data be continuously distributed. The information matrix for ϑ , treating φ as known, is

$$I_\varphi(\vartheta) = E_{F_\varphi(\vartheta)} \left[\left(\frac{\partial}{\partial \vartheta} f_\varphi(D_i; \vartheta) \right)^2 \right].$$

We impose the following regularity condition on the paths $\{F_\varphi(\cdot) : \varphi \in \Phi\}$:

Assumption 1. *For ϑ in an open neighborhood of zero and all $\varphi \in \Phi$; (i) $\sqrt{f_\varphi(d; \vartheta)}$ is continuously differentiable with respect to ϑ for all $d \in \mathcal{D}$; (ii) $I_\varphi(\vartheta)$ is finite and continuous in ϑ ; and (iii) $I_\varphi(0) \in \{0, 1\}$.*

Parts (i) and (ii) of Assumption 1 are standard conditions used in deriving asymptotic results, and hold in a wide variety of settings; see Chapter 7.2 in van der Vaart (1998) for further discussion. Part (iii) is a normalization which, for ϑ close to zero, allows us to interpret the magnitude of ϑ as a measure of the degree of misspecification.⁶

We consider local perturbations $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$, which imply that the degree of misspecification is of the same order as the Cramer-Rao lower bound on the standard deviation for unbiased estimators of ϑ , provided the usual regularity conditions for this bound hold.⁷ As we show in Section 3.2

⁶Formally, if we begin with a perturbation $\tilde{F}_\varphi(\tilde{\vartheta})$ satisfying parts (i) and (ii) of Assumption 1 but not part (iii), and define $F_\varphi(\vartheta) = \tilde{F}_\varphi\left(\vartheta/\sqrt{\tilde{I}_\varphi(0)}\right)$ for $\tilde{I}_\varphi(\tilde{\vartheta})$ the information matrix for $\tilde{\vartheta}$, then $F_\varphi(\cdot)$ satisfies parts (i)-(iii).

⁷When the information matrix is zero, this lower bound is infinite.

below, this in turn implies that the magnitude μ has an interpretation as a measure of the degree of misspecification based both on widely studied measures for the divergence between distributions and on the power of tests to detect this local perturbation.

Our second assumption requires that the set of perturbations be sufficiently rich.

Assumption 2. *The set of score functions $s_\varphi(D_i) = \frac{\partial}{\partial \vartheta} (\log f_\varphi(D_i; 0))$ includes all those consistent with Assumption 1, in the sense that for any $s(\cdot)$ with $E_{F_0}[s(D_i)] = 0$ and $E_{F_0}[s(D_i)^2] \in \{0, 1\}$ there exists $\varphi \in \Phi$ with $E_{F_0} \left[\left(s(D_i) - \frac{\partial}{\partial \vartheta} \log(f_\varphi(D_i; 0)) \right)^2 \right] = 0$.*

Assumption 2 requires that Φ imply a rich enough set of $F_\varphi(\cdot)$ that the corresponding score functions include all those consistent with Assumption 1.⁸ While allowing a rich set of perturbations is important to capture the many ways in which the model could be misspecified, if in a particular setting we know more about the form of misspecification, this will restrict Φ . In such settings, the bounds we derive below will be valid, but may no longer be tight.

Estimators

Our key assumption on the estimators \hat{c} and $\hat{\gamma}$ is that they behave, asymptotically, like averages of functions over the observations.

Assumption 3. *Under F_0^n ,*

$$(5) \quad \sqrt{n}(\hat{c} - c_0, \hat{\gamma} - \gamma_0) = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n \phi_c(D_i), \sum_{i=1}^n \phi_\gamma(D_i) \right) + o_p(1),$$

for functions $\phi_c(D_i)$ and $\phi_\gamma(D_i)$ such that $E_{F_0}[\phi_c(D_i)] = 0$, $E_{F_0}[\phi_\gamma(D_i)] = 0$, $E_{F_0}[\phi_c(D_i)^2] = \sigma_c^2$, and $E_{F_0}[\phi_\gamma(D_i)\phi_\gamma(D_i)'] = \Sigma_{\gamma\gamma}$, for σ_c^2 and $\Sigma_{\gamma\gamma}$ finite. Moreover, $p_\gamma + 1 < |\text{supp}_{F_0}(D_i)|$ for $\text{supp}_{F_0}(D_i)$ the support of D_i under F_0 .

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 (5) holds for an extremely wide range of estimators (see e.g. Ichimura and Newey 2015). This representation will be the key to our analysis in this section and the next, and immediately implies that \hat{c} and $\hat{\gamma}$ are jointly asymptotically normal

⁸That the score function $s(D_i)$ for any perturbation satisfying Assumption 1 has mean zero follows from Lemma 7.6 and Theorem 7.2 in van der Vaart (1998).

under F_0 as in (1). Sufficient conditions for asymptotic linearity when \hat{c} and $\hat{\gamma}$ are based on minimum distance estimators are given in Section 5. The second part of the assumption requires that the dimension of $(c, \gamma)'$ be smaller than the cardinality of the support of D_i , and holds trivially if D_i has at least one continuously distributed component. If this condition fails, our bounds remain valid but some interior values for the bias may not be achievable.

Verifying Conditions 1 and 2

We next show that Assumptions 1-3 imply Conditions 1 and 2.

Lemma 1. *Assumptions 1 and 3 imply Condition 1. In particular, under $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$*

$$\sqrt{n} \begin{pmatrix} \hat{c} - c_0 \\ \hat{\gamma} - \gamma_0 \end{pmatrix} \rightarrow_d N \left(\begin{pmatrix} \mu \bar{c}_\varphi \\ \mu \bar{\gamma}_\varphi \end{pmatrix}, \Sigma \right).$$

where

$$\begin{pmatrix} \bar{c}_\varphi \\ \bar{\gamma}_\varphi \end{pmatrix} = \begin{pmatrix} E_{F_0} [\phi_c(D_i) s_\varphi(D_i)] \\ E_{F_0} [\phi_\gamma(D_i) s_\varphi(D_i)] \end{pmatrix},$$

and

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

Lemma 1 shows that Assumptions 1 and 3 imply Condition 1, and characterizes the form of the bias. As in Andrews et al. (2017) the asymptotic variance is unaffected by the local perturbation $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$. The asymptotic distributions of $\sqrt{n}(\hat{c} - c_0)$ and $\sqrt{n}(\hat{\gamma} - \gamma_0)$, on the other hand, are shifted by $\mu \bar{c}_\varphi$ and $\mu \bar{\gamma}_\varphi$ respectively, where the effect on each estimator depends on the covariance of its influence function with the score.⁹ We next verify Condition 2.

Lemma 2. *For $(\bar{c}_\varphi, \bar{\gamma}_\varphi)$ and Σ as defined in Lemma 1, Assumption 2 implies Condition 2.*

Lemma 2 shows that Assumption 2 implies Condition 2. Without Assumption 2 we can still show that the left hand side of (3) is a subset of the right hand side, but it may be strictly contained, in which case our bounds on bias may not be tight.

⁹Lemma 1 considers behavior under local perturbations $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$. Rieder (1994, Chapter 6) provides stronger conditions on the estimators \hat{c} and $\hat{\gamma}$ such that an analogous result holds uniformly over shrinking nonparametric neighborhoods of F_0 .

3.2 Scaling of Perturbations

We have interpreted the parameter μ as a measure of the “size” of the perturbation $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$. Condition 1 and Lemma 1 show that this interpretation is reasonable for fixed φ , since the shift in the asymptotic distribution of $(\hat{c}, \hat{\gamma})$ scales with μ . For Proposition 1, however, we consider all perturbations consistent with a given μ , so it is important that μ measures the degree of misspecification in a way that is meaningful across values φ as well.

This section shows that this interpretation is justified under two natural measures for the degree of misspecification. First, we show that μ determines the highest possible asymptotic power for any test to distinguish the perturbation $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$ from the base distribution F_0^n . Second, we show that for a family of divergences measuring the difference between $F_\varphi\left(\frac{\mu}{\sqrt{n}}\right)$ and F_0 , including Kullback-Leibler divergence and squared Hellinger divergence, the scaled divergence converges to μ^2 under mild regularity conditions.

Throughout this section we limit attention to non-trivial perturbations, which we define as those with φ such that $E_{F_0}\left[s_\varphi(D_i)^2\right] \neq 0$, for $s_\varphi(D_i)$ the score as defined in Assumption 2. For trivial perturbations one can show (see the proof of Proposition 5) that $\bar{c}_\varphi = \bar{\gamma}_\varphi = 0$, so the perturbation has no first-order asymptotic effect on $(\hat{c}, \hat{\gamma})$. Trivial perturbations are likewise asymptotically indistinguishable from the base distribution according to both measures we consider in this section.

Asymptotic Distinguishability

One natural measure for the size of the perturbation $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$ is the power of tests of F_0^n against $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$.

Proposition 2. *Under Assumption 1 the most powerful level α test of the null hypothesis $H_0 : (D_1, \dots, D_n) \sim F_0^n$ against the non-trivial alternative $H_1 : (D_1, \dots, D_n) \sim F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$ has asymptotic power $1 - F_{N(0,1)}(c_\alpha - \mu)$ for c_α the $1 - \alpha$ quantile of the standard normal distribution.*

The proof of Proposition 2 shows that the most powerful test corresponds asymptotically to a z-test, where the z-statistic has mean μ under H_1 . Hence, we can interpret a value of $\mu = 1$ as corresponding to a one-standard-error shift in the z-statistic, putting μ on a conventional scale.

The optimal test derived in the proof of Proposition 2 relies on knowledge of F_0 and so is infeasible in practice. However, the problem of testing F_0^n against $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$ is statistically easier than the problem of testing $\{F^n(\eta) : c(\eta) = c_0, \eta \in H\}$ or $\{F^n(\eta) : \eta \in H\}$ against $F_\varphi^n\left(\frac{\mu}{\sqrt{n}}\right)$.

Hence, the power bound in Proposition 2 is an upper bound on the power of tests of either the hypothesis that the data generating process is consistent with the model and c_0 or the hypothesis that the model is correctly specified without restrictions on c_0 .

Asymptotic Divergence

Divergence measures provide an alternative quantification for the difference between F_0 and $F_\varphi\left(\frac{\mu}{\sqrt{n}}\right)$, and hence for the degree of misspecification. We consider divergences of the form

$$(6) \quad r(F_0, F_\varphi(\vartheta)) = E_{F_0} \left[\psi \left(\frac{dF_\varphi(\vartheta)}{dF_0} \right) \right] = E_{F_0} \left[\psi \left(\frac{f_\varphi(D_i; \vartheta)}{f_\varphi(D_i; 0)} \right) \right]$$

for $\psi(\cdot)$ a function. We impose the following condition on $\psi(\cdot)$.

Assumption 4. *The function $\psi(\cdot)$ is twice continuously differentiable with $\psi(1) = 0$ and $\psi''(1) = 2$.*

Given a twice continuously differentiable function $\tilde{\psi}(\cdot)$ with $\tilde{\psi}''(1) \neq 0$, there always exist (a, b) so that $\psi(\cdot) = a + b \cdot \tilde{\psi}(\cdot)$ satisfies Assumption 4. Hence, in addition to imposing differentiability the role of Assumption 4 is to fix the level and scale of $\psi(\cdot)$.

A common class of divergences is 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 was recently used to measure the degree of misspecification by Hansen and Sargent (2016), while Hellinger divergence was used by Kitamura et al. (2013). A Cressie-Read (1984) divergence between F_0 and $F_\varphi(\vartheta)$ takes

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

Remark. All of the Cressie-Read (1984) divergences satisfy Assumption 4.

We are interested in the divergence between F_0 and $F_\varphi\left(\frac{\mu}{\sqrt{n}}\right)$ in large samples, and impose the following regularity conditions.

Assumption 5. *$f_\varphi(D_i; \vartheta)$ is twice continuously differentiable in ϑ at 0, and there exists an open neighborhood \mathcal{N}_ϑ of zero such that*

$$E_{F_0} \left[\sup_{\vartheta \in \mathcal{N}_\vartheta} \left(\left| \frac{\partial}{\partial \vartheta} f_\varphi(D_i; \vartheta) \right| + \left| \frac{\partial^2}{\partial \vartheta^2} f_\varphi(D_i; \vartheta) \right| + \left| \frac{\frac{\partial}{\partial \vartheta} f_\varphi(D_i; \vartheta)}{f_\varphi(D_i; 0)} \right| \right) \right]$$

and

$$E_{F_0} \left[\sup_{\vartheta \in \mathcal{N}_\vartheta} \left| \psi' \left(\frac{f_\varphi(D_i; \vartheta)}{f_\varphi(D_i; 0)} \right) \frac{\frac{\partial^2}{\partial \mu^2} f_\varphi(D_i; \vartheta)}{f_\varphi(D_i; 0)} + \psi'' \left(\frac{f_\varphi(D_i; \vartheta)}{f_\varphi(D_i; 0)} \right) \left(\frac{\frac{\partial}{\partial \mu} f_\varphi(D_i; \vartheta)}{f_\varphi(D_i; 0)} \right)^2 \right| \right]$$

are finite.

Proposition 3. *Under Assumptions 1, 4, and 5,*

$$\lim_{n \rightarrow \infty} n \cdot r \left(F_0, F_\varphi \left(\frac{\mu}{\sqrt{n}} \right) \right) = \mu^2$$

for all non-trivial perturbations $F_\varphi \left(\frac{\mu}{\sqrt{n}} \right)$.

Proposition 3 shows that nontrivial perturbations have scaled divergence converging to μ^2 , where we scale by n to ensure a nontrivial limit. Hence, in large samples the Cressie-Read (1984) divergences yield the same ranking over perturbations as that implied by μ .

3.3 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 that the data now follow F^n , where F does not change with the sample size, and we may again have $F \notin \{F(\eta) : c(\eta) = c_0, \eta \in H\}$. In this case \hat{c} and $\hat{\gamma}$ will typically be inconsistent for c_0 and γ_0 , so let us denote their probability limits under F^n by $\tilde{c}(F)$ and $\tilde{\gamma}(F)$, respectively. We assume for ease of exposition that these probability limits exist on the neighborhoods of F_0 that we consider.

Suppose that for a divergence r defined as in Section 3.2 we are willing to assume that $r(F_0, F) \leq \bar{r}$ for \bar{r} a known scalar. For $\mathcal{F} = \{F_\varphi(\vartheta) : \vartheta \in \mathbb{R}_+, \varphi \in \Phi\}$ the set of potential distributions for the data under misspecification, the set of possible probability limits for $\hat{c} - c_0$ is

$$\tilde{\mathcal{B}}^{\bar{r}} = \{\tilde{c}(F) - c_0 : F \in \mathcal{F}, r(F_0, F) \leq \bar{r}\}.$$

This is the non-local analogue of the set of asymptotic biases \mathcal{B}^μ defined in Section 2. We can likewise consider the set of probability limits for $\hat{c} - c_0$ under forms of misspecification that do not

affect the probability limit of $\hat{\gamma}$:

$$\tilde{\mathcal{B}}_0^{\bar{r}} = \{\tilde{c}(F) - c_0 : F \in \mathcal{F}, r(F_0, F) \leq \bar{r}, \tilde{\gamma}(F) - \gamma_0 = 0\}.$$

This is the non-local analogue of the set \mathcal{B}_0^μ .

Provided that $|\tilde{\mathcal{B}}^{\bar{r}}|$ and $|\tilde{\mathcal{B}}_0^{\bar{r}}|$ are both finite and non-zero, we can define a non-local analogue $\tilde{\Delta}(\bar{r})$ of informativeness Δ by

$$\sqrt{1 - \tilde{\Delta}(\bar{r})} = \frac{|\tilde{\mathcal{B}}_0^{\bar{r}}|}{|\tilde{\mathcal{B}}^{\bar{r}}|}.$$

Intuitively, $\tilde{\Delta}(\bar{r})$ measures the extent to which, for base distribution F_0 and neighborhood size \bar{r} , limiting attention to forms of misspecification that do not affect $\hat{\gamma}$ limits the scope for inconsistency of the estimator \hat{c} .

Observe that, whereas under our assumptions Δ does not depend on the degree μ of the local perturbation, $\tilde{\Delta}(\bar{r})$ will typically depend on \bar{r} . Moreover, calculation of $\tilde{\Delta}(\bar{r})$ typically requires specifying the base distribution F_0 (since this isn't generally uniquely determined from F and \bar{r}), and may be computationally difficult.

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

4 Sensitivity and Informativeness

Proposition 1 considers the effect of limiting attention to perturbations with $\bar{\gamma}_\varphi = 0$. In some cases, however, researchers may be interested in forms of misspecification with a non-zero, but known, asymptotic effect on $\hat{\gamma}$. In such cases, our assumptions again imply a relationship between the biases in \hat{c} and $\hat{\gamma}$.

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

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

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

To build intuition, note that sensitivity characterizes the relationship between \hat{c} and $\hat{\gamma}$ in the

asymptotic distribution under the base model. Let \tilde{c} and $\tilde{\gamma}$ denote the normally distributed random variables to which $\sqrt{n}(\hat{c} - c_0)$ and $\sqrt{n}(\hat{\gamma} - \gamma_0)$ converge in distribution under F_0^n . The sensitivity Λ is simply the vector of coefficients from the population regression of \tilde{c} on $\tilde{\gamma}$. An element Λ_j of Λ is the effect of changing the realization of a particular $\tilde{\gamma}_j$ on the expected value of \tilde{c} , holding the other elements of $\tilde{\gamma}$ constant.

Andrews et al. (2017) show that for $\hat{c} = c(\hat{\eta})$, $\hat{\eta}$ a minimum distance estimator based on moments $\hat{g}(\eta)$, and $\hat{\gamma} = \hat{g}(\eta_0)$ the estimation moments evaluated at the true parameter value, sensitivity allows us to relate the effect of misspecification on $\hat{\gamma}$ (i.e., $\mu\bar{\gamma}_\phi$) to the effect on \hat{c} (i.e., $\mu\bar{c}_\phi$).

Proposition 4. (Andrews et al. 2017) For $\hat{c} = c(\hat{\eta})$, $\hat{\eta}$ a minimum distance estimator satisfying Assumption 7 below, and $\hat{\gamma} = \hat{g}(\eta_0)$, under Assumption 1 and local perturbations $F_\phi^n\left(\frac{\mu}{\sqrt{n}}\right)$ we have $\mu\bar{c}_\phi = \Lambda\mu\bar{\gamma}_\phi$.

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

Proposition 5. Suppose that Assumptions 1-3 hold, and consider the set of local perturbations with magnitude μ and $\mu\bar{\gamma}_\phi = \bar{\gamma}$. The asymptotic biases in \hat{c} associated with perturbations in this set are

$$\mathcal{B}_\gamma^\mu = \left[\Lambda\bar{\gamma} - \sigma_c\mu(\bar{\gamma})\sqrt{1-\Delta}, \Lambda\bar{\gamma} + \sigma_c\mu(\bar{\gamma})\sqrt{1-\Delta} \right],$$

for $\mu(\bar{\gamma}) = \sqrt{\mu^2 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma}}$, provided $\mu^2 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma} \geq 0$. If $\mu^2 - \bar{\gamma}'\Sigma_{\gamma\gamma}^{-1}\bar{\gamma} < 0$, then $\mathcal{B}_\gamma^\mu = \emptyset$.

The ratio of the width of \mathcal{B}_γ^μ to the width of the interval when $\bar{\gamma}_\phi$ is unconstrained is

$$\frac{|\mathcal{B}_\gamma^\mu|}{|\mathcal{B}^{\mu(\bar{\gamma})}|} = \sqrt{1-\Delta},$$

where $\mathcal{B}^{\mu(\bar{\gamma})} = [-\sigma_c\mu(\bar{\gamma}), \sigma_c\mu(\bar{\gamma})]$ is the set of asymptotic biases associated with all local perturbations of magnitude $\mu(\bar{\gamma})$.

Proposition 5 extends the results of Andrews et al. (2017) to the case where $\hat{\gamma}$ need not be a vector of estimation moments, and thus we may have $\Delta < 1$. It likewise extends Proposition 1 to perturbations with a non-zero effect on $\hat{\gamma}$, so $\mu\bar{\gamma}_\phi \neq 0$. The resulting set of first-order asymptotic biases $\mu\bar{c}_\phi$ for \hat{c} is centered at $\Lambda\bar{\gamma}$ with width proportional to $\sqrt{1-\Delta}$.

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

The second part of the proposition compares the range of biases under perturbations with magnitude μ and first-order bias $\bar{\gamma}$ for $\hat{\gamma}$ with the range of biases under unrestricted perturbations with magnitude $\mu(\bar{\gamma})$, and finds that the ratios of their widths is $\sqrt{1-\Delta}$. This result highlights that the degree of excess misspecification $\mu(\bar{\gamma})$ plays the same role in the class of perturbations with a known nonzero effect on $\hat{\gamma}$ as the degree of misspecification μ plays in the class of perturbations with either a zero or unrestricted effect on $\hat{\gamma}$.

Example. (continued) Recall the three examples from Section 2.1.

1. (Minimum distance) Here Λ is the sensitivity measure of Andrews et al. (2017). Since $\Delta = 1$, Proposition 5 (or the main result of Andrews et al. 2017) implies that $\mu\bar{c}_\phi = \Lambda\mu\bar{\gamma}_\phi$. Hence, if we know the asymptotic effect of model misspecification on $\hat{\gamma}$, we know its effect on \hat{c} as well.
2. (MLE) Here Λ will depend on the definition of \hat{c} . Since $\Delta < 1$, however, the value of $\mu\bar{\gamma}_\phi$ does not fully determine the asymptotic bias of \hat{c} . The range of asymptotic biases for \hat{c} is centered at $\Lambda\mu\bar{\gamma}_\phi$, and has width that depends on Δ and the degree of excess misspecification $\mu(\bar{\gamma})$.
3. (Irrelevant descriptive statistics) Here $\Lambda = 0$. Since $\Delta = 0$ as well, knowing that $\mu\bar{\gamma}_\phi = \bar{\gamma}$ does not restrict the asymptotic bias of \hat{c} except through $\mu(\bar{\gamma})$.

5 Implementation

Given an estimate $\hat{\Sigma}$ of Σ , it is straightforward to derive estimates $\hat{\Delta}$ and $\hat{\Lambda}$ of informativeness and sensitivity, respectively. In a wide range of applications, convenient estimates of Σ under F_0^n are available following standard asymptotic results (e.g., Newey and McFadden 1994). Given such an

estimate one can construct plug-in estimates

$$(7) \quad \hat{\Delta} = \frac{\hat{\Sigma}_{c\gamma} \hat{\Sigma}_{\gamma\gamma}^{-1} \hat{\Sigma}'_{c\gamma}}{\hat{\sigma}_c^2}, \quad \hat{\Lambda} = \hat{\Sigma}_{c\gamma} \hat{\Sigma}_{\gamma\gamma}^{-1}.$$

Provided $\hat{\Sigma}$ is consistent under F_0^n , consistency of $\hat{\Sigma}$, $\hat{\Delta}$, and $\hat{\Lambda}$ under any local perturbation follows immediately under our maintained assumptions that $\sigma_c^2 > 0$ and $\Sigma_{\gamma\gamma}$ has full rank.

Assumption 6. $\hat{\Sigma} \xrightarrow{p} \Sigma$ under F_0^n .

Proposition 6. Under Assumptions 1 and 6, $\hat{\Sigma} \xrightarrow{p} \Sigma$, $\hat{\Delta} \xrightarrow{p} \Delta$, and $\hat{\Lambda} \xrightarrow{p} \Lambda$ under any local perturbation $F_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)$.

5.1 Implementation with Minimum Distance Estimators

We have so far imposed only high-level assumptions (specifically Assumptions 3 and 6) 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. In this section we consider the case where c_0 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. Note that this encompasses the case, discussed in examples above, where the structural parameters η are estimated via MLE and the descriptive statistics $\hat{\gamma}$ are regression coefficients or comparisons of means. To facilitate application of our results, we provide sufficient conditions for Assumptions 3 and 6 in this setting.

Formally, suppose that we can decompose $\eta = (\theta, \omega)$ where θ is finite-dimensional and $c(\eta)$ depends on η only through θ . We can then, in a slight abuse of notation, write $c_0 = c(\theta_0)$, where θ_0 is defined by $\eta_0 = (\theta_0, \omega_0)$.

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

$$(8) \quad \min_{\theta \in \Theta} \hat{g}(\theta)' \hat{W} \hat{g}(\theta)$$

for Θ a compact set, $\hat{g}(\theta)$ a k_g -dimensional continuously differentiable vector of moments with Jacobian $\hat{G}(\theta)$, and \hat{W} a $k_g \times k_g$ -dimensional weighting matrix.

The researcher computes $\hat{\gamma}$ by solving

$$(9) \quad \min_{\gamma \in \Gamma} \hat{m}(\gamma)' \hat{U} \hat{m}(\gamma),$$

for Γ a compact set, $\hat{m}(\gamma)$ a k_m -dimensional continuously differentiable vector of moments with Jacobian $\hat{M}(\gamma)$, and \hat{U} a $k_m \times k_m$ -dimensional weighting matrix.

We impose several regularity conditions:

Assumption 7. Under F_0^n :

- (A) (i) both $\hat{g}(\theta)$ and $\hat{G}(\theta)$ converge uniformly in probability to continuous functions $g(\theta)$ and $G(\theta)$; (ii) $\hat{W} \rightarrow_p W$ for a positive semi-definite matrix W ; (iii) $g(\theta)' W g(\theta)$ has a unique minimum at θ_0 ; (iv) θ_0 lies in the interior of Θ ; (v) $\sqrt{n} \hat{g}(\theta_0) \rightarrow_d N(0, \Sigma_{gg})$ for full-rank Σ_{gg} ; and (vi) $G' W G = G(\theta_0)' W G(\theta_0)$ is nonsingular.
- (B) (i) both $\hat{m}(\gamma)$ and $\hat{M}(\gamma)$ converge uniformly in probability to continuous functions $m(\gamma)$ and $M(\gamma)$; (ii) $\hat{U} \rightarrow_p U$ for a positive semi-definite matrix U ; (iii) $m(\gamma)' U m(\gamma)$ has a unique minimizer γ_0 ; (iv) γ_0 lies in the interior of Γ ; (v) $\sqrt{n} \hat{m}(\gamma_0) \rightarrow_d N(0, \Sigma_{mm})$ for full-rank Σ_{mm} ; and (vi) $M' U M = M(\gamma_0)' U M(\gamma_0)$ is nonsingular.
- (C) The function $c(\cdot)$ is continuously differentiable with gradient $C(\theta)$ such that $C = C(\theta_0)$ is nonzero.
- (D) $\hat{g}(\theta_0)$ and $\hat{m}(\gamma_0)$ are asymptotically linear in the sense that

$$\sqrt{n}(\hat{g}(\theta_0), \hat{m}(\gamma_0)) = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n \phi_g(D_i), \sum_{i=1}^n \phi_m(D_i) \right) + o_p(1),$$

where $\phi_g(D_i)$ and $\phi_m(D_i)$ have mean zero and variance Σ_{gg} and Σ_{mm} .

- (E) For known estimators $\hat{\phi}_g(\cdot)$ and $\hat{\phi}_m(\cdot)$,

$$(10) \quad \frac{1}{n} \begin{pmatrix} \sum_i [\hat{\phi}_g(D_i) \hat{\phi}_g(D_i)'] & \sum_i [\hat{\phi}_g(D_i) \hat{\phi}_m(D_i)'] \\ \sum_i [\hat{\phi}_m(D_i) \hat{\phi}_g(D_i)'] & \sum_i [\hat{\phi}_m(D_i) \hat{\phi}_m(D_i)'] \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \Sigma_{gg} & \Sigma_{gm} \\ \Sigma_{mg} & \Sigma_{mm} \end{pmatrix},$$

where

$$\begin{pmatrix} \Sigma_{gg} & \Sigma_{gm} \\ \Sigma_{mg} & \Sigma_{mm} \end{pmatrix} = \begin{pmatrix} E_{F_0} [\phi_g(D_i) \phi_g(D_i)'] & E_{F_0} [\phi_g(D_i) \phi_m(D_i)'] \\ E_{F_0} [\phi_m(D_i) \phi_g(D_i)'] & E_{F_0} [\phi_m(D_i) \phi_m(D_i)'] \end{pmatrix}.$$

Parts (A) and (B) of Assumption 7 are standard regularity conditions. Part (C) allows us to apply the delta method to the estimator \hat{c} , and will hold in the wide range of situations in which the parameter of interest is a smooth, nontrivial function of model primitives. Part (D) is straightforward to verify in many situations, and holds with no remainder when $\hat{\theta}$ and $\hat{\gamma}$ are GMM or ML estimators. Part (E) requires that we be able to consistently estimate the asymptotic variance of $\sqrt{n}\hat{g}(\theta_0)$ and $\sqrt{n}\hat{m}(\gamma_0)$, and again holds under mild conditions. In particular, sufficient conditions for Assumption 7 parts (A), (B), and (E) for the case where $\hat{\theta}$ and $\hat{\gamma}$ are GMM estimators are imposed in Theorems 3.4 and 4.5 of Newey and McFadden (1994).

Assumption 7 implies Assumption 3:

Lemma 3. *Under Assumption 7 parts (A)-(D), if $p_\gamma + 1 < |\text{supp}_{F_0}(D_i)|$ then Assumption 3 holds for $\phi_c(D_i) = \Lambda_{cg}\phi_g(D_i)$ and $\phi_\gamma(D_i) = \Lambda_{\gamma m}\phi_m(D_i)$, where $\Lambda_{cg} = -C(G'WG)^{-1}G'W$ and $\Lambda_{\gamma m} = -(M'UM)^{-1}M'U$ are the sensitivities of \hat{c} with respect to $\hat{g}(\theta_0)$ and of $\hat{\gamma}$ with respect to $\hat{m}(\gamma_0)$ as defined in Andrews et al. (2017). Thus, $\sigma_c^2 = \Lambda_{cg}\Sigma_{gg}\Lambda'_{cg}$ and $\Sigma_{\gamma\gamma} = \Lambda_{\gamma m}\Sigma_{mm}\Lambda'_{\gamma m}$, $\sigma_c^2 > 0$, and $\Sigma_{\gamma\gamma}$ has full rank.*

Assumption 7 likewise implies Assumption 6:

Lemma 4. *Under Assumption 7 parts (A)-(E), Assumption 6 holds for*

$$(11) \quad \hat{\Sigma} = \frac{1}{n} \begin{pmatrix} \sum_i [\hat{\phi}_c(D_i) \hat{\phi}_c(D_i)'] & \sum_i [\hat{\phi}_c(D_i) \hat{\phi}_\gamma(D_i)'] \\ \sum_i [\hat{\phi}_\gamma(D_i) \hat{\phi}_c(D_i)'] & \sum_i [\hat{\phi}_\gamma(D_i) \hat{\phi}_\gamma(D_i)'] \end{pmatrix},$$

where for $\hat{C} = C(\hat{\theta})$, $\hat{G} = \hat{G}(\hat{\theta})$, and $\hat{M} = \hat{M}(\hat{\gamma})$,

$$\hat{\phi}_c(\cdot) = \hat{\Lambda}_{cg}\hat{\phi}_g(\cdot) = -\hat{C}(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{\phi}_g(\cdot)$$

and

$$\hat{\phi}_\gamma(\cdot) = \hat{\Lambda}_{\gamma m}\hat{\phi}_m(\cdot) = -(\hat{M}'\hat{U}\hat{M})^{-1}\hat{M}'\hat{U}\hat{\phi}_m(\cdot).$$

Corollary 1. *Under Assumptions 1 and 7, for $\hat{\Sigma}$ defined as in (11) and $\hat{\Delta}$ and $\hat{\Lambda}$ defined as in (7), $\hat{\Sigma} \xrightarrow{P} \Sigma$, $\hat{\Delta} \xrightarrow{P} \Delta$, and $\hat{\Lambda} \xrightarrow{P} \Lambda$ under any local perturbation $F_{\varphi}^n \left(\frac{\mu}{\sqrt{n}} \right)$.*

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

Lemma 4 provides a convenient recipe for estimation of Σ and hence of Δ and Λ . In the case where $\hat{\theta}$ and $\hat{\gamma}$ are GMM or ML estimators, we can write

$$\begin{aligned}\hat{g}(\theta) &= \frac{1}{n} \sum_{i=1}^n \phi_g(D_i; \theta) \\ \hat{m}(\gamma) &= \frac{1}{n} \sum_{i=1}^n \phi_m(D_i; \gamma)\end{aligned}$$

so Assumption 7 holds with $\phi_g(D_i) = \phi_g(D_i; \theta_0)$ and $\phi_m(D_i) = \phi_m(D_i; \gamma_0)$. If $\hat{\theta}$ is an MLE, then the function $\phi_g(D_i; \theta)$ is the score.¹⁰ Therefore, in the case of GMM, estimates $\hat{\phi}_g(\cdot) = \phi_g(\cdot; \hat{\theta})$ and $\hat{\phi}_m(\cdot) = \phi_m(\cdot; \hat{\gamma})$ are available immediately from the computation of the final objective of the solver for (8) and (9), respectively. In the case of MLE, the score is likewise often computed as part of the numerical gradient for the likelihood. In both cases Assumption 7 is implied by standard regularity conditions (as in, e.g., Newey and McFadden 1994).

The other elements of the calculation of $\hat{\phi}_c(\cdot)$ and $\hat{\phi}_m(\cdot)$ are also commonly precomputed. The weights \hat{W} and \hat{U} are directly involved in the calculation of the objectives in (8) and (9), respectively. The gradients \hat{G} and \hat{M} are used in standard formulae for asymptotic inference on θ_0 and γ_0 , and the gradient \hat{C} is used in delta-method calculations for asymptotic inference on c_0 .¹¹

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_0 , γ_0 , and c_0 .

¹⁰Note that we can characterize the first-order asymptotic behavior of $\hat{\theta}$ even when the first-order conditions do not uniquely determine $\hat{\theta}$ (e.g., when the likelihood has multiple local optima). See Section 1 of Newey and McFadden (1994) for discussion.

¹¹Note that in cases where the function $c(\theta)$ depends on the distribution of exogenous covariates, our formulation implicitly treats those covariates as fixed at the sample distribution for the purposes of estimating Δ and Λ . Appendix B.3 discusses how to allow for uncertainty in the distribution of covariates in a special case, and presents corresponding calculations for our applications.

Recipe. (GMM/MLE)

- 1) Estimate $\hat{\theta}$ and $\hat{\gamma}$ following (8) and (9), respectively and compute $\hat{c} = c(\hat{\theta})$.
- 2) Collect $\{\phi_g(D_i; \hat{\theta})\}_{i=1}^n$ and $\{\phi_m(D_i; \hat{\gamma})\}_{i=1}^n$ from the calculation of the objective functions in (8) and (9), respectively.
- 3) Collect the numerical gradients $\hat{G} = \hat{G}(\hat{\theta})$, $\hat{M} = \hat{M}(\hat{\gamma})$, and $\hat{C} = C(\hat{\theta})$ from the calculation of asymptotic standard errors for $\hat{\theta}$, $\hat{\gamma}$, and \hat{c} .
- 4) Compute $\hat{\Lambda}_{cg} = -\hat{C}(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}$ and $\hat{\Lambda}_{\gamma m} = -(\hat{M}'\hat{U}\hat{M})^{-1}\hat{M}'\hat{U}$ using the weights \hat{W} and \hat{U} from the objective functions in (8) and (9), respectively.
- 5) Compute $\hat{\phi}_c(D_i) = \hat{\Lambda}_{cg}\phi_g(D_i; \hat{\theta})$ 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 (11).
- 7) Compute $\hat{\Delta}$ and $\hat{\Lambda}$ as in (7).

6 Applications

6.1 The Effects of PROGRESA

Attanasio et al. (2012a) use survey data from Mexico to study the effect of PROGRESA, a randomized social experiment involving a conditional cash transfer aimed in part at increasing persistence in school. The paper estimates a parametric model via maximum likelihood. The paper uses the estimated model to conduct a counterfactual experiment 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 “exogeneous variability in [their] data that drives [their] results” as follows (p. 53):

The comparison between treatment and control villages and between eligible and ineligible households within these villages can only identify the effect of the existence

of the grant. However, the amount of the grant varies by the grade of the child. The fact that children of different ages attend the same grade offers a source of variation of the amount that can be used to identify the effect of the size of the grant. Given the demographic variables included in our model and given our treatment for initial conditions, this variation can be taken as exogenous. Moreover, the way that the grant amount changes with grade varies in a non-linear way, which also helps identify the effect.

Thus, the effect of the grant is identified by comparing across treatment and control villages, by comparing across eligible and ineligible households (having controlled for being “non-poor”), and by comparing across different ages within and between grades. (p. 53)

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

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

Table 1 reports the estimated informativeness of each vector of descriptive statistics. The estimated informativeness for the combined vector is 0.28. This is largely accounted for by the age-grade-specific treatment-control differences for eligible children.¹²

This result shows that the authors’ estimator does indeed depend on the treatment-control differences $\hat{\gamma}$, but that it also depends to a substantial degree on other features of the data orthogonal to $\hat{\gamma}$. A reader wishing to evaluate the credibility of the conclusions will gain only limited confidence

¹²The estimated informativeness is slightly higher, at 0.31, if we define \hat{c} to be the partial-equilibrium effect of the program 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 the left-hand panel of Figure 1). The corresponding parameter c_0 can be estimated without most of the structure of the model, 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”).

from believing that the elements of $\hat{\gamma}$ are valid estimates of treatment effects, and that the model links these treatment effects correctly to the counterfactual c_0 of interest. In particular, Proposition 1 tells us that knowing that $\hat{\gamma}$ is correctly specified reduces the width of the bounds on bias in \hat{c} by a factor of only $1 - \sqrt{1 - 0.28} \approx 0.15$. To gain more confidence in the estimates, a reader would need to evaluate the other assumptions in the model, specifically those that relate c_0 to features of the data orthogonal to $\hat{\gamma}$.

6.2 Newspaper Demand

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

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

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

Motivated by Gentzkow’s (2007a) discussion, we define three vectors $\hat{\gamma}$ of descriptive statistics. The first vector (“IV coefficient”) is the coefficient from a 2SLS regression of last-five-weekday print readership on last-five-weekday online readership, instrumenting for the latter with the set of instruments (Gentzkow 2007a, Table 4, Column 2, first row). The second vector (“panel coefficient”) is the coefficient from an OLS regression of last-one-day print readership on last-one-day online readership controlling for the full set of interactions between indicators for print readership and indicators for online readership in the last five weekdays. Each of these regressions includes the standard set of demographic controls from Gentzkow (2007a, Table 5). The third vector $\hat{\gamma}$ consists of both the IV coefficient and the panel coefficient. Thus, the first two vectors have dimension

1, and the third has dimension 2.

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

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

These results clarify the relative importance of the model’s assumptions. Gentzkow’s (2007a) discussion of identification highlights both the exclusion restrictions underlying the IV coefficient and the panel variation underlying the panel coefficient as sources of identification, and if anything places more emphasis on the former. Given this, a reader interested in assessing the credibility of the estimates might naturally focus most of her attention on the exclusion restrictions and other assumptions relevant to the IV coefficient.

Our results here suggest that this could be a mistake. While exclusion restrictions alone may be sufficient for identification of the substitution patterns in the model, the estimator \hat{c} is in fact much more related to the panel coefficient than to the IV coefficient, and knowing that the IV coefficient alone is correctly specified should have little effect on a reader’s confidence in the estimator \hat{c} . By contrast, confidence in the model’s interpretation of the panel coefficient—which we can loosely interpret as confidence in the model’s assumptions about the time structure of preference shocks—delivers significant additional confidence in the estimator. A reader may also wish to consider other assumptions beyond those that relate to the panel coefficient, given that the value of informativeness is well below one.

6.3 Long-term Care Insurance

Hendren (2013a) uses data on insurance eligibility and self-reported beliefs about the likelihood of different types of “loss” events (e.g., becoming disabled) to recover the distribution of underlying beliefs and rationalize why some groups are routinely denied insurance coverage. The paper estimates a parametric model via maximum likelihood. We focus here on Hendren’s (2013a) model of the market for long-term care (LTC) insurance, and define two estimates \hat{c} of interest.

The first \hat{c} , the *fraction focal point respondents*, is an estimated parameter of the model (Hen-

dren 2013a, Table A-V, row labeled “Fraction focal respondents,” Column labeled “LTC-Reject”). In Hendren’s (2013a) data, many respondents give “focal” responses of 0, 0.5, or 1 to survey elicitation of probabilistic beliefs. To allow for the possibility that these focal responses are not the respondents’ actual beliefs, Hendren’s (2013a) model assumes that with some probability each respondent is a “focal point respondent” whose response is 0, 0.5, or 1, depending on which of three intervals her true beliefs falls into. This probability is the fraction focal point respondents.

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

Hendren (2013a) writes that “the fraction of focal point respondents...and the focal point window...are identified from the distribution of focal points and the loss probability at each focal point” (p. 1752). Hendren (2013a) also explains that the parameters that determine the minimum pooled price ratio are identified from the relationship of elicited beliefs to 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 first two vectors $\hat{\gamma}$ will be especially informative for the fraction focal point respondents, and that the third vector will be especially informative for the minimum pooled price ratio.

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

Table 3 reports the estimated informativeness of each vector of descriptive statistics for each estimate of interest.

For the fraction focal point respondents, the estimated informativeness is 0.63. The estimated informativeness is 0.35 with respect to the fractions in focal point groups, 0.30 with respect to the fractions in non-focal-point groups, and 0.08 with respect to the fraction in each group needing LTC.

These results show that the focal point parameter is strongly related to the fraction of focal point responses, as expected from the structure of the model and from the author’s discussion.

For the minimum pooled price ratio, the estimated informativeness 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.

This suggests that, consistent with what one might expect given the author’s discussion, the estimated minimum pooled price ratio is strongly connected to the relationship between elicited beliefs and eventual use of LTC. A reader willing to entertain the possibility that the model is locally misspecified will gain meaningful confidence in the estimated minimum pooled price ratio if she is willing to entertain that the full vector $\hat{\gamma}$ of descriptive statistics is correctly specified under the model.

7 Conclusions

We propose a measure Δ of the informativeness of a vector $\hat{\gamma}$ of descriptive statistics for a parameter estimate \hat{c} of interest. Informativeness is the R^2 from a regression of \hat{c} on $\hat{\gamma}$ under their joint asymptotic distribution. For any given degree of local misspecification, we show that Δ , and only Δ , governs the ratio of the maximal asymptotic bias in \hat{c} under all possible local perturbations to the maximal asymptotic bias in \hat{c} under those perturbations which imply that $\hat{\gamma}$ is correctly specified. We provide a convenient recipe for computing a consistent estimator $\hat{\Delta}$ of Δ . We apply the recipe to three recent papers and interpret the results. We propose that applied researchers wishing to relate some descriptive statistics to some estimate of interest should report the corresponding $\hat{\Delta}$.

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.
- Armstrong, Timothy B. and Michal Kolesar. 2018. Sensitivity Analysis using Approximate Moment Condition Models. arXiv:1808.07387v1 [econ.EM].
- 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].
- 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.
- Gentzkow, Matthew. 2007a. Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review* 97(3): 713–744.
- Gentzkow, Matthew. 2007b. Supplementary data for Valuing new goods in a model with complementarity: Online newspapers. Accessed at <https://www.aeaweb.org/aer/data/june07/20050374_data.zip> in February 2016.
- Gentzkow, Matthew and Jesse M. Shapiro. 2015. Measuring the sensitivity of parameter estimates to sample statistics. NBER Working Paper No. 20673.
- Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson. 2014. Competition and ideological

- diversity: Historical evidence from US newspapers. *American Economic Review* 104(10): 3073–3114.
- Guggenberger, Patrik. 2012. On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption. *Econometric Theory* 28(2): 387–421.
- Hansen, Lars P., and Thomas J. Sargent. 2016. Sets of models and prices of uncertainty. NBER Working Paper No. 22000.
- 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.
- 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. Forthcoming. Temporary migration and endogenous risk-sharing in village India. *Journal of Political Economy*. doi: 10.1086/700763.
- 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. *Handbook of Econometrics 4*: 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.
- van der Vaart, Aad W. 1998. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

A Proofs

A.1 Proof of Proposition 1

We first consider the case with $\bar{\gamma}$ unrestricted and $\Delta < 1$, noting that $\Delta < 1$ implies that Σ has full rank.¹³ Define $\Lambda^* = \Sigma_{\gamma c} \sigma_c^{-2}$ and $B = \begin{bmatrix} 1 & 0 \\ -\Lambda^* & I \end{bmatrix}$. The triangular matrix B has full rank since its diagonal entries are strictly positive, so

$$\begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} = \left(B \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} \right)' (B \Sigma B')^{-1} B \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix}.$$

Note, however, that

$$B \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} = \begin{pmatrix} \bar{c} \\ \bar{\gamma} - \Lambda^* \bar{c} \end{pmatrix}$$

while

$$B \Sigma B' = \begin{bmatrix} \sigma_c^2 & 0 \\ 0 & \Sigma_{\gamma\gamma} - \Lambda^* \sigma_c^2 \Lambda^{*'} \end{bmatrix}.$$

Hence,

$$\begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} = \frac{\bar{c}^2}{\sigma_c^2} + (\bar{\gamma} - \Lambda^* \bar{c})' \left(\Sigma_{\gamma\gamma} - \Lambda^* \sigma_c^2 \Lambda^{*'} \right)^{-1} (\bar{\gamma} - \Lambda^* \bar{c}).$$

The matrix $\Sigma_{\gamma\gamma} - \Lambda^* \sigma_c^2 \Lambda^{*'}$ is the asymptotic variance of $\sqrt{n}(\hat{\gamma} - \Lambda^* \hat{c})$ and so is positive semi-definite. Since one can choose $\bar{\gamma}$ to set the second term to zero for all values \bar{c} , the range of values of \bar{c}_φ consistent with the restriction in Condition 2 is $[-\sigma_c, \sigma_c]$. The form of \mathcal{B}^μ then follows immediately from Condition 1.

We next consider the case with $\bar{\gamma}$ unrestricted and $\Delta = 1$. Let O be a full-rank matrix with first row proportional to Λ and remaining rows Λ^\perp such that $\Lambda^\perp \Sigma_{\gamma\gamma} \Lambda' = 0$.¹⁴ We therefore have

$$O \Sigma_{\gamma\gamma} O' = \begin{bmatrix} \sigma_c^2 & 0 \\ 0 & \Lambda^\perp \Sigma_{\gamma\gamma} \Lambda^{\perp'} \end{bmatrix}. \text{ Consequently, we can write}$$

$$\begin{aligned} \bar{\gamma}' \Sigma_{\gamma\gamma}^{-1} \bar{\gamma} &= (O \bar{\gamma})' (O \Sigma_{\gamma\gamma} O')^{-1} O \bar{\gamma} \\ &= \frac{(\Lambda \bar{c})^2}{\sigma_c^2} + \left(\Lambda^\perp \bar{\gamma} \right)' \left(\Lambda^\perp \Sigma_{\gamma\gamma} \Lambda^{\perp'} \right)^{-1} \Lambda^\perp \bar{\gamma}. \end{aligned}$$

¹³To see that this is case, note that we have already assumed Σ_{gg} has full rank, and $\Delta < 1$ implies that $\sqrt{n}(\hat{c} - c_0)$ cannot be expressed (asymptotically under F_0^n) as a linear combination of $\sqrt{n}(\hat{\gamma} - \gamma_0)$. Hence, when $\Delta < 1$ no element of the vector $(\sqrt{n}(\hat{c} - c_0), \sqrt{n}(\hat{\gamma} - \gamma_0))$ can be expressed (asymptotically under F_0^n) as a linear combination of the other elements, which is equivalent to Σ having full rank.

¹⁴For example, we can take any basis for the null space of $\Lambda \Sigma_{\gamma\gamma}$.

As above, $\Lambda^\perp \Sigma_{\gamma\gamma} \Lambda^{\perp'}$ is positive semi-definite, and we can always set $\Lambda^\perp \bar{\gamma} = 0$ by picking $\bar{\gamma}$ proportional to $\Sigma_{\gamma\gamma} \Lambda'$, so this decomposition makes clear that the range of values consistent with the restriction in Condition 2 is $[-\sigma_c, \sigma_c]$.

To consider the case with $\bar{\gamma} = 0$ and $\Delta < 1$, define $\Lambda = \Sigma_{c\gamma} \Sigma_{\gamma\gamma}^{-1}$ and $A = \begin{bmatrix} 1 & -\Lambda \\ 0 & I \end{bmatrix}$. A has full rank, so

$$\begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} = \left(A \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} \right)' (A \Sigma A')^{-1} A \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix}.$$

Note, however, that

$$A \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} = \begin{pmatrix} \bar{c} - \Lambda \bar{\gamma} \\ \bar{\gamma} \end{pmatrix}$$

while

$$A \Sigma A' = \begin{bmatrix} \sigma_c^2 - \Lambda \Sigma_{\gamma\gamma} \Lambda' & 0 \\ 0 & \Sigma_{\gamma\gamma} \end{bmatrix} = \begin{bmatrix} \sigma_c^2 (1 - \Delta) & 0 \\ 0 & \Sigma_{\gamma\gamma} \end{bmatrix}.$$

Hence,

$$\begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} = \frac{(\bar{c} - \Lambda \bar{\gamma})^2}{\sigma_c^2 (1 - \Delta)} + \bar{\gamma}' \Sigma_{\gamma\gamma}^{-1} \bar{\gamma}.$$

Thus, if we impose $\bar{\gamma} = 0$, then under Condition 2 the range of values of \bar{c} is $[-\sigma_c \sqrt{1 - \Delta}, \sigma_c \sqrt{1 - \Delta}]$, and the form of \mathcal{B}_0^μ follows immediately from Condition 1.

Finally, in the case where $\bar{\gamma} = 0$ and $\Delta = 1$, under Condition 2 we have $\bar{c} = 0$, and the result is immediate.

A.2 Proof of Lemma 1

By Lemma 7.6 of van der Vaart (1998), Assumption 1 implies that $\sqrt{f_\varphi(D_i; \vartheta)}$ is differentiable in quadratic mean in the sense that for all φ ,

$$\int \left(\sqrt{f_\varphi(d; \vartheta)} - \sqrt{f_\varphi(d; 0)} - \frac{1}{2} \vartheta s_\varphi(d) \sqrt{f_\varphi(d; 0)} \right)^2 d\nu(d) = o(\vartheta^2)$$

as $\vartheta \rightarrow 0$. Theorem 7.2 of van der Vaart (1998) then implies that under F_0^n we have a quadratic approximation to the log likelihood ratios

$$(12) \quad \log \left(\frac{dF_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)}{dF_0^n} \right) = \sum_{i=1}^n \log \left(\frac{f_\varphi \left(D_i; \frac{\mu}{\sqrt{n}} \right)}{f_\varphi(D_i; 0)} \right) = \frac{\mu}{\sqrt{n}} \sum_{i=1}^n s_\varphi(D_i) - \frac{\mu^2}{2} E_{F_0} [s_\varphi(D_i)^2] + o_p(1)$$

and that $E_{F_0} [s_\varphi(D_i)] = 0$. Since $E_{F_0} [s_\varphi(D_i)^2]$ is finite, Assumption 3, the Central Limit Theorem and Slutsky's Lemma imply that under F_0^n

$$\left(\log \left(\frac{dF_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)}{dF_0^n} \right) \quad \frac{1}{\sqrt{n}} \sum \phi_c(D_i) \quad \frac{1}{\sqrt{n}} \sum \phi_\gamma(D_i) \right) \rightarrow_d \left(\zeta \quad \xi_c \quad \xi_\gamma \right) \sim$$

$$N \left(\begin{pmatrix} -\frac{\mu^2}{2} E_{F_0} [s_\varphi(D_i)^2] \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mu^2 E_{F_0} [s_\varphi(D_i)^2] & \mu E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] & \mu E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)'] \\ \mu E_{F_0} [\phi_c(D_i) s_\varphi(D_i)] & E_{F_0} [\phi_c(D_i)^2] & E_{F_0} [\phi_c(D_i) \phi_\gamma(D_i)'] \\ \mu E_{F_0} [\phi_\gamma(D_i) s_\varphi(D_i)] & E_{F_0} [\phi_\gamma(D_i) \phi_c(D_i)] & E_{F_0} [\phi_\gamma(D_i) \phi_\gamma(D_i)'] \end{pmatrix} \right).$$

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

$$\left(\log \left(\frac{dF_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)}{dF_0^n} \right) \quad \frac{1}{\sqrt{n}} \sum \phi_c(D_i) \quad \frac{1}{\sqrt{n}} \sum \phi_\gamma(D_i) \right) \rightarrow_d \left(\zeta^* \quad \xi_c^* \quad \xi_\gamma^* \right) \sim$$

$$N \left(\begin{pmatrix} \frac{\mu^2}{2} E_{F_0} [s_\varphi(D_i)^2] \\ \mu E_{F_0} [\phi_c(D_i) s_\varphi(D_i)] \\ \mu E_{F_0} [\phi_\gamma(D_i) s_\varphi(D_i)] \end{pmatrix}, \begin{pmatrix} \mu^2 E_{F_0} [s_\varphi(D_i)^2] & \mu E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] & \mu E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)'] \\ \mu E_{F_0} [\phi_c(D_i) s_\varphi(D_i)] & E_{F_0} [\phi_c(D_i)^2] & E_{F_0} [\phi_c(D_i) \phi_\gamma(D_i)'] \\ \mu E_{F_0} [\phi_\gamma(D_i) s_\varphi(D_i)] & E_{F_0} [\phi_\gamma(D_i) \phi_c(D_i)] & E_{F_0} [\phi_\gamma(D_i) \phi_\gamma(D_i)'] \end{pmatrix} \right).$$

Contiguity of F_0^n and $F_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)$ implies that any quantity that converges in probability under F_0^n converges in probability to the same limit under $F_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)$. Combined with Assumption 3, however, this means that under $F_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)$

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

and thus that

$$\sqrt{n} \begin{pmatrix} \hat{c} - c_0 \\ \hat{\gamma} - \gamma_0 \end{pmatrix} \rightarrow_d N \left(\begin{pmatrix} \mu \bar{c}_\varphi \\ \mu \bar{\gamma}_\varphi \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \Sigma_{c\gamma} \\ \Sigma_{\gamma c} & \Sigma_{\gamma\gamma} \end{pmatrix} \right),$$

for

$$\begin{pmatrix} \bar{c}_\varphi \\ \bar{\gamma}_\varphi \end{pmatrix} = \begin{pmatrix} E_{F_0} [\phi_c(D_i) s_\varphi(D_i)] \\ E_{F_0} [\phi_\gamma(D_i) s_\varphi(D_i)] \end{pmatrix},$$

and

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

This completes the proof.

A.3 Proof of Lemma 2

We first consider the case with $\Delta < 1$, and show that for any $(\bar{c}, \bar{\gamma})$ consistent with (3) we can find $\varphi \in \Phi$ such that $(\bar{c}_\varphi, \bar{\gamma}_\varphi) = (\bar{c}, \bar{\gamma})$. First consider any

$$(13) \quad (\bar{c}, \bar{\gamma}) \in \left\{ (c, \gamma) \in \mathbb{R} \times \mathbb{R}^{p_\gamma} : \sqrt{\begin{pmatrix} c \\ \gamma \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} c \\ \gamma \end{pmatrix}} = 1 \right\}.$$

Define

$$s(D_i; \bar{c}, \bar{\gamma}) = \begin{pmatrix} \phi_c(D_i) & \phi_\gamma(D_i)' \end{pmatrix} \Sigma^{-1} \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix}$$

and note that $E_{F_0} [s(D_i; \bar{c}, \bar{\gamma})] = 0$ and

$$E_{F_0} [s(D_i; \bar{c}, \bar{\gamma})^2] = \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} = 1.$$

Hence, by Assumption 2 there exists $\varphi \in \Phi$ with $s_\varphi(D_i) = s(D_i; \bar{c}, \bar{\gamma})$ F_0 -almost everywhere. By construction

$$\begin{pmatrix} \bar{c}_\varphi \\ \bar{\gamma}_\varphi \end{pmatrix} = \begin{pmatrix} E_{F_0} [\phi_c(D_i) s_\varphi(D_i)] \\ E_{F_0} [\phi_\gamma(D_i) s_\varphi(D_i)] \end{pmatrix} = \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix},$$

so we can attain any $(\bar{c}, \bar{\gamma})$ satisfying (13).

Next, consider any

$$(\bar{c}, \bar{\gamma}) \in \left\{ (c, \gamma) \in \mathbb{R} \times \mathbb{R}^{p_\gamma} : \sqrt{\begin{pmatrix} c \\ \gamma \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} c \\ \gamma \end{pmatrix}} < 1 \right\},$$

and note that by Assumption 3 (and in particular the assumption that $\dim(\gamma) + 1 < |\text{supp}_{F_0}(D_i)|$), there exists a function $\varepsilon^*(\cdot)$ with $E_{F_0} [\varepsilon^*(D_i)] = 0$, $E_{F_0} [\varepsilon^*(D_i)^2] = 1$, $E_{F_0} [\varepsilon^*(D_i) \phi_c(D_i)] = 0$, $E_{F_0} [\varepsilon^*(D_i) \phi_\gamma(D_i)] = 0$, and therefore $E_{F_0} [\varepsilon^*(D_i) s(D_i; \bar{c}, \bar{\gamma})] = 0$. Let us consider $\varphi \in \Phi$ corre-

spending to

$$s_\varphi(D_i) = s(D_i; \bar{c}, \bar{\gamma}) + \sqrt{1 - \left(\frac{\bar{c}}{\bar{\gamma}} \right)' \Sigma^{-1} \left(\frac{\bar{c}}{\bar{\gamma}} \right)} \varepsilon^*(D_i),$$

noting that since the right hand side has mean zero and variance one, such a φ again exists by Assumption 2. The corresponding score $s_\varphi(D_i)$ again delivers the desired $(\bar{c}_\varphi, \bar{\gamma}_\varphi)$.

We next establish the result in the other direction (again for the case of $\Delta < 1$), and show that for any $\varphi \in \Phi$, $(\bar{c}_\varphi, \bar{\gamma}_\varphi)$ satisfy (3). Consider any φ , and note that we can write

$$s_\varphi(D_i) = s(D_i; \bar{c}_\varphi, \bar{\gamma}_\varphi) + \varepsilon_\varphi(D_i)$$

for $\varepsilon_\varphi(D_i) = s_\varphi(D_i) - s(D_i; \bar{c}_\varphi, \bar{\gamma}_\varphi)$. Note that by the definition of $s(D_i; \bar{c}_\varphi, \bar{\gamma}_\varphi)$,

$$\begin{pmatrix} E_{F_0} [\phi_c(D_i) s_\varphi(D_i)] \\ E_{F_0} [\phi_\gamma(D_i) s_\varphi(D_i)] \end{pmatrix} = \begin{pmatrix} E_{F_0} \left[\phi_c(D_i) s(D_i; \bar{c}_\varphi, \bar{\gamma}_\varphi) \right] \\ E_{F_0} \left[\phi_\gamma(D_i) s(D_i; \bar{c}_\varphi, \bar{\gamma}_\varphi) \right] \end{pmatrix},$$

and thus

$$\begin{pmatrix} E_{F_0} [\phi_c(D_i) \varepsilon_\varphi(D_i)] \\ E_{F_0} [\phi_\gamma(D_i) \varepsilon_\varphi(D_i)] \end{pmatrix} = 0.$$

Hence, $\varepsilon_\varphi(D_i)$ is orthogonal to $\phi_c(D_i)$ and $\phi_\gamma(D_i)$, and thus to $s(D_i; \bar{c}_\varphi, \bar{\gamma}_\varphi)$. In particular, we can interpret $s(D_i; \bar{c}_\varphi, \bar{\gamma}_\varphi)$ as the projection of $s_\varphi(D_i)$ onto $(\phi_c(D_i), \phi_\gamma(D_i))$, and $\varepsilon_\varphi(D_i)$ as the residual from this projection. Using this orthogonality, note that

$$\begin{aligned} \sqrt{E_{F_0} [s_\varphi(D_i)^2]} &= \sqrt{E_{F_0} [s(D_i; \bar{c}_\varphi, \bar{\gamma}_\varphi)^2] + E_{F_0} [\varepsilon_\varphi(D_i)^2]} \\ &\geq \sqrt{E_{F_0} [s(D_i; \bar{c}_\varphi, \bar{\gamma}_\varphi)^2]} = \sqrt{\left(\frac{\bar{c}_\varphi}{\bar{\gamma}_\varphi} \right)' \Sigma^{-1} \left(\frac{\bar{c}_\varphi}{\bar{\gamma}_\varphi} \right)}. \end{aligned}$$

Thus, for $\Delta < 1$, $\sqrt{E_{F_0} [s_\varphi(D_i)^2]} \leq 1$ implies that $(\bar{c}_\varphi, \bar{\gamma}_\varphi)$ satisfies (3). This completes the proof for the case with $\Delta < 1$.

We finally consider the case with $\Delta = 1$. In this case, we can repeat the argument above (dropping \bar{c} and replacing Σ with $\Sigma_{\gamma\gamma}$) to show that a value $\bar{\gamma}$ is consistent with $\sqrt{E_{F_0} [s_\varphi(D_i)^2]} \leq 1$ if and only if $\sqrt{\bar{\gamma}' \Sigma_{\gamma\gamma}^{-1} \bar{\gamma}} \leq 1$. The proof of Proposition 5 below shows, however, that $\bar{c}_\varphi = \Lambda \bar{\gamma}_\varphi$ when $\Delta = 1$, which completes the proof.

A.4 Proof of Proposition 2

By the Neyman-Pearson Lemma (see Theorem 3.2.1 in Lehmann and Romano 2005), the most powerful level- α test of $H_0 : (D_1, \dots, D_n) \sim F_0^n$ against $H_1 : (D_1, \dots, D_n) \sim F_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)$ rejects when the log likelihood ratio $\log \left(\frac{dF_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)}{dF_0^n} \right)$ exceeds a critical value $c_{\alpha, n}$ chosen to ensure rejection probability α under H_0 (and may randomize when the log likelihood ratio exactly equals the critical value).

From the proof of Lemma 1, however, we see that since we limit attention to non-trivial perturbations,

$$\log \left(\frac{dF_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)}{dF_0^n} \right) \rightarrow_d \begin{cases} N \left(-\frac{\mu^2}{2}, \mu^2 \right) & \text{under } F_0^n \\ N \left(\frac{\mu^2}{2}, \mu^2 \right) & \text{under } F_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right) \end{cases}.$$

Hence, since $c_{\alpha, n}$ corresponds to the $1 - \alpha$ quantile of the log likelihood ratio under the null, we see that it converges to the $1 - \alpha$ quantile of a $N \left(-\frac{\mu^2}{2}, \mu^2 \right)$ distribution. Thus,

$$\frac{\log \left(\frac{dF_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)}{dF_0^n} \right) - c_{\alpha, n}}{\mu} \rightarrow_d \begin{cases} N(-c_\alpha, 1) & \text{under } F_0^n \\ N(\mu - c_\alpha, 1) & \text{under } F_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right) \end{cases}$$

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

A.5 Proof of Proposition 3

Assumption 5 and Leibniz's rule implies for n sufficiently large we can exchange integration and differentiation twice in the definition of $r \left(F_0, F_\varphi \left(\frac{\mu}{\sqrt{n}} \right) \right)$, so by Taylor's Theorem

$$n \cdot r \left(F_0, F_\varphi \left(\frac{\mu}{\sqrt{n}} \right) \right) = n \cdot E_{F_0} \left[\begin{aligned} & \psi \left(\frac{f_\varphi(D_i; 0)}{f_\varphi(D_i; 0)} \right) + \psi' \left(\frac{f_\varphi(D_i; 0)}{f_\varphi(D_i; 0)} \right) \frac{\frac{\partial}{\partial \vartheta} f_\varphi(D_i; 0)}{f_\varphi(D_i; 0)} \frac{\mu}{\sqrt{n}} \\ & + \frac{1}{2} \left(\psi' \left(\frac{f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right) \frac{\frac{\partial^2}{\partial \vartheta^2} f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} + \psi'' \left(\frac{f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right) \left(\frac{\frac{\partial}{\partial \vartheta} f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right)^2 \right) \frac{\mu^2}{n} \end{aligned} \right]$$

for $\tilde{\vartheta}_n$ a sequence of values with $\tilde{\vartheta}_n \in \left[0, \frac{\mu}{\sqrt{n}} \right]$. Thus, since $\psi(1) = 0$ by Assumption 4,

$$n \cdot E_{F_0} \left[\psi \left(\frac{dF_\varphi \left(\frac{\mu}{\sqrt{n}} \right)}{dF_0} \right) \right] =$$

$$E_{F_0} \left[+\frac{\mu^2}{2} \left(\psi' \left(\frac{f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right) \frac{\frac{\partial^2}{\partial \vartheta^2} f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} + \psi'' \left(\frac{f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right) \left(\frac{\frac{\partial}{\partial \vartheta} f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right)^2 \right) \right].$$

Assumption 5 and Leibniz's rule implies that

$$\frac{\partial}{\partial \vartheta} \int f_\varphi(d; \vartheta) d\nu(d) = \int \frac{\partial}{\partial \vartheta} f_\varphi(d; \vartheta) d\nu(d).$$

This can also be verified using Assumption 1 together with Lemma 7.6 and Theorem 7.2 of van der Vaart (1998). Hence, we see that

$$E_{F_0} [s_\varphi(D_i)] = \int \frac{\partial}{\partial \vartheta} f_\varphi(d; 0) d\nu(d) = \frac{\partial}{\partial \vartheta} \int f_\varphi(d; 0) d\nu(d) = 0,$$

and thus that

$$n \cdot E_{F_0} \left[\psi \left(\frac{dF_\varphi \left(\frac{\mu}{\sqrt{n}} \right)}{dF_0} \right) \right] = \frac{\mu^2}{2} E_{F_0} \left[\psi' \left(\frac{f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right) \frac{\frac{\partial^2}{\partial \vartheta^2} f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} + \psi'' \left(\frac{f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right) \left(\frac{\frac{\partial}{\partial \vartheta} f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right)^2 \right].$$

The Dominated Convergence Theorem then implies that (since $\psi''(1) = 2$ by Assumption 4, and the expression in brackets is uniformly bounded on \mathcal{N}_ϑ by a random variable with finite expectation by Assumption 5)

$$\begin{aligned} & \frac{\mu^2}{2} E_{F_0} \left[\psi' \left(\frac{f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right) \frac{\frac{\partial^2}{\partial \vartheta^2} f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} + \psi'' \left(\frac{f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right) \left(\frac{\frac{\partial}{\partial \vartheta} f_\varphi(D_i; \tilde{\vartheta}_n)}{f_\varphi(D_i; 0)} \right)^2 \right] \rightarrow \\ & \frac{\mu^2}{2} E_{F_0} \left[\psi' \left(\frac{f_\varphi(D_i; 0)}{f_\varphi(D_i; 0)} \right) \frac{\frac{\partial^2}{\partial \vartheta^2} f_\varphi(D_i; 0)}{f_\varphi(D_i; 0)} + \psi'' \left(\frac{f_\varphi(D_i; 0)}{f_\varphi(D_i; 0)} \right) s_\varphi(D_i)^2 \right] \\ & = \mu^2 E_{F_0} \left[\frac{\psi'(1)}{2} \frac{\frac{\partial^2}{\partial \vartheta^2} f_\varphi(D_i; 0)}{f_\varphi(D_i; 0)} + s_\varphi(D_i)^2 \right]. \end{aligned}$$

However,

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

so

$$\lim_{n \rightarrow \infty} n \cdot r \left(F_0, F_\varphi \left(\frac{\mu}{\sqrt{n}} \right) \right) = \mu^2 E_{F_0} \left[s_\varphi(D_i)^2 \right] = \mu^2,$$

where the last equality follows from Assumption 1 and the fact that we have limited attention to non-trivial perturbations.

A.6 Proof of Proposition 5

For trivial perturbations (i.e. perturbations with $E_{F_0} \left[s_\varphi(D_i)^2 \right] = 0$) the Cauchy-Schwarz inequality implies that $\bar{c}_\varphi = 0$ and $\bar{\gamma}_\varphi = 0$, since for example $|\bar{c}_\varphi| \leq \sigma_c \sqrt{E_{F_0} \left[s_\varphi(D_i)^2 \right]}$. Hence, for the remainder of the proof we consider non-trivial perturbations with $E_{F_0} \left[s_\varphi(D_i)^2 \right] = 1$.

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

$$\varepsilon_\varphi(D_i) = s_\varphi(D_i) - s(D_i; \bar{\gamma}_\varphi).$$

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

$$\begin{aligned} \bar{c}_\varphi &= E_{F_0} \left[\phi_c(D_i) s_\varphi(D_i) \right] = E_{F_0} \left[\phi_c(D_i) \phi_\gamma(D_i) \right] \Sigma_{\gamma\gamma}^{-1} \bar{\gamma}_\varphi + E_{F_0} \left[\phi_c(D_i) \varepsilon_\varphi(D_i) \right] \\ &= \Lambda \bar{\gamma}_\varphi + E_{F_0} \left[\phi_c(D_i) \varepsilon_\varphi(D_i) \right]. \end{aligned}$$

Next, define

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

and note that

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

The Cauchy-Schwarz inequality then implies that

$$\begin{aligned} |E_{F_0} \left[\tilde{\phi}_c(D_i) \varepsilon_\varphi(D_i) \right]| &\leq \sqrt{E_{F_0} \left[\tilde{\phi}_c(D_i)^2 \right]} \sqrt{E_{F_0} \left[\varepsilon_\varphi(D_i)^2 \right]} \\ &= \sqrt{\sigma_c^2 - \Lambda \Sigma_{\gamma\gamma} \Lambda'} \sqrt{E_{F_0} \left[\varepsilon_\varphi(D_i)^2 \right]} \\ &= \sigma_c \sqrt{1 - \Delta} \sqrt{1 - \bar{\gamma}'_\varphi \Sigma_{\gamma\gamma}^{-1} \bar{\gamma}_\varphi}. \end{aligned}$$

Combining these results we see that for (μ, φ) such that $\mu \bar{\gamma}_\varphi = \bar{\gamma}$

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

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

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

so if $\mu \bar{\gamma}_\varphi = \bar{\gamma}$, we must have $\bar{\gamma}' \Sigma_{\gamma\gamma}^{-1} \bar{\gamma} \leq \mu^2$. Hence, if $\mu^2 - \bar{\gamma}' \Sigma_{\gamma\gamma}^{-1} \bar{\gamma} < 0$ then there exists no φ with $\mu \bar{\gamma}_\varphi = \bar{\gamma}$, so the set of possible values for $\mu \bar{c}_\varphi$ is empty.

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

$$(14) \quad \bar{c} \in \left[\Lambda \bar{\gamma} - \sigma_c \sqrt{1 - \Delta} \sqrt{\mu^2 - \bar{\gamma}' \Sigma_{\gamma\gamma}^{-1} \bar{\gamma}}, \Lambda \bar{\gamma} + \sigma_c \sqrt{1 - \Delta} \sqrt{\mu^2 - \bar{\gamma}' \Sigma_{\gamma\gamma}^{-1} \bar{\gamma}} \right]$$

there exists $\varphi \in \Phi$ with $\mu \bar{c}_\varphi = \bar{c}$ and $\mu \bar{\gamma}_\varphi = \bar{\gamma}$. To prove that this is the case, let us assume without loss of generality that $\mu = 1$ (the result in other cases corresponds to a rescaling of this case). If $\Delta < 1$, define

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

Note that

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

while

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

Moreover,

$$\begin{aligned} E_{F_0} [s^*(D_i; \bar{c}, \bar{\gamma})^2] &= E_{F_0} [s(D_i; \bar{\gamma})^2] + E_{F_0} [\tilde{\phi}_c(D_i)^2] \frac{(\bar{c} - \Lambda \bar{\gamma})^2}{\sigma_c^4 (1 - \Delta)^2} \\ &= \bar{\gamma}' \Sigma_{\gamma\gamma}^{-1} \bar{\gamma} + \frac{(\bar{c} - \Lambda \bar{\gamma})^2}{\sigma_c^2 (1 - \Delta)}. \end{aligned}$$

However, by (14) we know that

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

and thus that

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

so $E_{F_0} [s^*(D_i; \bar{c}, \bar{\gamma})^2] \leq 1$. Hence, for $\varepsilon^*(D_i)$ such that $E_{F_0} [\phi_\gamma(D_i) \varepsilon^*(D_i)] = 0$, $E_{F_0} [\phi_c(D_i) \varepsilon^*(D_i)] = 0$, and $E_{F_0} [\varepsilon^*(D_i)^2] = 1$ (which exists as argued in the proof of Lemma 2), by Assumption 2 there

exists φ with score corresponding to

$$s^*(D_i; \bar{c}, \bar{\gamma}) + \sqrt{1 - E_{F_0} [s^*(D_i; \bar{c}, \bar{\gamma})^2]} \varepsilon^*(D_i)$$

and this φ yields $E_{F_0} [\phi_c(D_i) s_\varphi(D_i)] = \bar{c}$, $E_{F_0} [\phi_\gamma(D_i) s_\varphi(D_i)] = \bar{\gamma}$, as desired. In cases with $\Delta = 1$, on the other hand, we can repeat the same argument with $s^*(D_i; \bar{c}, \bar{\gamma}) = s(D_i; \bar{\gamma})$. That $\mathcal{B}_{\bar{\gamma}}^\mu$ gives the range of asymptotic biases for \hat{c} under perturbations with $\mu \bar{\gamma}_\varphi = \bar{\gamma}$ then follows from Lemma 1.

The result for $\mathcal{B}^{\mu(\bar{\gamma})}$ is immediate from Lemma 2 and Proposition 1.

A.7 Proof of Proposition 6

The proof of Lemma 1 shows that the distribution of the data under $F_\varphi^n \left(\frac{\mu}{\sqrt{n}} \right)$ is mutually contiguous with that under F_0^n . Hence, to establish convergence in probability under all local perturbations, it suffices to establish convergence in probability under F_0^n . Consistency of $\hat{\Delta}$ and $\hat{\Lambda}$ under F_0^n is implied by the Continuous Mapping Theorem (see e.g. Theorem 2.3 of van der Vaart 1998) and the maintained assumptions that $\sigma_c^2 > 0$ and $\Sigma_{\gamma\gamma}$ has full rank.

A.8 Proof of Lemma 3

The proof of Proposition 1 in Andrews et al. (2017) along with Remark 1 in that paper (or, alternatively, the proof of Theorem 3.2 of Newey and McFadden 1994), implies that

$$\sqrt{n}(\hat{c} - c_0, \hat{\gamma} - \gamma_0) = \sqrt{n}(\Lambda_{cg} \hat{g}(\theta_0), \Lambda_{\gamma m} \hat{m}(\gamma_0)) + o_p(1).$$

The result then follows immediately from Assumption 7.

A.9 Proof of Lemma 4

Note, first, that $\hat{\Lambda}_{cg} \xrightarrow{p} \Lambda_{cg}$ and $\hat{\Lambda}_{\gamma m} \xrightarrow{p} \Lambda_{\gamma m}$ by Assumption 7 parts (A)-(C) along with the Continuous Mapping Theorem (see e.g. Theorem 2.3 of van der Vaart 1998). Next, for

$$\Sigma^* = \begin{pmatrix} \Sigma_{gg} & \Sigma_{gm} \\ \Sigma_{mg} & \Sigma_{mm} \end{pmatrix}$$

and $\hat{\Sigma}^*$ the estimator on the left hand side of (10), Assumption 7 part (E) implies that $\hat{\Sigma}^* \rightarrow_p \Sigma^*$. By Lemma 3,

$$\Sigma = \begin{bmatrix} E_{F_0} [\phi_c(D_i)^2] & E_{F_0} [\phi_c(D_i) \phi_\gamma(D_i)'] \\ E_{F_0} [\phi_\gamma(D_i) \phi_c(D_i)] & E_{F_0} [\phi_\gamma(D_i) \phi_\gamma(D_i)'] \end{bmatrix} = \begin{bmatrix} \Lambda_{cg} & 0 \\ 0 & \Lambda_{\gamma m} \end{bmatrix} \Sigma^* \begin{bmatrix} \Lambda_{cg} & 0 \\ 0 & \Lambda_{\gamma m} \end{bmatrix}'.$$

Finally, note that we can write $\hat{\Sigma}$ as

$$\hat{\Sigma} = \begin{bmatrix} \hat{\Lambda}_{cg} & 0 \\ 0 & \hat{\Lambda}_{\gamma m} \end{bmatrix} \hat{\Sigma}^* \begin{bmatrix} \hat{\Lambda}_{cg} & 0 \\ 0 & \hat{\Lambda}_{\gamma m} \end{bmatrix}',$$

so $\hat{\Sigma} \rightarrow_p \Sigma$ by the Continuous Mapping Theorem.

A.10 Proof of Corollary 1

Follows immediately from Lemma 4 and Proposition 6.

B Additional Results

B.1 Non-constant η

In the main text we assume that the true data generating process lies in a neighborhood of a fixed distribution $F_0 = F(\eta_0)$, and compare \hat{c} and $\hat{\gamma}$ to fixed values c_0 and γ_0 . In this section we show that our results are robust to instead allowing the “centering” value of η (and thus the corresponding c and γ) to vary with the sample size, local to a fixed η_0 .

Assumptions

Define \mathcal{H} as a set of values such that for all $h \in \mathcal{H}$, $\eta_0 + \zeta \cdot h \in H$ for $\zeta > 0$ sufficiently small. Define $F_{h,\varphi}(\zeta, \vartheta)$ to perturb $F(\eta_0 + \zeta \cdot h)$ by ϑ in the direction φ .¹⁵ We again assume the distributions $F_{h,\varphi}(\zeta, \vartheta)$ have densities $f_{h,\varphi}(\zeta, \vartheta)$ with respect to a dominating measure ν . The information matrix for (ζ, ϑ) , treating (h, φ) as known, is

$$I_{h,\varphi}(\zeta, \vartheta) = E_{F_{h,\varphi}(\zeta, \vartheta)} \begin{bmatrix} \left(\frac{\frac{\partial}{\partial \zeta} f_{h,\varphi}(D_i; \zeta, \vartheta)}{f_{h,\varphi}(D_i; \zeta, \vartheta)} \right)^2 & \frac{\frac{\partial}{\partial \zeta} f_{h,\varphi}(D_i; \zeta, \vartheta)}{f_{h,\varphi}(D_i; \zeta, \vartheta)} \frac{\frac{\partial}{\partial \vartheta} f_{h,\varphi}(D_i; \zeta, \vartheta)}{f_{h,\varphi}(D_i; \zeta, \vartheta)} \\ \frac{\frac{\partial}{\partial \zeta} f_{h,\varphi}(D_i; \zeta, \vartheta)}{f_{h,\varphi}(D_i; \zeta, \vartheta)} \frac{\frac{\partial}{\partial \vartheta} f_{h,\varphi}(D_i; \zeta, \vartheta)}{f_{h,\varphi}(D_i; \zeta, \vartheta)} & \left(\frac{\frac{\partial}{\partial \vartheta} f_{h,\varphi}(D_i; \zeta, \vartheta)}{f_{h,\varphi}(D_i; \zeta, \vartheta)} \right)^2 \end{bmatrix}.$$

¹⁵If $\eta_0 + \zeta \cdot h \notin H$ we may define this distribution arbitrarily.

Define $s_h(D_i) = \frac{\partial}{\partial \zeta} (\log f_{h,\varphi}(D_i; 0, 0))$ and $s_\varphi(D_i) = \frac{\partial}{\partial \vartheta} (\log f_{h,\varphi}(D_i; 0, 0))$. We impose the following assumption, which extends Assumption 1 to the present setting:

Assumption 8. For (ζ, ϑ) in an open neighborhood of zero and all $h \in \mathcal{H}$, $\varphi \in \Phi$; (i) $\sqrt{f_{h,\varphi}(d; \zeta, \vartheta)}$ is continuously differentiable with respect to (ζ, ϑ) for all $d \in \mathcal{D}$; (ii) $I_{h,\varphi}(\zeta, \vartheta)$ is finite and continuous in (ζ, ϑ) ; and (iii) $E_{F_0} \left[(s_\varphi(D_i))^2 \right] \in \{0, 1\}$.

We likewise extend Assumption 2, while Assumption 3 does not require modification.

Assumption 9. The set of score functions $s_\varphi(D_i)$ includes all those consistent with Assumption 8, in the sense that for any $s(\cdot)$ with $E_{F_0} [s(D_i)] = 0$ and $E_{F_0} [s(D_i)^2] \in \{0, 1\}$ there exists $\varphi \in \Phi$ with $E_{F_0} \left[\left(s(D_i) - \frac{\partial}{\partial \vartheta} \log (s_\varphi(D_i)) \right)^2 \right] = 0$.

Finally, we need an assumption to control the behavior of c and γ as η varies. Formally, we define $\gamma(\eta)$ as the probability limit of $\hat{\gamma}$ under $F^n(\eta)$ (which we assume exists for η sufficiently close to η_0) and impose the following:

Assumption 10. For any $h \in \mathcal{H}$ and $\tau \geq 0$, if $\eta_{0,n} = \eta_0 + \frac{\tau}{\sqrt{n}}h$, $c_{0,n} = c(\eta_{0,n})$, and $\gamma_{0,n} = \gamma(\eta_{0,n})$, then under $F^n(\eta_{0,n})$,

$$\sqrt{n} \begin{pmatrix} \hat{c} - c_{0,n} \\ \hat{\gamma} - \gamma_{0,n} \end{pmatrix} \rightarrow_d N(0, \Sigma)$$

for

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

This assumption implies that $(\hat{c}, \hat{\gamma})$ are regular (see e.g., Newey 1994) and is required for standard inference procedures based on $(\hat{c}, \hat{\gamma})$ to be valid under the model.

With these assumptions, we obtain a generalization of Lemmas 1 and 2.

Lemma 5. If Assumptions 3, 8, and 10 hold, then under $F_{h,\varphi}^n \left(\frac{\tau}{\sqrt{n}}, \frac{\mu}{\sqrt{n}} \right)$,

$$\sqrt{n} \begin{pmatrix} \hat{c} - c_{0,n} \\ \hat{\gamma} - \gamma_{0,n} \end{pmatrix} \rightarrow_d N \left(\begin{pmatrix} \mu \bar{c}_\varphi \\ \mu \bar{\gamma}_\varphi \end{pmatrix}, \Sigma \right),$$

for $(\bar{c}_\varphi, \bar{\gamma}_\varphi)$ and Σ as in Lemma 1.

Lemma 6. For $(\bar{c}_\varphi, \bar{\gamma}_\varphi)$ and Σ as defined in Lemma 1, Assumption 9 implies that for $\Delta < 1$ the set of values $(\bar{c}_\varphi, \bar{\gamma}_\varphi)$ associated with the set of all directions $\varphi \in \Phi$ is

$$\mathcal{B} = \left\{ (\bar{c}, \bar{\gamma}) \in \mathbb{R} \times \mathbb{R}^{p_\gamma} : \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} \bar{c} \\ \bar{\gamma} \end{pmatrix} \leq 1 \right\}$$

If $\Delta = 1$, this set of values is

$$\mathcal{B} = \left\{ (\bar{c}, \bar{\gamma}) \in \mathbb{R} \times \mathbb{R}^{p_\gamma} : \bar{c} = \Lambda \bar{\gamma}, \bar{\gamma}' \Sigma_{\gamma\gamma}^{-1} \bar{\gamma} \leq 1 \right\}$$

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

These lemmas verify the analogue of Conditions 1 and 2 for the setting with drifting η , so Proposition 1 immediately generalizes as well.

Proposition 7. Suppose Assumptions 3 and 8-10 hold, and consider the set of local perturbations $F_{h,\varphi}^n \left(\frac{\tau}{\sqrt{n}}, \frac{\mu}{\sqrt{n}} \right)$ with τ unrestricted and μ fixed. The asymptotic biases in \hat{c} associated with perturbations in this set are

$$\mathcal{B}^\mu = [-\sigma_c \mu, \sigma_c \mu].$$

The asymptotic biases in \hat{c} associated with the subset of these perturbations for which $\bar{\gamma}_\varphi = 0$ are

$$\mathcal{B}_0^\mu = [-\sigma_c \mu \sqrt{1-\Delta}, \sigma_c \mu \sqrt{1-\Delta}].$$

The ratio of the widths of these intervals is

$$\frac{|\mathcal{B}_0^\mu|}{|\mathcal{B}^\mu|} = \sqrt{1-\Delta}.$$

B.1.1 Proofs

Proof of Lemma 5 The proof for this result follows by the same argument as the proof of Lemma 1. Lemma 7.6 and Theorem 7.2 of van der Vaart (1998) imply that under F_0^n

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

and that $E_{F_0} [s_h(D_i)] = E_{F_0} [s_\varphi(D_i)] = 0$. By the same argument used in the proof of Lemma 1, Le Cam's Third Lemma then implies that under $F_{h,\varphi}^n \left(\frac{\tau}{\sqrt{n}}, \frac{\mu}{\sqrt{n}} \right)$,

$$(15) \quad \sqrt{n} \begin{pmatrix} \hat{c} - c_0 \\ \hat{\gamma} - \gamma_0 \end{pmatrix} \rightarrow_d N \left(\begin{pmatrix} \tau \bar{c}_h + \mu \bar{c}_\varphi \\ \tau \bar{\gamma}_h + \mu \bar{\gamma}_\varphi \end{pmatrix}, \Sigma \right),$$

for

$$\begin{pmatrix} \bar{c}_h \\ \bar{\gamma}_h \end{pmatrix} = \begin{pmatrix} E_{F_0} [\phi_c(D_i) s_h(D_i)] \\ E_{F_0} [\phi_\gamma(D_i) s_h(D_i)] \end{pmatrix}.$$

If we consider (15) in the special case with $\mu = 0$, Assumption 10 implies that under $F^n(\eta_{0,n})$,

$$\sqrt{n} \begin{pmatrix} \hat{c} - c_0 \\ \hat{\gamma} - \gamma_0 \\ \hat{c} - c_{0,n} \\ \hat{\gamma} - \gamma_{0,n} \end{pmatrix} \rightarrow_d N \left(\begin{pmatrix} \tau \bar{c}_h \\ \tau \bar{\gamma}_h \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & \Sigma \\ \Sigma & \Sigma \end{pmatrix} \right)$$

and hence that

$$\sqrt{n} \begin{pmatrix} c_{0,n} - c_0 \\ \gamma_{0,n} - \gamma_0 \end{pmatrix} \rightarrow \begin{pmatrix} \tau \bar{c}_h \\ \tau \bar{\gamma}_h \end{pmatrix}.$$

Consequently, under $F_{h,\varphi}^n \left(\frac{\tau}{\sqrt{n}}, \frac{\mu}{\sqrt{n}} \right)$,

$$\sqrt{n} \begin{pmatrix} \hat{c} - c_{0,n} \\ \hat{\gamma} - \gamma_{0,n} \end{pmatrix} \rightarrow_d N \left(\begin{pmatrix} \mu \bar{c}_\varphi \\ \mu \bar{\gamma}_\varphi \end{pmatrix}, \Sigma \right).$$

This completes the proof.

Proof of Lemma 6 This lemma follows by the same argument as in the proof of Lemma 2.

Proof of Proposition 7 Given the results of Lemmas 5 and 6, this proposition follows by the same argument as in the proof of Proposition 1.

B.2 Non-Local Misspecification

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

asymptotically linear functionals, in the sense that

$$(16) \quad \begin{aligned} \lim_{\vartheta \rightarrow 0} \left\| \tilde{c}(F_\varphi(\vartheta)) - c_0 - \vartheta E_{F_0} [s_\varphi \phi_c(D_i)] \right\| / \vartheta &= 0 \text{ for all } \varphi \in \Phi \\ \lim_{\vartheta \rightarrow 0} \left\| \tilde{\gamma}(F_\varphi(\vartheta)) - \gamma_0 - \vartheta E_{F_0} [s_\varphi \phi_\gamma(D_i)] \right\| / \vartheta &= 0 \text{ for all } \varphi \in \Phi. \end{aligned}$$

See Newey (1994) and Rieder (1994) for discussion.

Since (16) only restricts behavior as $\vartheta \rightarrow 0$ along fixed paths $F_\varphi(\cdot)$, rather than working with $\tilde{\Lambda}^{\bar{r}}$ as defined in the main text, here we instead work with misspecification measures defined using finite collections of paths. Specifically, for each $\varphi \in \Phi$ let

$$\bar{\vartheta}_\varphi(\bar{r}) = \inf \{ \vartheta \in \mathbb{R}_+ : r(F_0, F_\varphi(\vartheta)) \geq \bar{r} \}$$

denote the largest value of ϑ such that $r(F_0, F_\varphi(\vartheta)) < \bar{r}$ for all $\vartheta < \bar{\vartheta}_\varphi(\bar{r})$. Let $\Phi_+ \subset \Phi$ denote the set of values φ corresponding to non-trivial perturbations (that is, perturbations with $E_{F_0} [s_\varphi(D_i)^2] = 1$) which also satisfy Assumption 5. Let $\mathcal{Q} \subset \Phi_+$ denote a finite subset of Φ_+ , and let \mathcal{Q} denote the set of all such finite subsets. Finally, let

$$\tilde{\mathcal{B}}^{\bar{r}}(\mathcal{Q}) = \{ \tilde{c}(F_\varphi(\vartheta)) - c_0 : \varphi \in \mathcal{Q}, \vartheta < \bar{\vartheta}_\varphi(\bar{r}) \}$$

denote the analogue of $\tilde{\mathcal{B}}^{\bar{r}}$ based on the finite set of paths \mathcal{Q} , and for $\varepsilon > 0$ let

$$\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(\mathcal{Q}) = \left\{ \tilde{c}(F_\varphi(\vartheta)) - c_0 : \varphi \in \mathcal{Q}, \vartheta < \bar{\vartheta}_\varphi(\bar{r}), \|\tilde{\gamma}(F_\varphi(\vartheta)) - \gamma_0\| \leq \varepsilon \sqrt{\bar{r}} \right\},$$

denote the analogue of $\tilde{\mathcal{B}}_0^{\bar{r}}$ based on \mathcal{Q} which allows inconsistency in $\hat{\gamma}$ of at most $\varepsilon \sqrt{\bar{r}}$. Because $\tilde{\mathcal{B}}_0^{\bar{r}}(\mathcal{Q})$ may equal $\{0\}$ even for large \bar{r} due to the approximation error in (16), we consider limits as $\varepsilon \downarrow 0$ (i.e., as $\varepsilon \rightarrow 0$ from above).

Based on these objects, we define the analogue of $\tilde{\Delta}(\bar{r})$ as

$$\tilde{\Delta}(\bar{r}, \mathcal{Q}) = \sup_{Q_1 \in \mathcal{Q}} \inf_{Q_2 \in \mathcal{Q}} \lim_{\varepsilon \downarrow 0} \frac{|\text{Conv}(\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q_1))|}{|\text{Conv}(\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q_2))|},$$

provided the limit exists, where $\text{Conv}(A)$ denotes the convex hull of the set A . We work with convex hulls because the sets $\tilde{\mathcal{B}}^{\bar{r}}(Q)$ and $\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q)$ need not be intervals.

Proposition 8. *Suppose Assumptions 1-3 hold, and that the estimators \hat{c} and $\hat{\gamma}$ are regular. For $r(F_0, F) = E_{F_0} \left[\psi \left(\frac{dF}{dF_0} \right) \right]$ and ψ satisfying the conditions of Proposition 3 in the paper,*

$$\sup_{Q_1 \in \mathcal{Q}} \inf_{Q_2 \in \mathcal{Q}} \lim_{\varepsilon \downarrow 0} \lim_{\bar{r} \downarrow 0} \frac{|\text{Conv}(\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q_1))|}{|\text{Conv}(\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q_2))|} = \sqrt{1 - \Delta}.$$

It is important that we take the limit as $\bar{r} \downarrow 0$ inside the limit as $\varepsilon \downarrow 0$ and the sup and inf, since this order of limits allows us to take advantage of the approximation result (16).

Proof of Proposition 8 Note, first, that our Assumptions 1-3 imply the conditions of Theorem 2.1 of Newey (1994) other than regularity of $(\hat{c}, \hat{\gamma})$. Specifically, our Assumption 1 imposes that our paths are what Newey (1994) terms “regular.” Condition (i) of Theorem 2.1 in Newey (1994) is then immediate from our Assumption 2, up to the scale normalization we impose on the scores, which only affects the interpretation of the parameter ϑ . Condition (ii) follows from the same observation. Condition (iii) is implied by our Assumption 3. Regularity of $(\hat{c}, \hat{\gamma})$ is assumed, so Theorem 2.1 of Newey (1994) implies (16).

Note, next, that for any $\varphi \in \Phi_+$, the proof of Proposition 3 implies that

$$\lim_{\vartheta \downarrow 0} r(F_0, F_\varphi(\vartheta)) / \vartheta^2 = 1.$$

Hence, as $\bar{r} \downarrow 0$, $\bar{\vartheta}_\varphi(\bar{r}) / \sqrt{\bar{r}} \rightarrow 1$. For all $\varphi \in \Phi_+$, (16) implies that

$$\begin{aligned} \lim_{\bar{r} \downarrow 0} \sup_{\vartheta \leq \bar{\vartheta}_\varphi(\bar{r})} |\tilde{c}(F_\varphi(\vartheta)) - c_0 - \vartheta E_{F_0} [s_\varphi(D_i) \phi_c(D_i)]| / \vartheta &= 0 \\ \lim_{\bar{r} \downarrow 0} \sup_{\vartheta \leq \bar{\vartheta}_\varphi(\bar{r})} |\tilde{\gamma}(F_\varphi(\vartheta)) - \gamma_0 - \vartheta E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)]| / \vartheta &= 0, \end{aligned}$$

and thus that

$$\begin{aligned} &\left\{ \frac{1}{\sqrt{\bar{r}}} (\tilde{c}(F_\varphi(\vartheta)) - c_0, \tilde{\gamma}(F_\varphi(\vartheta)) - \gamma_0) : \vartheta \leq \bar{\vartheta}_\varphi(\bar{r}) \right\} \\ &\rightarrow \left\{ \tilde{\vartheta} (E_{F_0} [s_\varphi(D_i) \phi_c(D_i)], E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)]) : \tilde{\vartheta} \leq 1 \right\} \end{aligned}$$

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

$$\begin{aligned} &\left\{ \frac{1}{\sqrt{\bar{r}}} (\tilde{c}(F_\varphi(\vartheta)) - c_0, \tilde{\gamma}(F_\varphi(\vartheta)) - \gamma_0) : \varphi \in Q, \vartheta \leq \bar{\vartheta}_\varphi(\bar{r}) \right\} \\ &\rightarrow \left\{ \tilde{\vartheta} (E_{F_0} [s_\varphi(D_i) \phi_c(D_i)], E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)]) : \varphi \in Q, \tilde{\vartheta} \leq 1 \right\}. \end{aligned}$$

In particular, this implies that for $Q \in \mathcal{Q}$,

$$\frac{1}{\sqrt{\bar{r}}} \text{Conv}(\tilde{\mathcal{B}}^{\bar{r}}(Q)) \rightarrow \left[\left(\min_{\varphi \in Q} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] \right)_-, \left(\max_{\varphi \in Q} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] \right)_+ \right]$$

where $(A)_- = \min\{A, 0\}$ and $(A)_+ = \max\{A, 0\}$. Hence,

$$\frac{1}{\sqrt{\bar{r}}} |\text{Conv}(\tilde{\mathcal{B}}^{\bar{r}}(Q))| \rightarrow \left(\max_{\varphi \in Q} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] \right)_+ - \left(\min_{\varphi \in Q} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] \right)_-.$$

Matters are somewhat more delicate for $\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q)$. Note, in particular, that for $\varepsilon > 0$,

$$\begin{aligned} \frac{1}{\sqrt{\bar{r}}} \tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q) &\rightarrow \left\{ \tilde{\vartheta} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] : \varphi \in Q, \tilde{\vartheta} \leq 1, \tilde{\vartheta} \|E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)]\| \leq \varepsilon \right\} \\ &= \left\{ \tilde{\vartheta} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] : \varphi \in Q, \tilde{\vartheta} \leq \min \left\{ 1, \frac{\varepsilon}{\|E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)]\|} \right\} \right\}, \end{aligned}$$

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

$$\begin{aligned} &\frac{1}{\sqrt{\bar{r}}} |\text{Conv}(\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q))| \\ &\rightarrow \left| \left\{ \tilde{\vartheta} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] : \varphi \in Q, \tilde{\vartheta} \leq \min \left\{ 1, \frac{\varepsilon}{\|E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)]\|} \right\} \right\} \right|. \end{aligned}$$

Note, however, that by the Cauchy-Schwarz inequality and $E[s_\varphi(D_i)^2] \leq 1$ by Assumption 1, $E_{F_0}[s_\varphi(D_i) \phi_c(D_i)]$ is finite for all $\varphi \in \Phi$, so for any φ with $E_{F_0}[s_\varphi(D_i) \phi_\gamma(D_i)] \neq 0$,

$$\frac{\varepsilon}{\|E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)]\|} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] \rightarrow 0$$

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

$$\begin{aligned} &\left| \left\{ \tilde{\vartheta} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] : \varphi \in Q, \tilde{\vartheta} \leq \min \left\{ 1, \frac{\varepsilon}{\|E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)]\|} \right\} \right\} \right| \\ &\rightarrow |\{ \tilde{\vartheta} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] : \varphi \in Q_0, \tilde{\vartheta} \leq 1 \}|, \end{aligned}$$

for $Q_0 = \{ \varphi \in Q : E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)] = 0 \}$.

This immediately implies

$$\lim_{\varepsilon \downarrow 0} \lim_{\bar{r} \downarrow 0} \frac{1}{\sqrt{\bar{r}}} |\text{Conv}(\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q))| = \left(\max_{\varphi \in Q_0} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] \right)_+ - \left(\min_{\varphi \in Q_0} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] \right)_-.$$

Hence,

$$\lim_{\varepsilon \downarrow 0} \lim_{\bar{r} \downarrow 0} \frac{|\text{Conv}(\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q_1))|}{|\text{Conv}(\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q_2))|} = \frac{(\max_{\varphi \in Q_{1,0}} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)])_+ - (\min_{\varphi \in Q_{1,0}} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)])_-}{(\max_{\varphi \in Q_2} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)])_+ - (\min_{\varphi \in Q_2} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)])_-},$$

for $Q_{1,0} = \{ \varphi \in Q_1 : E_{F_0} [s_\varphi(D_i) \phi_\gamma(D_i)] = 0 \}$, provided the denominator on the right hand side

is non-zero.¹⁶

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

$$\begin{aligned} & \sup_{Q_1 \in \mathcal{Q}} \inf_{Q_2 \in \mathcal{Q}} \limlim_{\varepsilon \downarrow 0 \quad \bar{r} \downarrow 0} \frac{|\text{Conv}(\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q_1))|}{|\text{Conv}(\tilde{\mathcal{B}}^{\bar{r}}(Q_2))|} \\ &= \frac{\sup_{Q_0 \in \mathcal{Q}_0} \left\{ (\max_{\varphi \in Q_0} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)])_+ - (\min_{\varphi \in Q_0} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)])_- \right\}}{\sup_{Q \in \mathcal{Q}} \left\{ (\max_{\varphi \in Q} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)])_+ - (\min_{\varphi \in Q} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)])_- \right\}}. \end{aligned}$$

The proof of Proposition 5 shows, however, that

$$\max_{\varphi \in \Phi} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] = -\min_{\varphi \in \Phi} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] = \sigma_c$$

and

$$\max_{\varphi \in \Phi: E_{F_0} [s_\varphi \phi_\gamma(D_i)] = 0} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] = -\min_{\varphi \in \Phi: E_{F_0} [s_\varphi \phi_\gamma(D_i)] = 0} E_{F_0} [s_\varphi(D_i) \phi_c(D_i)] = \sigma_c \sqrt{1 - \Delta}.$$

Hence,

$$\sup_{Q_1 \in \mathcal{Q}} \inf_{Q_2 \in \mathcal{Q}} \limlim_{\varepsilon \downarrow 0 \quad \bar{r} \downarrow 0} \frac{|\text{Conv}(\tilde{\mathcal{B}}_\varepsilon^{\bar{r}}(Q_1))|}{|\text{Conv}(\tilde{\mathcal{B}}^{\bar{r}}(Q_2))|} = \sqrt{1 - \Delta},$$

as we wanted to show.

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

In Section 5, for cases where the function $c(\theta)$ depends on the distribution of exogenous covariates, our formulation implicitly treats those covariates as fixed at the sample distribution for the purposes of estimating Δ and Λ . Here we discuss how to allow for uncertainty in the distribution of covariates in a special case, and present corresponding calculations for our applications.

Suppose in particular that

$$(17) \quad \hat{c} = \frac{1}{n} \sum_i c(\hat{\theta}; D_i)$$

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

¹⁶If the denominator on the right hand side is zero, we define the limit as $+\infty$.

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

$$(18) \quad \hat{\phi}_c(D_i) = c(\hat{\theta}; D_i) + \hat{\Lambda}_{cg} \phi_g(D_i; \hat{\theta})$$

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

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

In the cases of Attanasio et al. (2012a) and Gentzkow (2007a), we can represent the calculation of \hat{c} in the form given in (17) and thus calculate $\hat{\Delta}$ using the modified estimated influence function in (18). In the case of Attanasio et al. (2012a), the estimates in Table 1 change from 0.283, 0.227, and 0.056, respectively, to 0.277, 0.221, and 0.055. In the case of Gentzkow (2007a), the estimates in Table 2 change from 0.514, 0.009, and 0.503, respectively, to 0.517, 0.008, and 0.507.

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

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

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

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

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

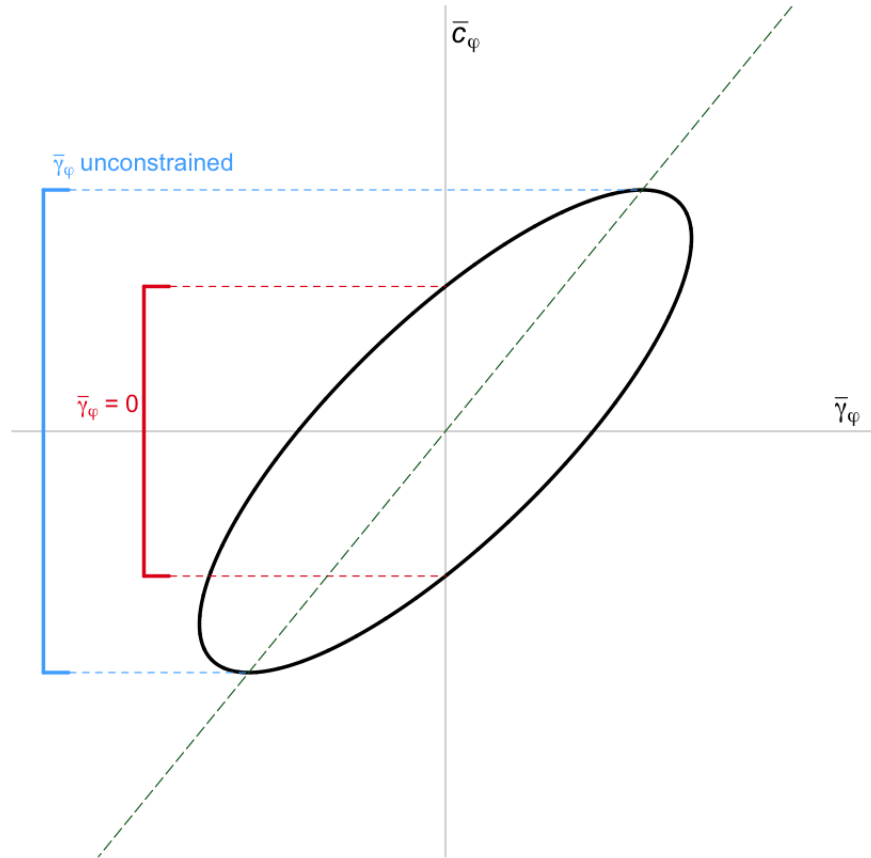
Notes: The table shows the estimated informativeness $\hat{\Delta}$ of three vectors $\hat{\gamma}$ of descriptive statistics for the estimated effect \hat{c} on the readership of the *Post* print edition if the *Post* online edition were removed from the choice set (Gentzkow 2007a, table 10, row labeled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

Table 3: Estimated informativeness of descriptive statistics for key estimates from Hendren (2013a)

Descriptive statistics $\hat{\gamma}$	Estimated informativeness $\hat{\Delta}$ for	
	Fraction focal point respondents	Minimum pooled price ratio
All	0.627	0.700
Fractions in focal point groups	0.351	0.005
Fractions in non-focal point groups	0.300	0.018
Fraction in each group needing LTC	0.076	0.676

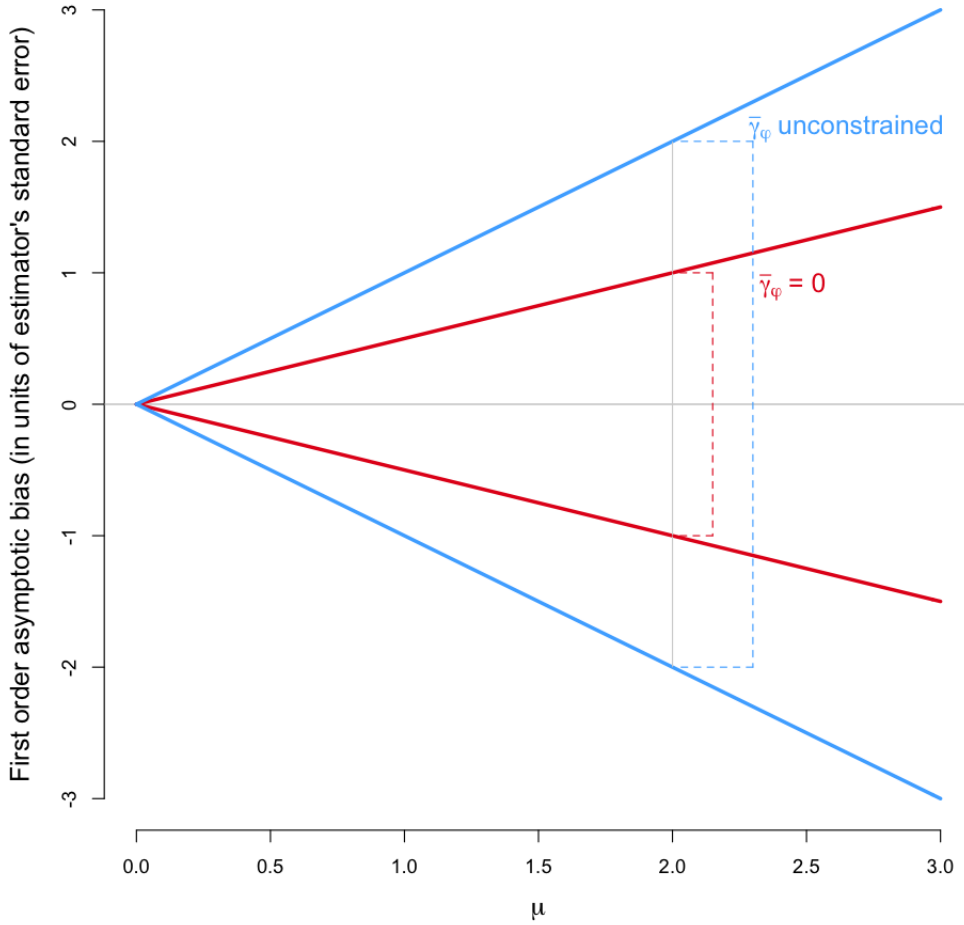
Notes: The table shows the estimated informativeness $\hat{\Delta}$ of four vectors $\hat{\gamma}$ of descriptive statistics for each of two estimates \hat{c} of interest from Hendren (2013a). The first estimate of interest \hat{c} is the “fraction focal point respondents” (Hendren 2013a, Table A-V, row labeled “Fraction focal respondents,” column labeled “LTC-Reject”). The second estimate of interest \hat{c} is the “minimum pooled price ratio” (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.

Figure 1: Set of possible values of $(\bar{c}_\varphi, \bar{\gamma}_\varphi)$ under local perturbations



Notes: Figure shows the ellipsoid defined in (3) for the case with $p_\gamma = 1$ and $\Delta = 0.64$. The interval labeled “ $\bar{\gamma}_\varphi$ unconstrained” characterizes the set of all possible values of \bar{c}_φ . The interval labeled “ $\bar{\gamma}_\varphi = 0$ ” characterizes the set of possible values of \bar{c}_φ when $\bar{\gamma}_\varphi = 0$. The diagonal line through the ellipsoid connects the points on the ellipsoid that achieve the minimum and maximum values of \bar{c}_φ .

Figure 2: Asymptotic bias \bar{c}_φ of estimator under local perturbations



Notes: Figure shows the limits of the intervals \mathcal{B}^μ (labeled " $\bar{\gamma}_\varphi$ unconstrained") and \mathcal{B}_0^μ (labeled " $\bar{\gamma}_\varphi = 0$ ") defined in Proposition 1 for the case of $\Delta = 0.75$. The intervals are normalized by the standard error σ_c , so that a value of 1 on the y-axis indicates a first-order asymptotic bias of one standard error σ_c .