On the Informativeness of Descriptive Statistics for Structural Estimates

Isaiah Andrews, MIT and NBER∗
Matthew Gentzkow, Stanford University and NBER
Jesse M. Shapiro, Brown University and NBER

June 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@mit.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), and the Silverman (1968) Family Career Development Chair at MIT. We thank Matias Cattaneo, Gary Chamberlain, Nathan Hendren, Yuichi Kitamura, Ariel Pakes, and participants at the Radcliffe Institute Conference on Statistics When the Model is Wrong 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 2017).

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 $(c_0, \gamma_0)$. 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 $\gamma_0$—for example, because it comes from a randomized experiment and the economic assumptions linking $\gamma_0$ 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 misspecified 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 \( \varphi \) (capturing the nature of the misspecification) and a magnitude \( \mu \) (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 \( \mu \) and (ii) the maximal asymptotic bias across the subset of these perturbations under which the descriptive statistics \( \hat{\gamma} \) are asymptotically unbiased estimators of \( \gamma_0 \). 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 \( \gamma_0 \).

Our main result shows that this ratio depends only on a scalar \( \Delta \), 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 \( \hat{c} \), and 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 \( \hat{\gamma} \), and 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 \( \Delta \) 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 \( \Delta \) is trivial.

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

As a more concrete illustration of our proposed approach, consider an example based on At-
tanasio 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 child’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 $\gamma$ 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 low, and confidence in the validity of the experiment and the way the model relates $\gamma$ 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 $\gamma_0$ 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 $\gamma_0$ of the model. In the example above, this means that the function $\tilde{c}(\cdot)$ is correctly specified. With a little reflection, the need for this requirement should be obvious. 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 $\Delta$ 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 $\Lambda$ of the sensitivity of a parameter estimate $\hat{c}$ to a vector of statistics $\hat{y}$, focusing on the case where $\hat{y}$ are estimation moments that fully determine the estimator $\hat{c}$ (and so $\Delta = 1$). Here, we ask the complementary question of to what extent a vector of descriptive statistics $\hat{y}$ determines the value of $\hat{c}$. In an extension, we generalize our main result to allow $\hat{y}$ to have possibly nonzero asymptotic bias $\bar{y}$. In this case, the set of possible asymptotic biases in $\hat{c}$ is an interval centered at $\Lambda \bar{y}$, 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 $\hat{y}$ 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 $\hat{y}$ are linear regression coefficients. In the third application, to Hendren (2013a), the parameter estimates $\hat{c}$ of interest are a model parameter and a key output governing the existence of insurance markets, and the descriptive statistics $\hat{y}$ summarize the joint distribution of self-reported probabilities and loss events. In each case, our choice of $\hat{y}$ is guided by the authors’ discussion. In each case, we report an estimate of $\Delta$ for various combinations of $\hat{c}$ and $\hat{y}$, and we discuss the implications for a reader’s confidence in the conclusions. These applications illustrate how estimates of $\Delta$ could 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 (e.g. Berkowitz et al. 2008; Guggenberger 2012) or robust estimation (e.g., Rieder 1994; Kitamura et al. 2013; Bonhomme and Weidner 2018). Rieder (1994) studies the choice of target parameters and

---

1Our work here draws on the analysis of “sensitivity to descriptive statistics” in Gentzkow and Shapiro (2015).
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. In contrast to the previous literature, our focus is 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). It is a common practice in applied research to describe the data features that “primarily identify” structural parameters or “drive” estimates of those parameters. As Keane (2010) and others have noted, such statements are not directly related to the formal notion of identification in econometrics. Their intended meaning is therefore up for grabs. If researchers are prepared to reinterpret these as statements linking correct specification of descriptive statistics to confidence in related structural estimates, then our approach provides a way to sharpen and quantify these statements at low cost to researchers.

2 Setup and Main Result

2.1 Setup

A researcher observes an i.i.d. sample $D_i \in \mathcal{D}$ for $i = 1, \ldots, n$. The researcher’s model implies that $D_i \sim F(\eta)$, for $\eta \in \mathcal{H}$ 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 $\eta_0$ denote the base value of $\eta$.

The key quantity of interest $c = c(\eta)$ is a scalar that may be an element of $\eta$, 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 $p_r \times 1$ vector of descriptive statistics $\hat{\gamma}$.

Andrews et al. (2017, footnotes 2 and 3) provide examples.
We assume that under $F_n^0$, the estimator $\hat{c}$ and the descriptive statistics $\hat{\gamma}$ are jointly asymptotically normal,

\[
\sqrt{n} \begin{pmatrix} \hat{c} - c_0 \\ \hat{\gamma} - \gamma_0 \end{pmatrix} \to_d N(0, \Sigma),
\]

where $\gamma_0$ is the probability limit of $\hat{\gamma}$ under $F_n^0$ and the asymptotic variance $\Sigma$ is finite. We assume throughout that the asymptotic variances $\sigma^2_c$ and $\Sigma_{\gamma\gamma}$ corresponding to $\hat{c}$ and $\hat{\gamma}$ are positive and full-rank, respectively. For convenience, we further assume for the moment that $\Sigma$ has full rank; we derive our main results without this assumption in Section 3. We let $\Sigma_{c\gamma}$ denote the off-diagonal block of $\Sigma$ corresponding to the asymptotic covariance of $\hat{c}$ and $\hat{\gamma}$.

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

\[
\Delta = \frac{\Sigma_{c\gamma} \Sigma^{-1}_{\gamma\gamma}}{\sigma^2_c}.
\]

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 $F_n^0$. It is immediate that $\Delta \in [0, 1]$.\(^3\)

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 $\mu$ is defined so that a perturbation of magnitude $\mu = 0$ coincides with the base distribution, and so that the ranking of perturbations implied by $\mu$ 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 $\mu$ has an interpretation in terms of the most powerful test of the base distribution against a perturbation.

More precisely, for each direction $\varphi$ in the possibly infinite-dimensional set $\Phi$, we define a

\(^3\)The assumption that $\Sigma$ has full rank formally rules out the case with $\Delta = 1$, but Section 3 derives our main results without this restriction. For the remainder of this section, when we discuss the case of $\Delta = 1$, it is with this understanding in mind.

\(^4\)Rieder (1994) defines a subset of these perturbations called “simple perturbations” and derives results for this class.
family of distributions $F_\phi(\vartheta)$ indexed by a scalar parameter $\vartheta \in \mathbb{R}^+$ such that $F_\phi(0) = F_0$. Each $F_\phi(\cdot)$ 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 $\varphi$ and magnitude $\mu$ is then defined to be the sequence of joint distributions $F_\phi^n\left(\frac{\mu}{\sqrt{n}}\right) = \times_n F_\phi\left(\frac{\mu}{\sqrt{n}}\right)$. When $\mu > 0$, the model may be misspecified in the sense that $F_\phi\left(\frac{\mu}{\sqrt{n}}\right) \not\in \{F(\eta) : c(\eta) = c_0, \eta \in H\}$, or even $F_\phi\left(\frac{\mu}{\sqrt{n}}\right) \not\in \{F(\eta) : \eta \in H\}$. 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 $c_0$, $\eta_0$, and $\gamma_0$ 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, $\eta$ 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 $\eta_0$ 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, $\gamma_0$ denotes the probability limit of $\hat{\gamma}$ under the base distribution, while $\gamma$ may or may not be defined outside the model.

**Example.** Say that $\eta$ is finite-dimensional and that $\hat{\gamma}$ are regression coefficients estimated on the sample, such as treatment-control differences from a randomized experiment. The quantity $c$ describes an unobserved counterfactual and is estimated with $\hat{c} = c(\hat{\eta})$ where $\hat{\eta}$ is an estimator of $\eta$ 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 $\hat{\eta} = \eta(\hat{\gamma}) + o_p\left(\frac{1}{\sqrt{n}}\right)$ where $\eta$ 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 \( \varphi \) and magnitude \( \mu \),

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

for some \( \bar{c}_\varphi \in \mathbb{R} \) and \( \bar{\gamma}_\varphi \in \mathbb{R}^p \), and for \( \Sigma \) 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 \( \Sigma \) 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 \( \mu \) of the perturbation and a term \( \bar{c}_\varphi \) that depends only on its direction.

**Condition 2.** \( \Sigma \) has full rank and the set of values \( (\bar{c}_\varphi, \bar{\gamma}_\varphi) \) associated with the set of all directions \( \varphi \in \Phi \) is

\[
\mathcal{B} = \left\{ (\bar{c}, \bar{\gamma}) \in \mathbb{R} \times \mathbb{R}^p : \sqrt{\left( \begin{array}{c} \bar{c} \\ \bar{\gamma} \end{array} \right)} \Sigma^{-1} \left( \begin{array}{c} \bar{c} \\ \bar{\gamma} \end{array} \right) \leq 1 \right\}.
\]

This set 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 \( \bar{c}_\varphi \) 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 \( \bar{c}_\varphi \) 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").

Our main result shows that the widths of these two intervals in the general case are simple functions of informativeness \( \Delta \), the magnitude of the perturbation \( \mu \), and the standard error \( \sigma_c \), and that their ratio depends only on \( \Delta \). Proofs for this, and other results, are in the Appendix.
Proposition 1. Suppose Conditions 1 and 2 hold, and consider the set of local perturbations with magnitude $\mu$. The asymptotic biases in $\hat{c}$ associated with perturbations in this set are

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

The asymptotic biases in $\hat{c}$ associated with the subset of these perturbations for which $\gamma_\phi \mu = 0$ are

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

The ratio of the widths of these intervals is

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

Thus, the extent to which knowing that $\gamma_\phi = 0$ 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 $\gamma_0$ tells us nothing about potential bias in $\hat{c}$. When $\Delta = 1$, knowing that $\hat{\gamma}$ is asymptotically unbiased for $\gamma_0$ 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 $\mathbb{B}^\mu$ and $\mathbb{B}_0^\mu$ (normalized with respect to the standard error $\sigma_c$ of $\hat{c}$) as a function of $\mu$. 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 $\gamma_0$ reduces the width of this interval to plus or minus $2\sqrt{1 - 0.75} = 1$ standard error. We return to the interpretation of $\mu$ 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 $\Delta$. We show below that it can be trivial to estimate $\Delta$ 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 $\hat{\gamma}$ should translate into confidence in the researcher’s inferences about $c_0$.

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 ($\gamma_\phi = 0$), then there is no asymptotic bias in $\hat{c}$ ($\gamma_\phi = 0$). Intuitively, this is because $\gamma_\phi = 0$
implies both that \( \hat{\gamma} \) is asymptotically unbiased for \( \gamma_0 \) and that the relation \( c_0 = c(\gamma(\gamma_0)) \) is correctly specified.

2. (MLE) Here, \( \Delta < 1 \). Therefore, by Proposition 1, \( \left| \frac{\partial \mu}{\partial \mu} \right| = \sqrt{1-\Delta} > 0 \). Knowing that \( \bar{\gamma}_\varphi = 0 \) 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 \( \bar{\gamma}_\varphi = 0 \) does not restrict \( \bar{\gamma}_\varphi \) at all. Intuitively, this is because \( \hat{c} \) depends entirely on features of the data unrelated to \( \hat{\gamma} \).

3 Regularity Conditions and Scaling of Perturbations

3.1 Regularity Conditions

This subsection develops asymptotic results that provide the foundation for our analysis. We introduce assumptions that imply Conditions 1 and 2, and then show that under these assumptions Proposition 1 generalizes to the case where \( \Sigma \) has reduced rank.

Perturbations

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

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

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

**Assumption 1.** For \( \vartheta \) in an open neighborhood of zero and all \( \varphi \in \Phi \): (i) \( \sqrt{f_\varphi(d; \vartheta)} \) is continuously differentiable with respect to \( \vartheta \) for all \( d \in \mathcal{D} \); (ii) \( I_\varphi(\vartheta) \) is finite and continuous in \( \vartheta \); 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 $\theta$ close to zero, allows us to interpret the magnitude of $\theta$ as a measure of the degree of misspecification.\(^5\)

We consider local perturbations $F^{\mu}_{\phi}\left(\frac{\mu}{\sqrt{n}}\right)$, 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 $\theta$, provided the usual regularity conditions for this bound hold.\(^6\) As we show in Section 3.2 below, this in turn implies that the magnitude $\mu$ 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 $\frac{\partial}{\partial \theta} \left(\log f_{\phi}(D_i;0)\right)$ includes all those consistent with Assumption 1, in the sense that for any $s(\cdot)$ with $E_{F_0}[s(D_i)] = 0$ and $E_{F_0}\left[s(D_i)^2\right] \in \{0,1\}$ there exists $\phi \in \Phi$ with $E_{F_0}\left[(s(D_i) - \frac{\partial}{\partial \phi} \log (f_{\phi}(D_i;0)))^2\right] = 0$.

Assumption 2 requires that $\Phi$ imply a rich enough set of $F_{\phi}(\cdot)$ that the corresponding score functions include all those consistent with Assumption 1.\(^7\) 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 $\Phi$. 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 $F^n_0$,

\[
(4)\quad \sqrt{n}(\hat{c} - c_0, \hat{\gamma} - \gamma_0) = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \phi_{\phi}(D_i), \sum_{i=1}^{n} \phi_{\gamma}(D_i)\right) + o_p(1),
\]

\(^5\)Formally, if we begin with a perturbation $F_{\phi}(\tilde{\theta})$ satisfying parts (i) and (ii) of Assumption 1 but not part (iii), and define $F_{\phi}(\theta) = \tilde{F}_{\phi}(\frac{\theta}{\sqrt{\tilde{I}_{\phi}(0)}})$ for $\tilde{I}_{\phi}(\tilde{\theta})$ the information matrix for $\tilde{\theta}$, then $F_{\phi}(\cdot)$ satisfies parts (i)-(iii).

\(^6\)When the information matrix is zero, this lower bound is infinite.

\(^7\)That 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).
for functions $\phi_c(D_i)$ and $\phi_\gamma(D_i)$ such that $E_{F_0} \left[ \phi_c(D_i) \right] = 0$, $E_{F_0} \left[ \phi_\gamma(D_i) \right] = 0$, $E_{F_0} \left[ \phi_c(D_i)^2 \right] = \sigma_c^2$, and $E_{F_0} \left[ \phi_\gamma(D_i) \phi_\gamma(D_i)' \right] = \Sigma_{\gamma\gamma}$, for $\sigma_c^2$ 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 (4) holds for an extremely wide range of estimators (see e.g. Ichimura and Newey 2017). 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 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. Define

$$s_\phi(D_i) = \frac{\partial}{\partial \vartheta} \log \left( f_{\phi}(D_i; \vartheta) \right) = \frac{\frac{\partial}{\partial \vartheta} f_{\phi}(D_i; \vartheta)}{f_{\phi}(D_i; \vartheta)}$$

to be the score of $f_{\phi}(D_i; \vartheta)$ with respect to $\vartheta$, evaluated at $\vartheta = 0$. We first verify Condition 1.

**Lemma 1.** Provided Assumptions 1 and 3 hold, under $F^n_\varphi \left( \frac{\mu}{\sqrt{n}} \right)$

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

where

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

and

$$\Sigma = \begin{pmatrix} \sigma_c^2 & \Sigma_{c\gamma} \\ \Sigma_{\gamma c} & \Sigma_{\gamma\gamma} \end{pmatrix} = \begin{pmatrix} E_{F_0} \left[ \phi_c(D_i)^2 \right] & E_{F_0} \left[ \phi_c(D_i) \phi_\gamma(D_i)' \right] \\ E_{F_0} \left[ \phi_\gamma(D_i) \phi_c(D_i) \right] & E_{F_0} \left[ \phi_\gamma(D_i) \phi_\gamma(D_i)' \right] \end{pmatrix}.$$
The asymptotic distributions of \( \sqrt{n}(\hat{c} - c_0) \) and \( \sqrt{n}(\hat{\gamma} - \gamma_0) \), on the other hand, are shifted by \( \mu \bar{c}_\phi \) and \( \mu \bar{\gamma}_\phi \) respectively, where the effect on each estimator depends on the covariance of its influence function with the score.\(^8\) We next verify Condition 2.

**Lemma 2.** If the matrix \( \Sigma \) has full rank and Assumption 2 holds, then for \((\bar{c}_\phi, \bar{\gamma}_\phi)\) as defined in Lemma 1

\[
\{ (\bar{c}_\phi, \bar{\gamma}_\phi) : \phi \in \Phi \} = \{ (\bar{c}, \bar{\gamma}) \in \mathbb{R} \times \mathbb{R}^p : \sqrt{\left( \begin{array}{c} \bar{c} \\ \bar{\gamma} \end{array} \right)'} \Sigma^{-1} \left( \begin{array}{c} \bar{c} \\ \bar{\gamma} \end{array} \right) \leq 1 \}.
\]

Lemma 2 shows that Assumption 2 implies Condition 2. Without Assumption 2 we can still show that the left hand side of (6) 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.

**Main Result**

Finally, we restate our main result, generalized to allow the possibility that \( \Sigma \) does not have full rank (as in the case when \( \Delta = 1 \)). This result differs from Proposition 1 in that it assumes the technical conditions in this section hold, but does not assume Lemma 2 (which requires full rank of \( \Sigma \)).

**Proposition 2.** Suppose Assumptions 1-3 are satisfied but \( \Sigma \) may not have full rank. Then the conclusions of Proposition 1 hold.

### 3.2 Scaling of Perturbations

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

\(^8\)Lemma 1 considers behavior under local perturbations \( F_\phi^n \left( \frac{\mu}{\sqrt{n}} \right) \). 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 \( F_0 \).
This section shows that this interpretation is justified under two natural measures for the degree of misspecification. First, we show that \( \mu \) determines the highest possible asymptotic power for any test to distinguish the perturbation \( F^n_\phi \left( \frac{\mu}{\sqrt{n}} \right) \) from the base distribution \( F^n_0 \). Second, we show that for a family of divergences measuring the difference between \( F_\phi \left( \frac{\mu}{\sqrt{n}} \right) \) and \( F_0 \), including Kullback-Leibler divergence and squared Hellinger divergence, the scaled divergence converges to \( \mu^2 \) under mild regularity conditions.

Throughout this section we limit attention to non-trivial perturbations, which we define as those with \( \phi \) such that \( E_{F_0} \left[ s_\phi (D_i)^2 \right] \neq 0 \), for \( s_\phi (D_i) \) the score as defined in (5). For trivial perturbations one can show (see the proof of Proposition 6) that \( \bar{c}_\phi = \bar{\gamma}_\phi = 0 \), 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 \( F^n_\phi \left( \frac{\mu}{\sqrt{n}} \right) \) is the power of tests of \( F^n_0 \) against \( F^n_\phi \left( \frac{\mu}{\sqrt{n}} \right) \).

**Proposition 3.** Under Assumption 1 the most powerful level \( \alpha \) test of the null hypothesis \( H_0 : (D_1, \ldots, D_n) \sim F^n_0 \) against the non-trivial alternative \( H_1 : (D_1, \ldots, D_n) \sim F^n_\phi \left( \frac{\mu}{\sqrt{n}} \right) \) has asymptotic power \( 1 - F_{N(0,1)} (c_\alpha - \mu) \) for \( c_\alpha \) the \( 1 - \alpha \) quantile of the standard normal distribution.

The proof of Proposition 3 shows that the most powerful test corresponds asymptotically to a z-test, where the z-statistic has mean \( \mu \) 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 \( \mu \) on a conventional scale.

The optimal test derived in the proof of Proposition 3 relies on knowledge of \( F_0 \) and so is infeasible in practice. However, the problem of testing \( F^n_0 \) against \( F^n_\phi \left( \frac{\mu}{\sqrt{n}} \right) \) 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 \( F^n_\phi \left( \frac{\mu}{\sqrt{n}} \right) \). Hence, the power bound in Proposition 3 is an upper bound on the power of tests of either the hypothesis that the DGP 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 \( F_\phi \left( \frac{\mu}{\sqrt{n}} \right) \), and hence for the degree of misspecification. The Cressie-Read (1984) family of divergences nests
many widely studied measures, including 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).

The Cressie-Read divergence between \( F_0 \) and \( F_\varphi (\vartheta) \) takes the form

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

for

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

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

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

and

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

are finite.

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

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

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

Proposition 4 shows that nontrivial perturbations have scaled divergence converging to \( \mu^2 \), where we scale by \( n \) to ensure a nontrivial limit. Hence, in large samples Cressie-Read (1984) divergences yield the same ranking over perturbations as that implied by \( \mu \).
4 Sensitivity and Informativeness

Propositions 1 and 2 consider the effect of limiting attention to perturbations with $\gamma = 0$. 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

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

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 $F_{\gamma 0}^n$. The sensitivity $\Lambda$ is simply the vector of coefficients from the population regression of $\tilde{c}$ on $\tilde{\gamma}$. An element $\Lambda_j$ of $\Lambda$ is the effect of changing the realization of a particular $\gamma_j$ 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., $\mu \gamma$) to the effect on $\hat{c}$ (i.e., $\mu c$).

**Proposition 5.** (Andrews et al. 2017) For $\hat{c} = c(\hat{\eta})$, $\hat{\eta}$ a minimum distance estimator satisfying Assumption 6 below, and $\hat{\gamma} = \hat{g}(\eta_0)$, under Assumption 1 and local perturbations $F_{\gamma 0}^n \left( \frac{\mu}{\sqrt{n}} \right)$ we have $\mu \hat{c}_\phi = \Lambda \mu \hat{\gamma}_\phi$.

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

**Proposition 6.** Suppose that Assumptions 1-3 hold, and consider the set of local perturbations with magnitude $\mu$ and $\mu \gamma_\phi = \gamma$. The asymptotic biases in $\hat{c}$ associated with perturbations in this set are

$$B^\mu_\gamma = \left[ \Lambda \gamma - \sigma_c (\gamma) \sqrt{1 - \Delta}, \Lambda \gamma + \sigma_c (\gamma) \sqrt{1 - \Delta} \right],$$
for \( \mu (\bar{\gamma}) = \sqrt{\mu^2 - \bar{\gamma}' \Sigma_{\bar{\gamma}}^{-1} \bar{\gamma}} \), provided \( \mu^2 - \bar{\gamma}' \Sigma_{\bar{\gamma}}^{-1} \bar{\gamma} \geq 0 \). If \( \mu^2 - \bar{\gamma}' \Sigma_{\bar{\gamma}}^{-1} \bar{\gamma} < 0 \), then \( \mathcal{B}_\bar{\gamma}^\mu = \emptyset \).

The ratio of the width of \( \mathcal{B}_\bar{\gamma}^\mu \) to the width of the interval when \( \gamma_\phi \) is unconstrained is

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

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 6 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 Propositions 1 and 2 to perturbations with a non-zero effect on \( \hat{\gamma} \), so \( \mu \hat{\gamma}_\phi \neq 0 \). The resulting set of first-order asymptotic biases \( \mu \hat{c}_\phi \) for \( \hat{c} \) is centered at \( \Lambda \bar{\gamma} \) with width proportional to \( \sqrt{1-\Delta} \).

Unlike in Propositions 1 and 2, the degree of misspecification now enters the width through \( \mu(\bar{\gamma}) = \sqrt{\mu^2 - \bar{\gamma}' \Sigma_{\bar{\gamma}}^{-1} \bar{\gamma}} \). Intuitively, \( \mu(\bar{\gamma}) \) measures the degree of excess misspecification beyond \( \bar{\gamma}' \Sigma_{\bar{\gamma}}^{-1} \bar{\gamma} \), which is the minimum necessary to allow \( \mu \hat{\gamma}_\phi = \bar{\gamma} \). 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 \( \mu \) and first-order bias \( \bar{\gamma} \) for \( \hat{\gamma} \) with the range under unrestricted perturbations with magnitude \( \mu(\bar{\gamma}) \), and finds that the ratios of their widths is \( \sqrt{1-\Delta} \). This 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 \( \mu \) 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 \( \Lambda \) is the sensitivity measure of Andrews et al. (2017). Since \( \Delta = 1 \), Proposition 6 (or the main result of Andrews et al. 2017) implies that \( \mu \hat{c}_\phi = \Lambda \mu \hat{\gamma}_\phi \). 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 \( \Lambda \) will depend on the definition of \( \hat{c} \). Since \( \Delta < 1 \), however, the value of \( \mu \hat{\gamma}_\phi \) does not fully determine the asymptotic bias of \( \hat{c} \). The range of asymptotic biases for \( \hat{c} \) is
centered at $\Lambda \mu \hat{\gamma}_\phi$, and has width that depends on $\Delta$ and the degree of excess misspecification $\mu(\hat{\gamma})$.

3. (Irrelevant descriptive statistics) Here $\Lambda = 0$. Since $\Delta = 0$ as well, knowing that $\mu \hat{\gamma}_\phi = \gamma$ does not restrict the asymptotic bias of $\hat{c}$ except through $\mu(\hat{\gamma})$.

5 Implementation

Given an estimate $\hat{\Sigma}$ of $\Sigma$, 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 $\Sigma$ under $F_n^0$ are available following standard asymptotic results (e.g., Newey and McFadden 1994). Given such an estimate one can construct plug-in estimates

$$
\hat{\Delta} = \frac{\hat{\Sigma}_{c}\hat{\gamma}_{\phi}^{-1}\hat{\phi}'}{\sigma_c^2}, \quad \hat{\Lambda} = \hat{\Sigma}_{c}\hat{\gamma}\hat{\gamma}^{-1}.
$$

Provided $\hat{\Sigma}$ is consistent under $F_n^0$, 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 5.** $\hat{\Sigma} \xrightarrow{p} \Sigma$ under $F_n^0$.

**Proposition 7.** Under Assumptions 1 and 5, $\hat{\Sigma} \xrightarrow{p} \Sigma, \hat{\Delta} \xrightarrow{p} \Delta$, and $\hat{\Lambda} \xrightarrow{p} \Lambda$ under any local perturbation $F_n^0\left(\frac{\mu}{\sqrt{n}}\right)$.

5.1 Implementation with Minimum Distance Estimators

We have so far imposed only high-level assumptions (specifically Assumptions 3 and 5) on $\hat{c}, \hat{\gamma}, \text{and } \hat{\Sigma}$. 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 $\eta$ 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 5 in this setting.
Formally, suppose that we can decompose \( \eta = (\theta, \omega) \) where \( \theta \) is finite-dimensional and \( c(\eta) \) depends on \( \eta \) only through \( \theta \). We can then, in a slight abuse of notation, write \( c_0 = c(\theta_0) \), where \( \theta_0 \) is defined by \( \eta_0 = (\theta_0, \omega_0) \).

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

\[
\min_{\theta \in \Theta} \hat{g}(\theta)' \hat{W} \hat{g}(\theta)
\]

for \( \Theta \) 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

\[
\min_{\gamma \in \Gamma} \hat{m}(\gamma)' \hat{U} \hat{m}(\gamma),
\]

for \( \Gamma \) 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 6.** 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) \( \hat{W} \to_p W \) for a positive semi-definite matrix \( W \); (iii) \( g(\theta)' W g(\theta) \) has a unique minimum at \( \theta_0 \); (iv) \( \theta_0 \) lies in the interior of \( \Theta \); (v) \( \sqrt{n} \hat{g}(\theta_0) \to_d N(0, \Sigma_{gg}) \) for full-rank \( \Sigma_{gg} \); 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) \( \hat{U} \to_p U \) for a positive semi-definite matrix \( U \); (iii) \( m(\gamma)' U m(\gamma) \) has a unique minimizer \( \gamma_0 \); (iv) \( \gamma_0 \) lies in the interior of \( \Gamma \); (v) \( \sqrt{n} \hat{m}(\gamma_0) \to_d N(0, \Sigma_{mm}) \) for full-rank \( \Sigma_{mm} \); 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

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

20
where \( \phi_g(D_i) \) and \( \phi_m(D_i) \) have mean zero and variance \( \Sigma_{gg} \) and \( \Sigma_{mm} \).

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

\[
\frac{1}{n} \left( \begin{array}{cc}
\sum_i \left[ \hat{\phi}_g(D_i) \hat{\phi}_g(D_i)' \right] & \sum_i \left[ \hat{\phi}_g(D_i) \hat{\phi}_m(D_i)' \right] \\
\sum_i \left[ \hat{\phi}_m(D_i) \hat{\phi}_g(D_i)' \right] & \sum_i \left[ \hat{\phi}_m(D_i) \hat{\phi}_m(D_i)' \right]
\end{array} \right) \xrightarrow{p} \left( \begin{array}{cc}
\Sigma_{gg} & \Sigma_{gm} \\
\Sigma_{mg} & \Sigma_{mm}
\end{array} \right),
\]

where

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

Parts (A) and (B) of Assumption 6 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 6 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 6 implies Assumption 3:

**Lemma 3.** Under Assumption 6 parts (A)-(D), Assumption 3 holds for \( \phi_c(D_i) = \Lambda_{cg} \phi_g(D_i) \) and \( \phi_\gamma(D_i) = \Lambda_{gm} \phi_m(D_i) \), where \( \Lambda_{cg} = -C (G'WG)^{-1} G'W \) and \( \Lambda_{gm} = -(M'UM)^{-1} M'U \) are the sensitivities of \( c \) with respect to \( g(\theta_0) \) and of \( \gamma \) with respect to \( m(\gamma_0) \) as defined in Andrews et al. (2017). Thus, \( \sigma_c^2 = \Lambda_{cg} \Sigma_{cg} \Lambda_{cg}' \) and \( \Sigma_{\gamma\gamma} = \Lambda_{gm} \Sigma_{mm} \Lambda_{gm}' \), \( \sigma_c^2 > 0 \), and \( \Sigma_{\gamma\gamma} \) has full rank.

Assumption 6 likewise implies Assumption 5:

**Lemma 4.** Under Assumption 6 parts (A)-(E), Assumption 5 holds for

\[
\hat{\Sigma} = \frac{1}{n} \left( \begin{array}{cc}
\sum_i \left[ \hat{\phi}_c(D_i) \hat{\phi}_c(D_i)' \right] & \sum_i \left[ \hat{\phi}_c(D_i) \hat{\phi}_\gamma(D_i)' \right] \\
\sum_i \left[ \hat{\phi}_\gamma(D_i) \hat{\phi}_c(D_i)' \right] & \sum_i \left[ \hat{\phi}_\gamma(D_i) \hat{\phi}_\gamma(D_i)' \right]
\end{array} \right),
\]

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

\[
\hat{\phi}_c(\cdot) = \Lambda_{cg} \hat{\phi}_g(\cdot) = -C \left( \hat{G}' \hat{W} \hat{G} \right)^{-1} \hat{G}' \hat{W} \hat{\phi}_g(\cdot)
\]
and
\[ \hat{\phi}_g(\cdot) = \hat{\Lambda}_g \hat{\phi}_m(\cdot) = - (\hat{M}'\hat{U}\hat{M})^{-1} \hat{M}'\hat{U} \hat{\phi}_m(\cdot). \]

**Corollary 1.** Under Assumption 6, for \( \hat{\Sigma} \) defined as in (11) and \( \hat{\Lambda} \) and \( \hat{\Lambda} \) defined as in (7), \( \hat{\Sigma} \xrightarrow{p} \Sigma, \hat{\Lambda} \xrightarrow{p} \Delta, \) and \( \hat{\Lambda} \xrightarrow{p} \Lambda \) under any local perturbation \( F_{\phi}(\mu/\sqrt{n}) \).

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 \( \Sigma \) and hence of \( \Delta \) and \( \Lambda \). In the case where \( \hat{\theta} \) and \( \hat{\gamma} \) are GMM or ML estimators, we can write
\[
\hat{g}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \phi_g(D_i; \theta) \quad \text{and} \quad \hat{m}(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \phi_m(D_i; \gamma)
\]
so Assumption 6 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.\(^9\) 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 6 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 \( \theta_0 \) and \( \gamma_0 \), and the gradient \( \hat{C} \) is used in delta-method calculations for asymptotic inference on \( c_0 \).\(^10\)

In this sense, in many applications estimation of \( \Sigma \) will involve only manipulation of vectors and matrices already computed as part of estimation of, and inference on, the parameters \( \theta_0, \gamma_0, \) and \( c_0 \).

---

\(^9\)Note 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.

\(^10\)Note 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 \( \Delta \) and \( \Lambda \).
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 \( \hat{\Lambda} \) 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 by reallocating program funds in a budget neutral fashion.

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.

Motivated by this discussion, we define three vectors \( \gamma \) 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 \( \gamma \) for the estimate \( \hat{c} \) following the recipe in Section 5.1. Because model estimation is via maximum likelihood and \( \gamma \) 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.11

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

---

11The 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 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
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 0.64. This is accounted for almost entirely by the panel coefficient, which alone has estimated informativeness of 0.62. 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 read to mean 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-
In Hendren’s (2013a) data, many respondents give “focal” responses of 0, 0.5, or 1 to survey elicitations 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 0.99. The estimated informativeness is 0.98 with respect to the fractions in focal point groups, 0.83 with respect to the fractions in non-focal-point groups, and 0.01 with respect to the fraction in each group needing LTC.\footnote{The estimated informativeness of the three component vectors sums to more than one. This can happen because the different vectors $\hat{\gamma}$ are correlated with one another.}

These results show that the focal point parameter is indeed tightly tied to the fraction of focal point responses. This is not surprising given the structure of the model. However, it confirms our expectation that in a case where an estimate has a simple and direct relationship to one or more descriptive statistics, estimated informativeness will be very high.

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.

\section{Conclusions}

We propose a measure $\Delta$ 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 $\Delta$, and only $\Delta$, 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 $\Delta$. 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


3073–3114.
A  Proofs

A.1  Proof of Proposition 1

We first consider the case with \( \bar{\gamma} \) unrestricted. 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

\[
\begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix} = \sum^{-1} \begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix} = \begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix} (B \sum B')^{-1} B \begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix}.
\]

Note, however, that

\[
B\begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix} = \begin{bmatrix} \bar{c} \\ \bar{\gamma} - \Lambda^* \bar{c} \end{bmatrix}
\]

while

\[
B \sum B' = \begin{bmatrix} \sigma_c^2 & 0 \\ 0 & \sum_{\gamma \gamma} - \Lambda^* \sigma_c^2 \Lambda^* \end{bmatrix}.
\]

Hence,

\[
\begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix} = \frac{\bar{c}^2}{\sigma_c^2} + (\bar{\gamma} - \Lambda^* \bar{c})' \left( \sum_{\gamma \gamma} - \Lambda^* \sigma_c^2 \Lambda^* \right)^{-1} (\bar{\gamma} - \Lambda^* \bar{c}).
\]

Since one can choose \( \bar{\gamma} \) to set the second term to zero for all values \( \bar{c} \), the range of values of \( \bar{c}_\phi \) consistent with the restriction in Condition 2 is \([-\sigma_c, \sigma_c]\). The form of \( \mathcal{R}^\mu \) then follows immediately from Condition 1.

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

\[
\begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix} = \sum^{-1} \begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix} = \begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix} (A \sum A')^{-1} A \begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix}.
\]

Note, however, that

\[
A\begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix} = \begin{bmatrix} \bar{c} - \Lambda \bar{\gamma} \\ \bar{\gamma} \end{bmatrix}
\]

while

\[
A \sum A' = \begin{bmatrix} \sigma_c^2 - \Lambda \sum_{\gamma \gamma} \Lambda' & 0 \\ 0 & \sum_{\gamma \gamma} \end{bmatrix} = \begin{bmatrix} \sigma_c^2 (1 - \Delta) & 0 \\ 0 & \sum_{\gamma \gamma} \end{bmatrix}.
\]

Hence,

\[
\begin{bmatrix} \bar{c} \\ \bar{\gamma} \end{bmatrix} = \frac{(\bar{c} - \Lambda \bar{\gamma})^2}{\sigma_c^2 (1 - \Delta)} + \bar{\gamma} \sum_{\gamma \gamma}^{-1} \bar{\gamma}.
\]
Thus, if we impose $\bar{\gamma} = 0$, then under Condition 2 the range of values of $\bar{c}_\varphi$ is $[-\sigma_c \sqrt{1 - \Delta}, \sigma_c \sqrt{1 - \Delta}]$, and the form of $\mathcal{B}_0^\mu$ follows immediately from Condition 1.

### A.2 Proof of Lemma 1

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

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

as $\vartheta \to 0$. 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

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

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

$$
\left( \log \left( \frac{dF_{\varphi}^n \left( \frac{\mu}{\sqrt{n}} \right)}{dF_0^n} \right), \frac{1}{\sqrt{n}} \sum \phi_c (D_i), \frac{1}{\sqrt{n}} \sum \phi_\gamma (D_i) \right) \to_d \left( \eta, \xi_c, \xi_\gamma \right)
$$

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 $-\frac{1}{2}$ 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 $F_{\varphi}^n \left( \frac{\mu}{\sqrt{n}} \right)$,

$$
\left( \log \left( \frac{dF_{\varphi}^n \left( \frac{\mu}{\sqrt{n}} \right)}{dF_0^n} \right), \frac{1}{\sqrt{n}} \sum \phi_c (D_i), \frac{1}{\sqrt{n}} \sum \phi_\gamma (D_i) \right) \to_d \left( \eta^*, \xi_{c^*}, \xi_{\gamma^*} \right)
$$
\[ N \left( \begin{pmatrix} \frac{\mu^2}{\Sigma} E_{F_0} \left[ s_{\varphi}(D_i) \right]^2 \\ \mu E_{F_0} \left[ \phi_c(D_i) s_{\varphi}(D_i) \right] \\ \mu E_{F_0} \left[ \phi_{\gamma}(D_i) s_{\varphi}(D_i) \right] \end{pmatrix}, \begin{pmatrix} \mu^2 E_{F_0} \left[ s_{\varphi}(D_i) \right]^2 & \mu E_{F_0} \left[ s_{\varphi}(D_i) \phi_c(D_i) \right] & \mu E_{F_0} \left[ s_{\varphi}(D_i) \phi_{\gamma}(D_i) \right] \\ \mu E_{F_0} \left[ \phi_c(D_i) s_{\varphi}(D_i) \right] & E_{F_0} \left[ \phi_c(D_i) \right]^2 & E_{F_0} \left[ \phi_c(D_i) \phi_{\gamma}(D_i) \right] \\ \mu E_{F_0} \left[ \phi_{\gamma}(D_i) s_{\varphi}(D_i) \right] & E_{F_0} \left[ \phi_{\gamma}(D_i) \phi_c(D_i) \right] & E_{F_0} \left[ \phi_{\gamma}(D_i) \phi_{\gamma}(D_i) \right] \end{pmatrix} \right) \]

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

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

and thus that

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

for

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

and

\[
\begin{pmatrix} \sigma_{c}^2 & \Sigma_{c\gamma} \\ \Sigma_{c\gamma} & \Sigma_{\gamma\gamma} \end{pmatrix} = \begin{pmatrix} E_{F_0} \left[ \phi_c(D_i) \right]^2 & E_{F_0} \left[ \phi_c(D_i) \phi_{\gamma}(D_i) \right] \\ E_{F_0} \left[ \phi_{\gamma}(D_i) \phi_c(D_i) \right] & E_{F_0} \left[ \phi_{\gamma}(D_i) \phi_{\gamma}(D_i) \right] \end{pmatrix}.
\]

This completes the proof.

### A.3 Proof of Lemma 2

We first show that for any \((\bar{c}, \bar{\gamma})\) consistent with (6), we can find \( \varphi \in \Phi \) such that \((\bar{c}_{\varphi}, \bar{\gamma}_{\varphi}) = (\bar{c}, \bar{\gamma})\).

First consider any

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

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 \( E_{F_0} [s(D_i; \overline{c}, \overline{\gamma})] = 0 \) and
\[
E_{F_0} \left[ s(D_i; \overline{c}, \overline{\gamma})^2 \right] = \left( \frac{\overline{c}}{\gamma} \right)' \Sigma^{-1} \left( \frac{\overline{c}}{\gamma} \right) = 1.
\]
Hence, by Assumption 2 there exists \( \varphi \in \Phi \) with \( s_{\varphi}(D_i) = s(D_i; \overline{c}, \overline{\gamma}) \) \( F_0 \)-almost everywhere. By construction
\[
\begin{pmatrix}
\overline{c}_{\varphi} \\
\overline{\gamma}_{\varphi}
\end{pmatrix} = \begin{pmatrix}
E_{F_0} \left[ \phi_c(D_i) s_{\varphi}(D_i) \right] \\
E_{F_0} \left[ \phi_{\gamma}(D_i) s_{\varphi}(D_i) \right]
\end{pmatrix} = \begin{pmatrix}
\overline{c} \\
\overline{\gamma}
\end{pmatrix},
\]
so we can attain any \((\overline{c}, \overline{\gamma})\) satisfying (13).

Next, consider any
\[
(\overline{c}, \overline{\gamma}) \in \left\{ (c, \gamma) \in \mathbb{R} \times \mathbb{R}^p : \sqrt{\left( \frac{c}{\gamma} \right)' \Sigma^{-1} \left( \frac{c}{\gamma} \right)} < 1 \right\},
\]
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 \( E_{F_0}[\varepsilon^*(D_i) s(D_i; \overline{c}, \overline{\gamma})] = 0 \). Let us consider \( \varphi \in \Phi \) corresponding to
\[
s_{\varphi}(D_i) = s(D_i; \overline{c}, \overline{\gamma}) + \sqrt{1 - \left( \frac{\overline{c}}{\overline{\gamma}} \right)' \Sigma^{-1} \left( \frac{\overline{c}}{\overline{\gamma}} \right)} \varepsilon^*(D_i),
\]
noting that since the right hand side has mean zero and variance one, such a \( \varphi \) again exists by Assumption 2. The corresponding score \( s_{\varphi}(D_i) \) again delivers the desired \((\overline{c}_{\varphi}, \overline{\gamma}_{\varphi})\).

We next show that for any \( \varphi \in \Phi \), \((\overline{c}_{\varphi}, \overline{\gamma}_{\varphi})\) satisfy (6). Consider any \( \varphi \), and note that we can write
\[
s_{\varphi}(D_i) = s(D_i; \overline{c}_{\varphi}, \overline{\gamma}_{\varphi}) + \varepsilon_{\varphi}(D_i)
\]
for \( \varepsilon_{\varphi}(D_i) = s_{\varphi}(D_i) - s(D_i; \overline{c}_{\varphi}, \overline{\gamma}_{\varphi}) \). Note that by the definition of \( s(D_i; \overline{c}_{\varphi}, \overline{\gamma}_{\varphi}) \),
\[
\begin{pmatrix}
E_{F_0} \left[ \phi_c(D_i) s_{\varphi}(D_i) \right] \\
E_{F_0} \left[ \phi_{\gamma}(D_i) s_{\varphi}(D_i) \right]
\end{pmatrix} = \begin{pmatrix}
E_{F_0} \left[ \phi_c(D_i) s \left( D_i; \overline{c}_{\varphi}, \overline{\gamma}_{\varphi} \right) \right] \\
E_{F_0} \left[ \phi_{\gamma}(D_i) s \left( D_i; \overline{c}_{\varphi}, \overline{\gamma}_{\varphi} \right) \right]
\end{pmatrix},
\]
and thus
\[
\begin{pmatrix}
E_{F_0} \left[ \phi_c(D_i) \varepsilon_{\varphi}(D_i) \right] \\
E_{F_0} \left[ \phi_{\gamma}(D_i) \varepsilon_{\varphi}(D_i) \right]
\end{pmatrix} = 0.
\]
Hence, \( \varepsilon_{\varphi}(D_i) \) is orthogonal to \( \phi_c(D_i) \) and \( \phi_{\gamma}(D_i) \), and thus to \( s(D_i; \overline{c}_{\varphi}, \overline{\gamma}_{\varphi}) \). In particular, we
can interpret \( s(D_i; \overline{c}_\phi, \overline{\gamma}_\phi) \) as the projection of \( s_\phi(D_i) \) onto \( (\phi_c(D_i), \phi_\gamma(D_i)) \), and \( \varepsilon_\phi(D_i) \) as the residual from this projection. Using this orthogonality, note that
\[
\sqrt{E_{F_0} \left[ s_\phi(D_i)^2 \right]} = \sqrt{E_{F_0} \left[ s(D_i; \overline{c}_\phi, \overline{\gamma}_\phi)^2 \right] + E_{F_0} \left[ \varepsilon_\phi(D_i)^2 \right]}
\]
\[
\geq \sqrt{E_{F_0} \left[ s(D_i; \overline{c}_\phi, \overline{\gamma}_\phi)^2 \right]} = \sqrt{\left( \frac{\overline{c}_\phi}{\overline{\gamma}_\phi} \right) \Sigma^{-1} \left( \frac{\overline{c}_\phi}{\overline{\gamma}_\phi} \right) \Sigma \left( \frac{\overline{c}_\phi}{\overline{\gamma}_\phi} \right)}.
\]
Thus, a value \((\overline{c}, \overline{\gamma})\) is consistent with \( \sqrt{E_{F_0} \left[ s_\phi(D_i)^2 \right]} \leq 1 \) if and only if it satisfies (6).

### A.4 Proof of Proposition 2

The result follows as a special case of Proposition 6 where we set \( \overline{\gamma} = 0 \).

### A.5 Proof of Proposition 3

By the Neyman-Pearson Lemma (see Theorem 3.2.1 in Lehmann and Romano 2005), the most powerful level-\( \alpha \) test of \( H_0 : (D_1, \ldots, D_n) \sim F_{\mu_n} \) against \( H_1 : (D_1, \ldots, D_n) \sim F_{\phi_n} \left( \frac{\mu}{\sqrt{n}} \right) \) rejects when the log likelihood ratio \( \log \left( \frac{dF_n}{dF_0} \left( \frac{\mu}{\sqrt{n}} \right) \right) / dF_0^n \) exceeds a critical value \( c_{\alpha,n} \) chosen to ensure rejection probability \( \alpha \) 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,
\[
\log \left( \frac{dF_n}{dF_0} \left( \frac{\mu}{\sqrt{n}} \right) \right) \rightarrow_d \begin{cases} 
N \left( -\frac{\mu^2}{2}, \mu^2 \right) & \text{under } F_0^n \\
N \left( \frac{\mu^2}{2}, \mu^2 \right) & \text{under } F_{\phi_n} \left( \frac{\mu}{\sqrt{n}} \right)
\end{cases}
\]
Thus, 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,
\[
\log \left( \frac{dF_n}{dF_0} \left( \frac{\mu}{\sqrt{n}} \right) \right) - c_{\alpha,n} \rightarrow_d \begin{cases} 
N \left( -c_\alpha, 1 \right) & \text{under } F_0^n \\
N \left( \mu - c_\alpha, 1 \right) & \text{under } F_{\phi_n} \left( \frac{\mu}{\sqrt{n}} \right)
\end{cases}
\]
for \( c_\alpha \) the \( 1 - \alpha \) quantile of a standard normal distribution, from which the result follows.
**A.6 Proof of Proposition 4**

Assumption 4 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

\[
n \cdot r \left( F_0, F_\varphi \left( \frac{\mu}{\sqrt{n}} \right) \right) =
\]

\[
n \cdot E_{F_0} \left[ \psi \left( \frac{\partial F_\varphi \left( \frac{\mu}{\sqrt{n}} \right)}{\partial \varphi} \right) \right] =
\]

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

Assumption 4 and Leibniz’s rule implies that

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

Hence, we see that

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

and thus that

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

\[
\frac{\mu^2}{2} E_{F_0} \left[ \psi' \left( \frac{f_\varphi (D_i; \tilde{\vartheta}_n)}{f_\varphi (D_i; 0)} \right) \frac{\partial^2 f_\varphi (D_i; \tilde{\vartheta}_n)}{\partial \varphi^2} + \psi'' \left( \frac{f_\varphi (D_i; \tilde{\vartheta}_n)}{f_\varphi (D_i; 0)} \right) \left( \frac{\partial f_\varphi (D_i; \tilde{\vartheta}_n)}{f_\varphi (D_i; 0)} \right)^2 \right] .
\]
The Dominated Convergence Theorem then implies that (since $\psi''(1) = 2$ and the expression in brackets is uniformly bounded on $\mathcal{N}$)

$$
\frac{\mu^2}{2} E_{F_0} \left[ \psi' \left( \frac{f_{\bar{D}_i} (D_i; \bar{c}_n)}{f_{\bar{D}_i} (D_i; 0)} \right) \frac{\partial^2 f_{\bar{D}_i} (D_i; \bar{c}_n)}{\partial \bar{c}_n^2} + \psi'' \left( \frac{f_{\bar{D}_i} (D_i; \bar{c}_n)}{f_{\bar{D}_i} (D_i; 0)} \right) \left( \frac{\partial f_{\bar{D}_i} (D_i; \bar{c}_n)}{f_{\bar{D}_i} (D_i; 0)} \right)^2 \right] \rightarrow
$$

$$
\frac{\mu^2}{2} E_{F_0} \left[ \psi' \left( \frac{f_{\bar{D}_i} (D_i; 0)}{f_{\bar{D}_i} (D_i; 0)} \right) \frac{\partial^2 f_{\bar{D}_i} (D_i; 0)}{\partial \bar{c}_n^2} + \psi'' \left( \frac{f_{\bar{D}_i} (D_i; 0)}{f_{\bar{D}_i} (D_i; 0)} \right) s_{\bar{D}_i} (D_i) \right]
$$

$$
= \mu^2 E_{F_0} \left[ -\frac{1}{\lambda + 1} \frac{\partial^2 f_{\bar{D}_i} (D_i; 0)}{f_{\bar{D}_i} (D_i; 0)} + s_{\bar{D}_i} (D_i) \right].
$$

However,

$$
E_{F_0} \left[ \frac{\partial^2 f_{\bar{D}_i} (D_i; 0)}{f_{\bar{D}_i} (D_i; 0)} \right] = \int \frac{\partial^2}{\partial \bar{c}_n^2} f_{\bar{D}_i} (d; 0) \, d\nu (d) = \frac{\partial^2}{\partial \bar{c}_n^2} \int f_{\bar{D}_i} (d; 0) \, d\nu (d) = 0,
$$

so

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

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

### A.7 Proof of Proposition 6

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

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

$$
\epsilon_{\phi} (D_i) = s_{\bar{D}_i} (D_i) - s (D_i; \bar{\gamma}_\phi).
$$

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

$$
\bar{c}_{\phi} = E_{F_0} \left[ \phi_{\epsilon} (D_i) \epsilon_{\phi} (D_i) \right] = E_{F_0} \left[ \phi_{\epsilon} (D_i) \phi_{\bar{\gamma}} (D_i) \right] \Sigma_{\bar{\gamma}}^{-1} \bar{\gamma}_\phi + E_{F_0} \left[ \phi_{\epsilon} (D_i) \epsilon_{\phi} (D_i) \right]
$$

$$
= \Lambda \bar{\gamma}_\phi + E_{F_0} \left[ \phi_{\epsilon} (D_i) \epsilon_{\phi} (D_i) \right].
$$

37
Next, define
\[ \bar{\phi}_c(D_i) = \phi_c(D_i) - \Lambda\phi_y(D_i) \]
and note that
\[ E_{F_0} [\phi_c(D_i)\varepsilon\phi(D_i)] = E_{F_0} [\bar{\phi}_c(D_i)\varepsilon\phi(D_i)]. \]

The Cauchy-Schwarz inequality then implies that
\[
|E_{F_0} [\bar{\phi}_c(D_i)\varepsilon\phi(D_i)]| \leq \sqrt{E_{F_0} [\bar{\phi}_c(D_i)^2]} \sqrt{E_{F_0} [\varepsilon\phi(D_i)^2]}
\]
\[
= \sqrt{\sigma_c^2 - \Delta\Sigma\gamma\gamma} \sqrt{E_{F_0} [\varepsilon\phi(D_i)^2]}
\]
\[
= \sigma_c \sqrt{1 - \Delta} \sqrt{1 - \bar{\gamma}\Sigma\gamma^{-1}\bar{\gamma}}.
\]

Combining these results we see that for \((\mu, \varphi)\) such that \(\mu\bar{\gamma}_\varphi = \bar{\gamma}\)
\[
\mu\bar{\epsilon}_\varphi \in \left[\Lambda\bar{\gamma} - \sigma_c \sqrt{1 - \Delta} \sqrt{\mu^2 - \bar{\gamma}\Sigma\gamma^{-1}\bar{\gamma}}, \Lambda\bar{\gamma} + \sigma_c \sqrt{1 - \Delta} \sqrt{\mu^2 - \bar{\gamma}\Sigma\gamma^{-1}\bar{\gamma}}\right],
\]
which are the bounds stated in the proposition. In particular, if \(\mu^2 - \bar{\gamma}\Sigma\gamma^{-1}\bar{\gamma} < 0\) then there exists no \(\varphi\) with \(\mu\bar{\gamma}_\varphi = \bar{\gamma}\), so the set of possible values for \(\mu\bar{\epsilon}_\varphi\) 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
\[
(14) \quad \bar{c} \in \left[\Lambda\bar{\gamma} - \sigma_c \sqrt{1 - \Delta} \sqrt{\mu^2 - \bar{\gamma}\Sigma\gamma^{-1}\bar{\gamma}}, \Lambda\bar{\gamma} + \sigma_c \sqrt{1 - \Delta} \sqrt{\mu^2 - \bar{\gamma}\Sigma\gamma^{-1}\bar{\gamma}}\right]
\]
there exists \(\varphi \in \Phi\) with \(\mu\bar{\epsilon}_\varphi = \bar{c}\) 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).

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

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

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

However, by (14) we know that

\[
|\hat{c} - \Lambda \hat{\gamma}| \leq \sigma_c \sqrt{1 - \Delta} \sqrt{1 - \hat{\gamma} \Sigma_{\hat{\gamma}}^{-1} \hat{\gamma}}
\]

and thus that

\[
\frac{(\hat{c} - \Lambda \hat{\gamma})^2}{\sigma_c^2 (1 - \Delta)} \leq \left(1 - \hat{\gamma} \Sigma_{\hat{\gamma}}^{-1} \hat{\gamma}\right)
\]

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

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

and this \(\varphi\) yields \(E_{F_0} \left[\phi_c (D_i) s_{\varphi} (D_i)\right] = \hat{c}, E_{F_0} \left[\phi_{\gamma} (D_i) s_{\varphi} (D_i)\right] = \hat{\gamma}\), as desired. That \(\mathcal{B}^\mu_{\hat{\gamma}}\) gives the range of asymptotic biases for \(\hat{c}\) under perturbations with \(\mu \hat{\gamma}_{\varphi} = \hat{\gamma}\) then follows from Lemma 1.

To derive the corresponding result for \(\mathcal{B}^\mu (\hat{\gamma})\), note first that by the Cauchy-Schwarz inequality

\[
|E_{F_0} \left[\phi_c (D_i) s_{\varphi} (D_i)\right]| \leq \sigma_c,
\]

which immediately implies that \(\mu \hat{c}_{\varphi} \in \mathcal{B}^\mu\) for all \(\mu\) and all \(\varphi \in \Phi\). To show that these bounds are tight, define \(\hat{s} (D_i; \hat{c}) = \phi_c (D_i) \frac{\varepsilon}{\sigma_c}\), and note that \(E_{F_0} \left[\phi_c (D_i) \hat{s} (D_i; \hat{c})\right] = \hat{c}\) by construction. For any \(\varepsilon \in [-\sigma_c, \sigma_c]\), Assumption 2 implies that there exists \(\varphi \in \Phi\) such that \(s_{\varphi} (D_i)\) is almost surely equal to

\[
\hat{s} (D_i; \hat{c}) + \sqrt{1 - E_{F_0} \left[\hat{s} (D_i; \hat{c})^2\right] \varepsilon^* (D_i)},
\]

for \(\varepsilon^* (D_i)\) as defined above. Note, however, that \(\hat{c}_{\varphi} = E_{F_0} \left[\phi_c (D_i) s_{\varphi} (D_i)\right] = \hat{c}\) by construction. The result then follows immediately after scaling by \(\mu\).

### A.8 Proof of Proposition 7

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 \(\hat{\Lambda}\) 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 \(\Sigma_{\hat{\gamma} \hat{\gamma}}\) has full rank.
A.9 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

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

The result then follows immediately from Assumption 6.

A.10 Proof of Lemma 4

Note, first, that $\hat{\Lambda}_{cg} \xrightarrow{p} \Lambda_{cg}$ and $\hat{\Lambda}_{gm} \xrightarrow{p} \Lambda_{gm}$ by Assumption 6 parts (A)-(C) along with the Continuous Mapping Theorem (see e.g. Theorem 2.3 of van der Vaart 1998). Note, next, that 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 6 part (E) implies that $\hat{\Sigma}^* \rightarrow_p \Sigma^*$.

Note, next, that by Lemma 3,

$$\Sigma = \begin{bmatrix} E_{F_0} \left[ \phi_c(D_i)^2 \right] & E_{F_0} \left[ \phi_c(D_i) \phi_y(D_i) \right] \\ E_{F_0} \left[ \phi_y(D_i) \phi_c(D_i) \right] & E_{F_0} \left[ \phi_y(D_i) \phi_y(D_i) \right] \end{bmatrix} = \begin{bmatrix} \Lambda_{cg} & 0 \\ 0 & \Lambda_{gm} \end{bmatrix} \Sigma^* \begin{bmatrix} \Lambda_{cg} & 0 \\ 0 & \Lambda_{gm} \end{bmatrix}'.$$

Note that we can write $\hat{\Sigma}$ as

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

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

A.11 Proof of Corollary 1

Follows immediately from Lemma 4 and Proposition 7.
Table 1: Estimated informativeness of descriptive statistics for the effect of a counterfactual rebudgeting of PROGRESA (Attanasio et al. 2012a)

<table>
<thead>
<tr>
<th>Descriptive statistics</th>
<th>Estimated informativeness $\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.283</td>
</tr>
<tr>
<td>Impact on eligibles</td>
<td>0.227</td>
</tr>
<tr>
<td>Impact on ineligibles</td>
<td>0.056</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Descriptive statistics</th>
<th>Estimated informativeness $\hat{\Delta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.635</td>
</tr>
<tr>
<td>IV coefficient</td>
<td>0.011</td>
</tr>
<tr>
<td>Panel coefficient</td>
<td>0.621</td>
</tr>
</tbody>
</table>

Notes: The table shows the estimated informativeness $\hat{\Delta}$ of three vectors $\hat{\gamma}$ of descriptive statistics for the estimated effect $\hat{\gamma}$ 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 (2013)

<table>
<thead>
<tr>
<th>Descriptive statistics $\hat{\gamma}$</th>
<th>Estimated informativeness $\hat{\Delta}$ for</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fraction focal point respondents</td>
</tr>
<tr>
<td>All</td>
<td>0.987</td>
</tr>
<tr>
<td>Fractions in focal point groups</td>
<td>0.979</td>
</tr>
<tr>
<td>Fractions in non-focal point groups</td>
<td>0.825</td>
</tr>
<tr>
<td>Fraction in each group needing LTC</td>
<td>0.014</td>
</tr>
</tbody>
</table>

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 \((\bar{c}_\phi, \bar{\gamma}_\phi)\) under local perturbations

Notes: Figure shows the ellipsoid defined in (3) for the case with \(p_\gamma = 1\). The interval labeled “\(\bar{\gamma}_\phi\) unconstrained” characterizes the set of all possible values of \(\bar{c}_\phi\). The interval labeled “\(\bar{\gamma}_\phi = 0\)” characterizes the set of possible values of \(\bar{c}_\phi\) when \(\bar{\gamma}_\phi = 0\). The diagonal line through the ellipsoid connects the points on the ellipsoid that achieve the minimum and maximum values of \(\bar{c}_\phi\).
Notes: Figure shows the limits of the intervals $\mathcal{B}^\mu$ (labeled “$\mathcal{Y}_\phi$ unconstrained”) and $\mathcal{B}_0^\mu$ (labeled “$\mathcal{Y}_\phi = 0$”) defined in Proposition 1 for the case of $\Delta = 0.75$. The intervals are normalized by the standard error $\sigma_c$, so that a value of 1 on the y-axis indicates a first-order asymptotic bias of one standard error $\sigma_c$. 

Figure 2: Asymptotic bias $\bar{c}_\phi$ of estimator under local perturbations