Measuring the Sensitivity of Parameter Estimates to Estimation Moments

Isaiah Andrews
MIT and NBER

Matthew Gentzkow
Stanford and NBER

Jesse M. Shapiro
Brown and NBER[*]

March 2017

Abstract

We propose a local measure of the relationship between parameter estimates and the moments of the data they depend on. Our measure can be computed at negligible cost even for complex structural models. We argue that reporting this measure can increase the transparency of structural estimates, making it easier for readers to predict the way violations of identifying assumptions would affect the results. When the key assumptions are orthogonality between error terms and excluded instruments, we show that our measure provides a natural extension of the omitted variables bias formula for nonlinear models. We illustrate with applications to published articles in several fields of economics.

[*]Conversations with Kevin M. Murphy inspired and greatly improved this work. We are grateful also to Josh Angrist, Steve Berry, Alan Bester, Stephane Bonhomme, Dennis Carlton, Raj Chetty, Tim Conley, Ron Goettler, Brett Gordon, Phil Haile, Christian Hansen, Frank Kleibergen, Pat Kline, Mark Li, Asad Lodhia, Magne Mogstad, Adam McCloskey, Yaroslav Mukhin, Pepe Olea, Matt Taddy, E. Glen Weyl, and seminar audiences at Berkeley, Brown, Columbia, University of Chicago, Harvard, University of Michigan, MIT, NBER, Northwestern, NYU, Princeton, Stanford, University of Toronto, and Yale for advice and suggestions, and to our dedicated research assistants for important contributions to this project. We thank the following authors for their assistance in working with their code and data: Mariacristina De Nardi, Eric French, and John B. Jones; Stefano DellaVigna, John List, and Ulrike Malmendier; Ron Goettler and Brett Gordon; Pierre-Olivier Gourinchas and Jonathan Parker; Nathaniel Hendren; Chris Knittel and Konstantinos Metaxoglou; Michael Mazzeo; Boris Nikolov and Toni Whited; Greg Kaplan; and Amil Petrin. This research was funded in part by the Initiative on Global Markets, the George J. Stigler Center for the Study of the Economy and the State, the Ewing Marion Kauffman Foundation, the Centel Foundation / Robert P. Reuss Faculty Research Fund, the Neubauer Family Foundation, and the Kathryn C. Gould Research Fund, all at the University of Chicago Booth School of Business, the Alfred P. Sloan Foundation, the Silverman (1968) Family Career Development Chair at MIT, the Stanford Institute for Economic Policy Research, the Brown University Population Studies and Training Center, and the National Science Foundation. E-mail: iandrews@mit.edu, gentzkow@stanford.edu, jesse_shapiro_1@brown.edu.

1 Introduction

One of the drawbacks commonly attributed to structural empirical methods is a lack of transparency. Heckman (2010) writes that "the often complex computational methods that are required to implement [structural estimation] make it less transparent" (358). Angrist and Pischke (2010) note that it is often "hard to see precisely which features of the data drive the ultimate results" (21).

In this paper, we suggest a way to improve the transparency of common structural estimators. We consider a researcher who computes an estimator $\hat{\theta}$ of a finite-dimensional parameter θ with true value θ_0. Under the researcher's maintained assumptions a_0, $\hat{\theta}$ is consistent and asymptotically normal. Not all readers of the research accept a_0, however, and different readers entertain different alternatives. To assess the potential bias in $\hat{\theta}$ under some alternative $a \neq a_0$, a reader needs to know two things: how a would change the moments of the data that the estimator uses as inputs, and how changes in these moments affect the estimates. We say that research is *transparent* to the extent that it makes these steps easy, allowing a reader to assess the potential bias for a range of alternatives $a \neq a_0$ she finds relevant.

Linear regression analysis is popular in part because it is transparent. Estimates depend on a set of intuitive variances and covariances, and it is straightforward to assess how these moments would change under violations of the identifying assumptions. Well-understood properties of linear models—most prominently, the omitted variables bias formula—make it easy for readers to guess how these changes translate into bias in the estimates. We do not need to have access to the data to know that a regression of wages on education would be biased upward by omitted skill, and we can form a guess about how much if we have a prior on the likely covariance properties of the omitted variable.

Our analysis is designed to make this kind of transparency easier to deliver for nonlinear models. We derive a measure of the sensitivity of an estimator to perturbations of different moments of the data, exploiting the same local linearization used to derive standard asymptotics. If a reader can predict the effect of an alternative a on the moments, our measure allows her to translate this into predicted bias in the estimates. We show that the measure can be used to predict the effect of omitted variables in a large class of nonlinear models—providing an analogue of the omitted variables bias formula for these settings—and also to predict the effect of many other potential violations of identifying assumptions. Because our approximation is local, the predictions will be valid for alternatives a that are close to a_0 in an appropriate sense.

We assume that $\hat{\theta}$ minimizes a criterion function $\hat{g}(\theta)'\hat{W}\hat{g}(\theta)$, where $\hat{g}(\theta)$ is a vector of moments or other statistics, $\hat{W}$ is a weight matrix, and both are functions of the realized data. This class of minimum distance estimators (MDEs) includes generalized method of moments (GMM), classical minimum distance (CMD), maximum likelihood (MLE), and their simulation-based ana-

logues (Newey and McFadden 1994), and so encompasses most of the workhorse methods of structural point estimation.

For any a in a set $\mathscr{A}$ of alternative assumptions, we follow the literature on local misspecification (e.g., Newey 1985; Conley et al. 2012) and define a local perturbation of the model in the direction of a such that the degree of misspecification shrinks with the size of the sample. For any such perturbation, we assume that $\sqrt{n}\hat{g}(\theta_0)$ converges in distribution to a random variable $\tilde{g}(a)$. We show that $\sqrt{n}\left(\hat{\theta}-\theta_0\right)$ then converges in distribution to a random variable $\tilde{\theta}(a)$ and $\hat{\theta}$ has first-order asymptotic bias:

$$\mathbb{E}\left(\tilde{\theta}(a)\right) \quad = \quad \Lambda\mathbb{E}\left(\tilde{g}(a)\right),$$

for a matrix Λ. An analogous relationship holds when the outcome of interest is a function of $\hat{\theta}$, such as a counterfactual experiment or welfare calculation.

The matrix Λ, which we call *sensitivity,* plays a central role in our analysis. It can be written as $\Lambda = -\left(G'WG\right)^{-1}G'W$, where G is the Jacobian of the probability limit of $\hat{g}(\theta)$ at θ_0 and W is the probability limit of $\hat{W}$. Since standard approaches to inference on θ employ plug-in estimates of G and W, sensitivity can be consistently estimated at essentially zero computational cost in most applications.

Intuitively, Λ is a local approximation to the mapping from moments to estimated parameters. A reader interested in an alternative a can use Λ to predict its effect on the results, provided she can form a guess as to the induced bias in the moments $\mathbb{E}(\tilde{g}(a))$. We argue theoretically, and illustrate in our applications, that predicting the way a affects the moments is straightforward in many cases of interest.

One leading special case is where $\hat{g}(\theta)$ is additively separable into a term $\hat{s}$ dependent on the data (but not the parameters) and a term $s(\theta)$ dependent on the parameters (but not the data). This class includes CMD, additively separable GMM or simulated method of moments, and indirect inference. Here, the key identifying assumptions a_0 imply that $\hat{s}$ converges in probability to the model analogues $s(\theta_0)$. Natural alternatives a involve misspecification of $s(\theta_0)$ and mismeasurement of $\hat{s}$. It is often straightforward to say how a given alternative a would impact the asymptotic behavior of the moments $\hat{g}(\theta_0) = \hat{s} - s(\theta_0)$. If the researcher reports Λ in her paper, a reader can use Λ to predict the effect of such alternatives on the estimator.

A second special case is where $\hat{g}(\theta)$ is the product of a vector of instruments Z and a vector of structural residuals $\hat{\zeta}(\theta)$, so $\hat{\theta}$ is a nonlinear instrumental variables (IV) estimator. Here, the key identifying assumptions a_0 specify orthogonality between Z and $\hat{\zeta}(\theta)$. We show that in this case Λ can be used to construct a nonlinear-model analogue to the omitted variables bias formula that can be reported directly in a research paper. This allows readers to predict the effect of any a

from a class of perturbations that introduce omitted variables correlated with the instruments. Just as with the standard omitted variables bias formula, the key input the reader must provide is the hypothesized coefficients from a regression of the omitted variable on the instruments. Our results for this case generalize the findings of Conley et al. (2012) on the effect of local misspecification in a linear IV setup.

We illustrate the utility of our approach with three applications. The first is to DellaVigna et al.'s (2012) model of charitable giving. The authors use a field experiment in conjunction with a structural model to distinguish between altruistic motives and social pressure as drivers of giving. They find that social pressure is an important driver and that the average household visited by their door-to-door solicitors is made worse off by the solicitation. We compute the sensitivity of the estimated social pressure to the moments used in estimation, and find that a key driver is the extent to which donations bunch at exactly \$10. This is consistent with the model's baseline assumptions, under which (i) households pay a social pressure cost if they give less than \$10, but pay no cost if they give \$10 or more, and (ii) there are no reasons to bunch at \$10 absent social pressure. We then show how a reader can use our sensitivity measure to assess the bias if the second assumption is relaxed—e.g., if some fraction of households give \$10 because it is a convenient cash denomination. We find that the estimated social pressure is biased upward in this case.

Our second application is to Gourinchas and Parker's (2002) model of lifecycle consumption. The model allows both consumption-smoothing ("lifecycle") and precautionary motives for savings. The authors find that precautionary incentives dominate at young ages, while lifecycle motives dominate later in life, providing a rationale for the observed combination of a hump-shaped consumption profile and high marginal propensity to consume out of income shocks at young ages. We show that our sensitivity measure provides intuition about the consumption profiles the model interprets as evidence of smoothing and precautionary motives respectively. We then show how a reader could use our measure to assess sensitivity to violations of two key assumptions: separability of consumption and leisure in utility, and the absence of unobserved income sources. We show that realistic violations of separability could meaningfully affect the results. For example, varying shopping intensity as in Aguiar and Hurst (2007) would mean that the estimates understate the importance of precautionary motives relative to lifecycle savings. We also show that the presence of within-family transfers, a potential source of unobserved income, would have a similar effect.

Our final application is to Berry et al.'s (1995, henceforth "BLP") model of automobile demand and pricing. The model yields estimates of the markups firms charge on specific car models. These markups are a measure of market power and an input into evaluation of policies such as trade restrictions (BLP 1999), mergers (Nevo 2000), and the introduction of new goods (Petrin 2002). The moments $\hat{g}(\theta)$ used to estimate the model are products of vehicle characteristics—used as

instruments—with shocks to demand and marginal cost, and the key identifying assumption is that the instruments are orthogonal to the shocks. We show how a reader could use our sensitivity measure to assess a range of violations of these assumptions including economies of scope and correlation between demand errors and the composition of product lines. We find that each of these violations could lead to economically meaningful bias in the estimated markups.

We emphasize two limitations to our approach. The first is that our sensitivity measure is a local approximation. For small deviations away from the baseline assumptions a_0, we can be confident it will deliver accurate predictions. For larger deviations, it may still provide valuable intuition, subject to the usual limitations of linear approximation. When there are specific large deviations of interest, we recommend that authors evaluate them using standard sensitivity analysis. The transparency our measure offers is a complement to this, allowing readers to build additional intuition about the impact of a broad set of alternatives. In the online appendix, we compare our local sensitivity measure to a measure of global sensitivity for DellaVigna et al. (2012) and BLP (1995).

The second limitation is that the units of Λ are contingent on the units of $\hat{g}(\theta)$. Changing the measurement of an element $\hat{g}_j(\theta)$ from, say, dollars to euros, changes the corresponding elements of Λ. This does not affect the bias a reader would estimate for specific alternative assumptions, but it does matter for qualitative conclusions about the relative importance of different moments.

The remainder of the paper is organized as follows. Section 2 situates our approach relative to prior literature. Section 3 defines sensitivity and characterizes its properties. Section 4 derives results for the special cases of CMD and IV. Section 5 develops an alternative notion of sensitivity that does not rely on large-sample approximations. Section 6 considers estimation. Section 7 presents our applications, and section 8 concludes. Appendix A discusses some common alternatives, and the online appendix extends our main results along several dimensions.

2 Relationship to Prior Literature

Transparency as defined here serves a distinct purpose from either traditional (global) sensitivity analysis or estimation under partial identification. In sensitivity analysis, a researcher shows how the results change under particular prominent alternatives a. Transparency is different because it allows readers to consider a large space of alternatives, including those not anticipated by the researcher in advance. In estimation under partial identification, a researcher computes bounds on θ_0 assuming only that some set $\tilde{A}$ contains a valid collection of assumptions. This does not replace transparency because the implied bounds could be very wide if we take $\tilde{A}$ to include all possible alternatives of interest, and because bounds do not tell a given reader which element of the identified set corresponds to her own beliefs.

What our measure captures is also distinct from identification. A model is identified if, under its assumptions, alternative values of the parameters imply different distributions of observable data (Matzkin 2013). This is a binary property, and a property of a model rather than of an estimator. Our analysis takes as given that a model is identified, and describes the way a specific estimator maps data features into results. We see this as a complement to, not a substitute for, formal analysis of identification.

That said, we do think that some informal discussions of identification that have appeared in structural papers under the heading of identification may be usefully reframed in terms of sensitivity. These discussions often describe the extent to which particular parameters are "identified by" specific moments of the data. As Keane (2010) notes, these discussions are hard to understand as statements about identification in the formal sense.[1] Because identification is a binary property, claims that a moment is the "main" or "primary" source of identification have no obvious formal meaning.[2] Authors often acknowledge the imprecision of their statements by saying they discuss identification "loosely," "casually," or "heuristically."[3] Sensitivity gives a formal, quantitative language in which to describe the relative importance of different moments for determining the value of specific parameters, and we think it may be closer to the concept that many authors have in mind when discussing identification informally. Transparency as we define it provides a rationale for why such discussions are valuable.

Our work has a number of antecedents. Our approach is related to influence function calculations for determining the distribution of estimators (Huber and Ronchetti 2009), and is particularly close to the large literature on local misspecification (e.g., Newey 1985; Berkowitz et al. 2008; Guggenberger 2012; Conley et al. 2012; Nevo and Rosen 2012; Kitamura et al. 2013; Glad and Hjort 2016; Kristensen and Salanié forthcoming). Our results also relate to the literature on sensi-

[1] Keane (2010) writes: "Advocates of the 'experimentalist' approach often criticize structural estimation because, they argue, it is not clear how parameters are 'identified'. What is meant by 'identified' here is subtly different from the traditional use of the term in econometric theory — i.e., that a model satisfies technical conditions insuring a unique global maximum for the statistical objective function. Here, the phrase 'how a parameter is identified' refers instead to a more intuitive notion that can be roughly phrased as follows: What are the key features of the data, or the key sources of (assumed) exogenous variation in the data, or the key a priori theoretical or statistical assumptions imposed in the estimation, that drive the quantitative values of the parameter estimates, and strongly influence the substantive conclusions drawn from the estimation exercise?" (6).

[2] Altonji et al. (2005) write: "Both [exclusion restrictions and functional form restrictions] contribute to identification.... We explore whether the source of identification is *primarily coming from* the exclusion restrictions or *primarily coming from* the functional form restrictions" (814). Goettler and Gordon (2011) write: "The demand-side parameters... are *primarily identified by* [a set of moments].... The supply-side parameters... are *primarily identified by* [a different set of moments]" (1161). DellaVigna et al. (2012) write: "Though the parameters are estimated jointly, it is possible to address the *main sources of identification* of individual parameters" (37). (Emphasis added.)

[3] Einav et al. (2015) write: "*Loosely speaking*, identification [of three key parameters] relies on three important features of our model and data..." (869). Crawford and Yurukoglu (2012) write: "One may *casually* think of [a set of moments] as 'empirically identifying' [a set of parameters]" (662). Gentzkow et al. (2014) offer a "*heuristic*" discussion of identification which they conclude by saying: "Although [we treat] the different steps as separable, the... parameters are in fact jointly determined and jointly estimated" (3097). (Emphasis added.)

tivity analysis (including Leamer 1983; Sobol 1993; Saltelli et al. 2008; Chen et al. 2011). Our focus is on local, rather than global, deviations from the assumed model, and the sample sensitivity we derive in section 5 is a natural local sensitivity measure from the perspective of this literature.

Relative to the existing literature on local misspecification, our main contribution is the proposal to report sensitivity alongside structural estimates, as a way to increase transparency and make it easier for readers to build intuition about the forms of misspecification they find most important. In this sense, our approach is similar to Müller's (2012) measure of prior sensitivity for Bayesian models, which allows readers to adjust reported results to better reflect their own priors. A second contribution of this paper is to characterize the finite-sample derivative of the minimum distance estimator with respect to perturbations of the estimation moments, and to show that this derivative's limiting value is the sensitivity matrix.

In appendix A, we discuss two alternative approaches that have appeared in the literature. One is to ask how parameter estimates change when a moment of interest is dropped from the estimation. We show that the limiting value of this change is the product of our sensitivity measure and the degree of misspecification of the dropped moment. The other is to ask how the value of the moments simulated from the model change when we vary a particular parameter. We show that this has a limiting value proportional to a generalized inverse of our measure.

3 Measure

We have observations $D_i \in \mathscr{D}$ for $i = 1, ..., n$, which comprise a sample $D \in \mathscr{D}^n$. A set of identifying assumptions a_0 implies that D_i follows $F(\cdot|\theta, \psi)$, where θ is a P-dimensional parameter of interest with true value θ_0 and ψ is a possibly infinite-dimensional nuisance parameter with true value ψ_0. When it does not introduce ambiguity, we abbreviate the distribution $F(\cdot|\theta_0, \psi_0)$ of D_i under this model by F, and the sequence of distributions of the sample by $F_n \equiv \{\times_n F\}_n$.

The estimator $\hat{\theta}$ solves

$$\min_{\theta \in \Theta} \hat{g}(\theta)' \hat{W} \hat{g}(\theta), \tag{1}$$

where Θ is a compact subset of $\mathbb{R}^P$ known to contain θ_0 in its interior. The object $\hat{g}(\theta)$ is a J-dimensional function of parameters and data continuously differentiable in θ with Jacobian $\hat{G}(\theta)$. We assume that under F_n, and thus under the assumptions a_0, (i) $\sqrt{n}\hat{g}(\theta_0) \xrightarrow{d} N(0, \Omega)$; (ii) $\hat{W}$ converges in probability to a positive semi-definite matrix W; (iii) $\hat{g}(\theta)$ and $\hat{G}(\theta)$ converge uniformly in probability to continuous functions $g(\theta)$ and $G(\theta)$; and (iv) $G'WG = G(\theta_0)'WG(\theta_0)$ is nonsingular. We further assume that $g(\theta)'Wg(\theta)$ has a unique minimum at θ_0. Under these assumptions, $\hat{\theta}$ is consistent, asymptotically normal, and asymptotically unbiased with variance

$\Sigma = (G'WG)^{-1} G'W\Omega WG(G'WG)^{-1}$ (Newey and McFadden 1994).

Definition. *The **sensitivity** of $\hat{\theta}$ to $\hat{g}(\theta_0)$ is*

$$\Lambda = -\left(G'WG\right)^{-1} G'W.$$

Example. (OLS) Suppose the data are $D_i = (Y_i, X_i)$. The baseline assumptions a_0 imply that

$$Y_i = X_i'\theta_0 + \varepsilon_i, \tag{2}$$

with $\mathbb{E}(\varepsilon_i|X_i) = 0$. The regression coefficient of Y on X can be written as a GMM estimator with $\hat{g}(\theta) = \frac{1}{n}\sum_i X_i(Y_i - X_i'\theta)$ and $W = I$. Thus, linear regression is a special case of minimum distance estimation as in (1). Noting that $G = -\mathbb{E}(X_iX_i') = -\Omega_{XX}$, we have $\Lambda = \Omega_{XX}^{-1}$.

While the estimator $\hat{\theta}$ is derived under the assumptions a_0, we may be concerned that the data generating process is in fact described by alternative assumptions a. We follow the literature on local misspecification (e.g., Newey 1985; Conley et al. 2012) and focus on perturbations that allow the degree of misspecification to shrink with the size of the sample. Define a family of distributions indexed by $\mu \in [0,1]$,

$$F(\mu) \equiv F(\cdot|\theta_0, \psi_0, \mu),$$

such that $F(0) = F(\cdot|\theta_0, \psi_0)$ denotes the distribution of the data under a_0 and $F(1) = F(\cdot|\theta_0, \psi_0, 1)$ denotes the distribution of the data under a. One such $F(\mu)$, for instance, assumes that a fraction μ of the observations are drawn from a distribution consistent with a, while the remaining $1-\mu$ are drawn from a distribution consistent with a_0.

We say that a sequence $\{\mu_n\}_{n=1}^{\infty}$ is a *local perturbation* if under $F_n(\mu_n)$: (i) $\hat{\theta} \xrightarrow{p} \theta_0$; (ii) $\sqrt{n}\hat{g}(\theta_0)$ converges in distribution to a random variable $\tilde{g}$; (iii) $\hat{g}(\theta)$ and $\hat{G}(\theta)$ converge uniformly in probability to $g(\theta)$ and $G(\theta)$; and (iv) $\hat{W} \xrightarrow{p} W$. Any sequence μ_n such that $F_n(\mu_n)$ is contiguous to $F_n(0)$ (see van der Vaart 1998) and under which $\sqrt{n}\hat{g}(\theta_0)$ has a well-defined limiting distribution is a local perturbation. Under this approach, we wish to relate changes in the expectation of $\tilde{g}$ to the first-order asymptotic bias of the estimator, which we generally abbreviate to "asymptotic bias" for ease of exposition.

Example. (OLS, cont'd) Suppose that under alternative assumptions a, the data are in fact generated by

$$Y_i = X_i'\theta_0 + V_i + \varepsilon_i,$$

where the scalar V_i is an omitted variable potentially correlated with X_i and $\mathbb{E}(\varepsilon_i|X_i) = 0$ still. The mean of the OLS moment condition is $\mathbb{E}[\hat{g}(\theta_0)] = \mathbb{E}[X_iV_i] = \Omega_{XV}$, where Ω_{AB} denotes $\mathbb{E}[A_iB_i']$ for vectors A and B.

To define a local perturbation corresponding to this alternative, let $F(\mu)$ be the distribution of data from the model

$$Y_i = X_i'\theta_0 + \mu V_i + \varepsilon_i, \tag{3}$$

and consider the sequence $\mu_n = \frac{1}{\sqrt{n}}$. Analyzing the behavior of $\hat{\theta}$ under this assumption, we can show that $\sqrt{n}\hat{g}(\theta_0)$ converges to a random variable $\tilde{g}$ with expectation Ω_{XV}, and $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ converges to a random variable $\tilde{\theta}^{OLS}$ with expectation

$$\begin{aligned} \mathbb{E}\left(\tilde{\theta}^{OLS}\right) &= \Omega_{XX}^{-1}\Omega_{XV} \\ &= \Lambda\mathbb{E}(\tilde{g}). \end{aligned}$$

The expression $\Omega_{XX}^{-1}\Omega_{XV}$ is the large-sample analogue of the standard omitted variables bias formula. Sensitivity Λ thus gives an expression for asymptotic omitted variables bias analogous to the usual finite-sample expression.

The standard omitted variables bias formula shows that to predict the bias in the estimator for a specific omitted variable, a reader need only be able to form a guess as to the coefficients $\Omega_{XX}^{-1}\Omega_{XV}$ from a regression of the omitted variable on the endogenous regressors. The matrix Ω_{XX}^{-1}—our sensitivity measure Λ in this case—translates the deviation Ω_{XV} in the moments into bias in the estimator. Our main result extends this logic to our more general setup.

Proposition 1. *For any local perturbation* $\{\mu_n\}_{n=1}^{\infty}$, $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ *converges in distribution under* $F_n(\mu_n)$ *to a random variable* $\tilde{\theta}$ *with*

$$\tilde{\theta} = \Lambda\tilde{g}$$

almost surely. This implies in particular that the first-order asymptotic bias $\mathbb{E}\left(\tilde{\theta}\right)$ *is given by*

$$\mathbb{E}\left(\tilde{\theta}\right) = \Lambda\mathbb{E}(\tilde{g}).$$

Proof. See appendix. □

Two extensions are immediate.

Remark 1. In some cases, we are interested in the sensitivity of a counterfactual or welfare calculation that depends on $\hat{\theta}$, rather than the sensitivity of $\hat{\theta}$ per se. Suppose $c(\cdot)$ is a continuously differentiable function not dependent on the data, with non-zero gradient $C = C(\theta_0) = \frac{\partial}{\partial\theta}c(\theta_0)$ at θ_0. Then under any local perturbation, the delta method implies that $\sqrt{n}\left(c\left(\hat{\theta}\right) - c(\theta_0)\right)$ converges in distribution to $\tilde{c} = C\Lambda\tilde{g}$. We will refer to $C\Lambda$ as the sensitivity of $c\left(\hat{\theta}\right)$.

Remark 2. We may be interested in the sensitivity of some elements of the parameter vector holding other elements constant. Decomposing θ into subvectors (θ_1, θ_2), the conditional sensitivity of the first subvector, fixing the second, is

$$\Lambda_1 = -\left(G_1' W G_1\right)^{-1} G_1' W,$$

for $G_1 = \frac{\partial}{\partial \theta_1} g\left(\theta_{1,0}, \theta_{2,0}\right)$, where $\theta_{1,0}$ and $\theta_{2,0}$ are the true values of θ_1 and θ_2 respectively. Conditional sensitivity Λ_1 measures the asymptotic bias of $\hat{\theta}_1$ under local perturbations when $\hat{\theta}_2$ is held fixed at $\theta_{2,0}$.

An alternative to our local perturbation approach is to consider how the probability limit of $\hat{\theta}$ changes under a fixed alternative a—that is, to consider misspecification that does not vanish as the sample size grows large. We show in the online appendix that if the probability limits of $\hat{\theta}$ and $\hat{g}\left(\theta_0\right)$ under assumptions a are $\theta\left(a\right)$ and $g\left(a\right)$ respectively, we have

$$\begin{aligned} \theta\left(a\right) - \theta_0 &\approx \Lambda\left[g\left(a\right) - g\left(a_0\right)\right] \\ &= \Lambda g\left(a\right), \end{aligned}$$

for a close to a_0 in an appropriate sense. This probability limit approach has the drawback that the fixed misspecification becomes arbitrarily large relative to sampling error in a large sample, making it difficult to apply this approach to adjust inference for the induced bias. For this reason, we follow the literature in focusing on local perturbations. However, the intuition delivered by the two approaches is similar.

Other extensions can also be developed. Our MDE setup directly accommodates maximum likelihood or M-estimators with $\hat{\theta} = \arg\min_\theta \frac{1}{n}\sum_i m\left(D_i, \theta\right)$ if we take $\hat{g}\left(\theta\right)$ to be the first-order conditions of the objective and assume that these suffice to identify θ. Our results can also be extended to accommodate, say, models with local maxima or minima in the objective following the reasoning in Newey and McFadden (1994, section 1). The online appendix shows how to extend our asymptotic results to the case where the sample moments $\hat{g}\left(\theta\right)$ are non-differentiable, as in many simulation-based estimators.

4 Special Cases

Two special cases encompass the applications we present below and provide a template for many other cases of interest. Particular transformations of Λ are sometimes more readily interpretable in certain applications. Below we provide guidance on what we think researchers should report in each case.

4.1 Classical Minimum Distance

The first case of interest is where $\hat{g}(\theta) = \hat{s} - s(\theta)$ for sample statistics $\hat{s}$ and corresponding predictions $s(\theta)$ under the model. We refer to this class collectively as classical minimum distance estimators. Our definition of this case includes estimation by additively separable GMM, simulated method of moments, and indirect inference (Gourieroux et al. 1993; Smith 1993). Examples include the estimators of DellaVigna et al. (2012) and Gourinchas and Parker (2002) which we discuss below, as well as a large number of other papers in industrial organization (e.g., Goettler and Gordon 2011), labor (e.g., Voena 2015), finance (e.g., Nikolov and Whited 2014), and macro (e.g., Christiano et al. 2005).

Definition. *$\hat{\theta}$ is a **classical minimum distance (CMD) estimator** if $\hat{g}(\theta) = \hat{s} - s(\theta)$, where $\mathbb{E}(\hat{s}) = s(\theta_0)$ and $s(\cdot)$ is a function that does not depend on the data.*

When $\hat{\theta}$ is a CMD estimator, sensitivity is $\Lambda = (S'WS)^{-1} S'W$, where S is the matrix of partial derivatives of $s(\theta)$ evaluated at θ_0. A natural category of perturbations to consider in this case are additive shifts of the moment functions due to either misspecification of $s(\theta)$ or measurement error in $\hat{s}$. In such cases, we obtain a simple characterization of the asymptotic bias of the CMD estimator.

Proposition 2. *Suppose that $\hat{\theta}$ is a CMD estimator and under $F_n(\mu)$ $\hat{s} = \tilde{s} + \mu\hat{\eta}$, where $\hat{\eta}$ converges in probability to a vector of constants η and the distribution of $\tilde{s}$ does not depend on μ. Take $\mu_n = \frac{1}{\sqrt{n}}$, and suppose that $\hat{W} \xrightarrow{p} W$ under $F_n(\mu_n)$. Then $\mathbb{E}\left(\tilde{\theta}\right) = \Lambda\eta$.*

Proof. See appendix. □

Since the data affects the CMD estimator through the vector of sample statistics $\hat{s}$, in this setting we suggest either reporting an estimate of Λ (if the units of the elements of $s(\theta)$ are naturally comparable), or else multiplying each element Λ_{pj} of Λ by the standard deviation $\sqrt{\Omega_{jj}}$ of the j^{th} moment, so the elements can be interpreted as the effect of a one-standard-deviation change in the moment on the parameters. A reader can then estimate the asymptotic bias associated with any alternative assumptions a, provided she can build intuition about the way they change the statistics $\hat{s}$.

Example. (Indirect Inference) Suppose that each element $\hat{s}_j$ is the coefficient from a descriptive regression of some outcome Y_{ij} on some predictor X_{ij}, with Y_{ij} and X_{ij} functions of the underlying data D_i. Suppose that the model is exactly identified. Under assumptions a_0, $\mathbb{E}\left[Y_{ij}|X_{ij}\right] = s_j(\theta_0)X_{ij}$ for all j, so $\mathbb{E}(\hat{s}) = s(\theta_0)$. Sensitivity is $\Lambda = S^{-1}$.

Under alternative a, the model omits important correlates of Y_{ij}; in a sample of size n, $\mathbb{E}\left[Y_{ij}|X_{ij}\right] = s_j(\theta_0)X_{ij} + \frac{1}{\sqrt{n}}V_{ij}$ for an omitted variable V_{ij}. Applying proposition 2, the asymptotic bias of the

estimator is

$$\mathbb{E}\left(\tilde{\theta}\right) = S^{-1}\Omega_{XX}^{-1}\Omega_{XV}.$$

In this case, sensitivity links the omitted variables bias in the individual regression coefficients $\hat{s}_j$ to the induced asymptotic bias in $\hat{\theta}$.

4.2 Instrumental Variables

The second case of interest is where the parameters of interest are estimated by nonlinear instrumental variables, with moments formed by interacting the instruments with estimated structural errors. Among the examples of this case are the BLP application discussed below and a large set of related demand models, as well as other structural models employing instrumental variables for identification.

Definition. *$\hat{\theta}$ is an **instrumental variables (IV) estimator** if $\hat{g}(\theta) = \frac{1}{n}\sum_i Z_i \otimes \hat{\zeta}_i(\theta)$, where Z_i is a vector of instruments and $\hat{\zeta}_i(\theta)$ is a function of data and parameters with $\mathbb{E}\left(\hat{\zeta}_i(\theta_0)|Z_i\right) = 0$ under F_n.*[4]

When $\hat{\theta}$ is an IV estimator, sensitivity is $\Lambda = -\left(\Omega'_{Z\tilde{X}} W \Omega_{Z\tilde{X}}\right)^{-1}\Omega'_{Z\tilde{X}} W$, where $\Omega_{Z\tilde{X}} = \mathbb{E}(Z_i\tilde{X}_i')$ and $\tilde{X}_i$ are the "pseudo-regressors" $\partial\hat{\zeta}_i(\theta_0)/\partial\theta$.

A natural perturbation to consider in this case is the introduction of an omitted variable V_i that causes the errors ζ_i to be correlated with the instruments Z_i. We provide sufficient conditions for this form of misspecification to be a local perturbation. These conditions apply more generally than nonlinear IV.

Assumption 1. *The observed data $D_i = [Y_i, X_i]$ consist of i.i.d. draws of endogenous variables Y_i and exogenous variables X_i, where $Y_i = h(X_i, \zeta_i; \theta)$ is a one-to-one transformation of the vector of structural errors ζ_i given X_i and θ with inverse $\hat{\zeta}(Y_i, X_i; \theta) = \hat{\zeta}_i(\theta)$. There is also an unobserved (potentially omitted) variable V_i. Under F_n: (i) ζ_i is continuously distributed with full support conditional on X_i; (ii) (ζ_i, X_i, V_i) has a density f with respect to some base measure ν; (iii) $\sqrt{f(\zeta_i, X_i, V_i)}$ is continuously differentiable in ζ_i; (iv) we have*

$$0 < \mathbb{E}\left[\left(\frac{V_i' \frac{\partial}{\partial\zeta} f(\zeta_i, X_i, V_i)}{f(\zeta_i, X_i, V_i)}\right)^2\right] < \infty;$$

[4]For notational simplicity we have assumed that all the instruments Z_i are interacted with each element of $\hat{\zeta}_i(\theta)$. The results derived below continue to apply, however, if we use different instrument sets for different elements of $\hat{\zeta}_i(\theta)$.

and (v) the moments are asymptotically linear in the sense that

$$\sqrt{n}\hat{g}(\theta_0) = \frac{1}{\sqrt{n}}\sum_i \varphi(\zeta_i, X_i, V_i, \theta_0) + o_p(1),$$

where $\varphi(\zeta_i, X_i, V_i, \theta_0)$ *has finite variance.*

The main substantive restriction imposed by assumption 1 is that the structural errors have full support and map one-to-one to the outcomes Y_i. This is satisfied, for example, in BLP (1995) and similar models of aggregate demand. The remaining assumptions are regularity conditions that hold in a wide range of contexts.

Proposition 3. *Suppose that* $\hat{\theta}$ *is an IV estimator satisfying assumption 1, and that under* $F_n(\mu)$ *we have* $\hat{\zeta}_i(\theta_0) = \tilde{\zeta}_i + \mu V_i$*, where* V_i *is an omitted variable with* $\frac{1}{n}\sum_i Z_i \otimes V_i \xrightarrow{p} \Omega_{ZV} \neq 0$ *and the distribution of* $\tilde{\zeta}_i$ *does not depend on* μ*. Then, taking* $\mu_n = \frac{1}{\sqrt{n}}$*, we have* $\mathbb{E}\left(\tilde{\theta}\right) = \Lambda\Omega_{ZV}$*.*

Proof. See appendix. □

Proposition 3 directly generalizes the omitted variables bias formula to locally misspecified nonlinear models. If we consider any just-identified instrumental variables model, then we can restate the conclusion of proposition 3 as

$$\mathbb{E}\left(\tilde{\theta}\right) = -\Omega_{Z\hat{X}}^{-1}\Omega_{ZV}.$$

This is a more general analogue of the omitted variables bias formula: rather than the coefficients from a regression of the omitted variable on the regressors, the asymptotic bias is now given by the coefficients from a two-stage least squares regression of the omitted variable on the pseudo-regressors, using Z as instruments.

If a researcher reports Λ, a reader can predict the asymptotic bias due to any omitted variable provided she can predict its covariance Ω_{ZV} with the instruments. To simplify the reader's task further, we recommend that researchers report an estimate of $\Lambda\Omega_{ZZ}$, possibly multiplied by a scaling matrix that makes the units more comparable across elements of Z. Given $\Lambda\Omega_{ZZ}$, the additional input the reader must provide is the coefficients from a regression of the omitted variable on the excluded instruments—exactly the same input needed to apply the omitted variables bias formula for OLS.

Remark 3. Suppose $\gamma = \Omega_{ZZ}^{-1}\Omega_{ZV}$ are the coefficients from a regression of the omitted variable V_i on the instruments Z_i. Then under the hypotheses of proposition 3, $\mathbb{E}\left(\tilde{\theta}\right) = (\Lambda\Omega_{ZZ})\gamma$.

As a final example, we re-derive the asymptotic bias expression of Conley et al. (2012) for the linear IV model with locally invalid instruments.

Example. (2SLS) Suppose the data are $D_i = [Y_i, X_i, Z_i]$ and the expression for Y_i under the assumed model is the same as in equation (2) with $\mathbb{E}(\varepsilon_i|Z_i) = 0$ and $\mathbb{E}(\varepsilon_i|X_i) \neq 0$. The 2SLS estimator can be written as a GMM estimator with $\hat{g}(\theta) = \frac{1}{n}\sum_i Z_i \otimes (Y_i - X_i'\theta)$ and $\hat{W} = \left(\frac{1}{n}\sum_i Z_i Z_i'\right)^{-1}$. Sensitivity Λ in this case is $\Lambda = \left(\Omega_{ZX}'\Omega_{ZZ}^{-1}\Omega_{ZX}\right)^{-1}\Omega_{ZX}'\Omega_{ZZ}^{-1}$. Conley et al. (2012) consider a perturbed model in which ε_i is replaced by $\frac{1}{\sqrt{n}}Z_i\gamma + \varepsilon_i$. Applying remark 3, we see that the asymptotic bias of 2SLS is

$$\mathbb{E}\left(\tilde{\theta}\right) = \left(\Omega_{ZX}'\Omega_{ZZ}^{-1}\Omega_{ZX}\right)^{-1}\Omega_{ZX}'\gamma.$$

This is the expression Conley et al. (2012) derive in section III.C.

5 Sample Sensitivity

In our analysis thus far we have focused on the sensitivity of the asymptotic behavior of an estimator to changes in identifying assumptions. A distinct but related question is how our estimator $\hat{\theta}$ would change if we used the alternative assumptions a in estimation. In this section, we derive a sensitivity measure which answers this question, and show that it coincides asymptotically with Λ.

Suppose that under the alternative assumptions a we can calculate the probability limit of our moment conditions at the true parameter value, $g(a) = \operatorname{plim}\hat{g}(\theta_0)$. We can use this knowledge to calculate "corrected" moments $\hat{g}^a(\theta) = \hat{g}(\theta) - g(a)$, which under assumptions a again have mean zero at the true parameter value. For a CMD model where we think measurement error biases the first entry of $\hat{s}$ upwards by one unit, for instance, we can take $g(a)$ to be the vector with one in the first entry and zeros everywhere else. Likewise, for an IV model where we think there is an omitted variable V_i that is correlated with the instruments, we can take $g(a) = \Omega_{ZZ}\gamma$ for γ again the regression coefficient of V_i on Z_i.

It is natural to ask how the estimator $\hat{\theta}^a$ derived under a differs from the estimator $\hat{\theta}$ derived under a_0. To provide an approximate answer to this question which can be used to consider many different alternatives a, as in our analysis of asymptotic bias we will consider local approximations. In particular, suppose we can construct a family of moment functions

$$\hat{g}(\theta, \mu) = (1-\mu)\cdot\hat{g}(\theta) + \mu\cdot\hat{g}^a(\theta).$$

Define $\hat{\theta}(\mu)$ to solve

$$\min_{\theta\in\Theta}\hat{g}(\theta,\mu)'\hat{W}\hat{g}(\theta,\mu). \tag{4}$$

For this section, we assume that $\hat{g}(\theta)$ is twice continuously differentiable on Θ.

Define the *sample sensitivity* of $\hat{\theta}$ to $\hat{g}\left(\hat{\theta}\right)$ as

$$\hat{\Lambda}_S = -\left(\hat{G}\left(\hat{\theta}\right)'\hat{W}\hat{G}\left(\hat{\theta}\right)+\hat{A}\right)^{-1}\hat{G}\left(\hat{\theta}\right)'\hat{W},$$

where

$$\hat{A} = \left[\begin{array}{ccc} \left(\frac{\partial}{\partial\theta_1}\hat{G}\left(\hat{\theta}\right)'\right)\hat{W}\hat{g}\left(\hat{\theta}\right) & \ldots & \left(\frac{\partial}{\partial\theta_P}\hat{G}\left(\hat{\theta}\right)'\right)\hat{W}\hat{g}\left(\hat{\theta}\right) \end{array}\right].$$

Sample sensitivity measures the derivative of $\hat{\theta}$ with respect to perturbations of the moments without any assumptions on the data generating process. Specifically, if $\hat{\theta}$ is the unique solution to (1) and lies in the interior of Θ, then

$$\frac{\partial}{\partial\mu}\hat{\theta}(0) = \hat{\Lambda}_S\left(\hat{g}^a(\theta)-\hat{g}(\theta)\right), \tag{5}$$

whenever $\hat{G}\left(\hat{\theta}\right)'\hat{W}\hat{G}\left(\hat{\theta}\right)+\hat{A}$ is non-singular. (This is proved in the online appendix as a consequence of a more general result.) Thus, if we consider a first-order approximation we obtain

$$\hat{\theta}^a - \hat{\theta} \approx \hat{\Lambda}_S\left(\hat{g}^a(\theta)-\hat{g}(\theta)\right),$$

which is analogous to our proposition 1, except that rather than approximating the asymptotic bias of an estimator, we are now approximating the estimator's finite-sample value relative to a correctly specified alternative.

As is intuitively reasonable, the sample sensitivity $\hat{\Lambda}_S$ relates closely to Λ introduced above. To formalize this relationship, we make an additional technical assumption.

Assumption 2. *For $1 \leq p \leq P$ and $\mathscr{B}_\theta$ a ball around θ_0, $\sup_{\theta\in\mathscr{B}_\theta}\|\frac{\partial}{\partial\theta_p}\hat{G}(\theta)\|$ is asymptotically bounded.*[5]

This condition is satisfied if, for example, $\frac{\partial}{\partial\theta_p}\hat{G}(\theta)$ converges to a continuous function $\frac{\partial}{\partial\theta_p}G(\theta)$ uniformly on $\mathscr{B}_\theta$. Assumption 2 is sufficient to ensure that $\hat{A}\xrightarrow{p}0$. Since the sample analogues of G and W converge to their population counterparts, $\hat{\Lambda}_S$ converges to Λ.

Proposition 4. *Consider a local perturbation μ_n such that assumption 2 holds under $F_n(\mu_n)$. $\hat{\Lambda}_S\xrightarrow{p}\Lambda$ under $F_n(\mu_n)$ as $n\to\infty$.*

Proof. See appendix. □

Remark 4. In contrast to proposition 1, the statement in equation (5) does not rely on asymptotic approximations. Consequently, we can use $\hat{\Lambda}_S$ even in settings where conventional asymptotic approximations are unreliable, such as models with weak instruments or highly persistent data. In

[5] In particular, for any $\varepsilon > 0$, there exists a finite constant $r(\varepsilon)$ such that $\limsup_{n\to\infty} Pr\left\{\sup_{\theta\in\mathscr{B}_\theta}\|\frac{\partial}{\partial\theta_p}\hat{G}(\theta)\| > r(\varepsilon)\right\} < \varepsilon$.

such cases, however, the connection between $\hat{\Lambda}_S$ and Λ generally breaks down, and neither measure necessarily provides a reliable guide to bias.

6 Estimation

Because consistent estimators of G and W are typically needed to perform inference on θ, a consistent plug-in estimator of sensitivity is available at essentially no additional computational cost.

Definition. *Define* ***plug-in sensitivity*** *to be*

$$\hat{\Lambda} = -\left(\hat{G}\left(\hat{\theta}\right)'\hat{W}\hat{G}\left(\hat{\theta}\right)\right)^{-1}\hat{G}\left(\hat{\theta}\right)'\hat{W}.$$

Proposition 5. *For any local perturbation* $\{\mu_n\}_{n=1}^{\infty}$, $\hat{\Lambda} \xrightarrow{p} \Lambda$ *under* $F_n(\mu_n)$.

Proof. By assumption $\hat{G}(\theta) \xrightarrow{p} G(\theta)$ uniformly in θ, so consistency of $\hat{\theta}$ implies that $\hat{G}\left(\hat{\theta}\right) \xrightarrow{p} G$. Since $G'WG$ has full rank and $\hat{W} \xrightarrow{p} W$, the result follows by the continuous mapping theorem. □

Analogous results apply to transformations of sensitivity, such as the measure $\Lambda\Omega_{ZZ}$ suggested for instrumental variables models.

Turn next to inference. Under standard regularity conditions the bootstrap will provide a valid approximation to the sampling variability of $\hat{\Lambda}$. To illustrate, we present bootstrap confidence intervals on functions of Λ for our application to BLP (1995) below. An important caveat is that, under local perturbations, $\hat{\Lambda}$ has asymptotic bias of order $\frac{1}{\sqrt{n}}$ (just as $\hat{\theta}$ does). Thus, the location (but not the width) of bootstrap confidence intervals is distorted and their coverage is not correct.

7 Applications

7.1 Charitable Giving

DellaVigna et al. (2012) use data from a field experiment to estimate a model of charitable giving. In the experiment, solicitors go door-to-door and either ask households to donate or ask households to complete a survey. The two charities in the experiment are the East Carolina Hazard Center (ECU) and the La Rabida Children's Hospital (La Rabida). In some treatments, households are warned ahead of time via a flyer that a solicitor will be coming to their home, and in others they are both warned and given a chance to opt out. Households' responses to these warnings, as well as variation across treatments in amounts given and survey completion, pin down preference parameters that allow the authors to assess the welfare effects of solicitation. The main findings

are that social pressure is an important driver of giving and that the average visited household is made worse off by the solicitation.

The model is a two-period game between a solicitor and a household. In the first period, the solicitor may notify the household of the upcoming solicitation, in which case the household can undertake costly effort to avoid it. If the household does not avoid the solicitation, then the household chooses an amount to donate to the charity. The household may receive utility from giving due to altruism (concern for the total resources of the charity) or warm glow (direct utility from giving). The household may also experience social pressure, which is modeled as a cost that decreases linearly in the donation up to a threshold amount d^*, after which social pressure is zero. The game is solved via backward induction, with households rationally anticipating future social pressure. The threshold d^* is taken to be the sample median donation amount of \$10.

The estimator solves (1) with moments

$$\hat{g}(\theta) = \hat{s} - s(\theta),$$

where the statistics $\hat{s}$ include the share of households opening the door in each treatment, the share giving donations in various ranges in the charity treatments, the share completing the survey in the survey treatments, and the share opting out when this was allowed, and $s(\theta)$ is the expected value of each statistic under the model, computed numerically by quadrature. The parameter vector θ includes determinants of the distribution of altruism and the social pressure cost of choosing not to give. Key parameters, including the cost of social pressure, are allowed to differ between the two charities ECU and La Rabida. The weight matrix $\hat{W}$ is equal to the diagonal of the inverted variance-covariance matrix of the observed statistics $\hat{s}$. Under the assumed model F_n, $\mathbb{E}(\hat{s}) = s(\theta_0)$. This is a CMD estimator as defined above.

A reader of the paper might be concerned that several of the model's assumptions, including the functional forms for the distribution of altruism, the utility function, and the social pressure cost, may not hold exactly. We can apply our measure to make the mapping from moments to estimates more transparent, and so allow a reader to estimate the asymptotic bias under various violations of these assumptions.

We consider a perturbed model under which $\hat{s} = \tilde{s} + \mu\eta$ where η is a vector of constants and the distribution of $\tilde{s}$ does not depend on μ. By proposition 2, under the local perturbation $\mu_n = \frac{1}{\sqrt{n}}$, the first-order asymptotic bias is then $\mathbb{E}\left(\tilde{\theta}\right) = \Lambda\eta$. We estimate Λ with its plug-in using estimates of G and W provided to us by the authors.[6] We focus on the sensitivity of the estimated social

[6]We are grateful to Stefano DellaVigna and his co-authors for providing these inputs. We received the parameter vector $\hat{\theta}$, covariance matrix $\hat{\Omega}$, Jacobian $\hat{G}$, and weight matrix $\hat{W}$ resulting from 12 runs of an adaptive search algorithm. These values differ very slightly from those reported in the published paper, which correspond to 500 runs. To evaluate specific forms of misspecification, we code our own implementation of the prediction function $s(\theta)$ and

pressure in the ECU charity solicitations. We show analogous results for the La Rabida social preference parameter in the online appendix.

Figure 1 plots the column of the estimated Λ corresponding to the per-dollar social pressure cost θ^{cost} of not giving to ECU.[7] The estimated value of this parameter is \$0.14 with a standard error of \$0.08 (DellaVigna et al. 2012). Because the moments are probabilities, we scale the estimated Λ so that it can be read as the effect of a one-percentage-point violation of the given moment condition on the asymptotic bias in θ^{cost}.

Figure 1 provides useful qualitative lessons about the estimator. We indicate with solid circles the elements that DellaVigna et al. (2012) single out as important for this parameter: donations at \$10, donations less than \$10, and the share of people opening the door in the treatment where they were warned by a flyer. DellaVigna et al. (2012) write: "The [social pressure] is identified from two main sources of variation: home presence in the flyer treatment ... and the distribution of small giving (the higher the social pressure, the more likely is small giving and in particular bunching at [\$10])" (38). Figure 1 lines up well with these expectations, reinterpreted as statements about sensitivity rather than identification. Estimated social pressure is increasing in the share of people bunching at \$10 and decreasing in the share donating less than \$10. Estimated social pressure is also decreasing in the share of people opening the door in the flyer treatment, reflecting the model's prediction that a household that anticipates high social pressure costs should not open the door. The absolute magnitude of sensitivity is highest for bunching at \$10. These qualitative patterns might lead a reader to be particularly concerned about alternative assumptions a that affect the likelihood that households give exactly \$10.

To illustrate the way sensitivity can be used to assess specific alternatives, suppose households have reasons other than social pressure to give exactly \$10, for example because this is a convenient cash denomination. In particular, suppose that 99 percent of households obey the model, while 1 percent of households obey the model in all ways except that they choose an exogenous donation amount $\tilde{d}$ (e.g., \$10) conditional on giving. The values of η implied by this alternative can be easily computed using the expected values $s\left(\hat{\theta}\right)$ of the statistics $\hat{s}$ reported in the appendix of the original article.[8] Figure 1 can then be used to estimate the implied asymptotic bias in estimated social pressure.

---

confirm that our calculation of $s\left(\hat{\theta}\right)$ closely matches the published results.

[7] We plot the sensitivities with respect to the elements of $\hat{g}\left(\theta_0\right)$ associated with the ECU treatments.

[8] Consider the steps for computing η when $\tilde{d} = 10$. We begin by altering the expected values $s(\hat{\theta})$ of the statistics $\hat{s}$ reported by DellaVigna et al. (2012) in two ways. First, for each ECU treatment we set the probability of giving \$10 to the total predicted probability of giving. Second, we set the probabilities for giving positive amounts other than \$10 to zero. We then compute η by multiplying the difference between our alternative predicted probabilities and the original ones by 0.01, the share of model violators. To illustrate, the component of η for the probability of giving exactly \$10 under the flyer treatment is $0.01 \times (0.0451 - 0.0056)$. Multiplying by the sensitivity of ECU social pressure cost to this probability, which equals 7.455, gives 0.0029. Asymptotic bias is just the sum of such values—a large majority of which are zero—over all moments.

Figure 2 shows the implied asymptotic bias for a range of alternative values of the exogenous gift amount $\tilde{d}$. As expected, the largest asymptotic bias arises when $\tilde{d} = 10$, exactly the threshold at which DellaVigna et al.'s (2012) model assumes that social pressure ceases. Other exogenous giving levels imply much smaller asymptotic bias. The asymptotic bias at $\tilde{d} = 10$ is equal to 0.008, implying that the estimated social pressure is overstated by roughly five percent of the baseline estimate. If the share of households giving exogenously at \$10 were 10 percent, the projected asymptotic bias would be 0.08, implying the estimated social pressure is overstated by more than 50 percent of the baseline estimate. The online appendix compares these local estimates of sensitivity to a global analogue of sample sensitivity.

The authors could of course have estimated this specific alternative model and reported it as part of their robustness analysis. The value of figure 1 is that it allows readers to evaluate this and a wide range of other alternatives themselves. The qualitative patterns provide guidance about which kinds of violations of the model's assumptions are likely to be most important, and the quantitative values provide an estimate of the magnitude of the asymptotic bias for specific alternatives.

7.2 Lifecycle Consumption

Gourinchas and Parker (2002) estimate a structural model of lifecycle consumption with uncertain income. In the model, households' saving decisions are driven by both precautionary and lifecycle motives. The estimates suggest that precautionary motives dominate up to the mid-40s, with consumers acting as "buffer stock" agents who seek to maintain a target level of assets and consume any additional income over that threshold. Lifecycle savings motives (i.e., saving to smooth consumption at retirement) dominate at older ages, with consumers acting in rough accordance with the permanent income hypothesis. The results provide an economic rationale for both the hump-shaped consumption profile and the high marginal propensity to consume out of income shocks at young ages observed in the data.

Households in the model live and work for a known, finite number of periods. In each period of working life each household receives exogenous labor income that is the product of permanent and transitory components. The permanent component evolves (in logs) as a random walk with drift. The transitory component is an i.i.d. shock that is either zero or is lognormally distributed. Households choose consumption in each period of working life to maximize the expected discounted sum of an isoelastic felicity function, and receive a reduced-form terminal payoff for retirement wealth.

The data D are aggregated to a vector $\hat{s}$ consisting of average log consumption at each age e, adjusted in a preliminary stage for differences in family size, cohort, and regional unemployment rates. The parameters of interest θ are the discount factor, the coefficient of relative risk aversion,

and two parameters governing the payoff in retirement. The model also depends on a second vector of parameters χ, including the real interest rate and the parameters of the income generating process, for which the authors compute estimates $\hat{\chi}$ of the true values χ_0 in a first stage. Under the assumed model F_n

$$\hat{s}_e = s_e(\theta_0, \chi_0) + \varepsilon_e,$$

where $s_e(\theta, \chi)$ is the average log consumption predicted by the model and ε_e is a measurement error satisfying $\mathbb{E}(\varepsilon_e) = 0$ for all e.

The estimator $\hat{\theta}$ solves (1) with moments

$$\hat{g}(\theta) = \hat{s} - s(\theta, \hat{\chi}).$$

The weight matrix $\hat{W}$ is a constant that does not depend on the data. Following the authors' initial approach to inference (Gourinchas and Parker 2002, Table III), we proceed as if $\hat{\chi}$ is also a constant that does not depend on the data.[9] The estimator is then a CMD estimator as defined above.

The condition $\mathbb{E}(\hat{s}) = s(\theta_0, \chi_0)$ depends on a number of underlying economic assumptions. A central one is that consumption and leisure are separable. This implies that the level of income in a given period is not correlated with the marginal utility of consumption. Subsequent literature, however, has shown that working can affect marginal utility in important ways. Aguiar and Hurst (2007) show that shopping intensity increases when consumers work less, implying that lower income increases the marginal utility a consumer can obtain from a given expenditure on consumption. Aguiar and Hurst (2013) show that a meaningful portion of consumption goes to work related expenses, implying a second reason for non-separability. Since work time and work-related expenses both vary systematically with age, these forces would change the age-consumption profile relative to what the Gourinchas and Parker (2002) model would predict.

Another important assumption is that there are no unobserved components of income that vary systematically over the lifecycle. If younger consumers receive transfers from their families, for example, consumption relative to income would look artificially high at young ages. An example is the in-kind housing support from parents studied by Kaplan (2012). Gourinchas and Parker (2002) note that their data exhibit consumption in excess of income in the early years of adulthood (something that is impossible under the assumptions of their model), and they speculate that this could be explained by such unobserved transfers.

We show how a reader can use sensitivity to asses the asymptotic bias introduced by violations of these assumptions. We focus on the sensitivity of the two key preference parameters—the

[9] If we instead let $\hat{\chi}$ depend on the data, the analysis below and, by lemma 1, its interpretation in terms of misspecification are preserved, provided that the distribution of $\hat{\chi}$ does not vary with the perturbation parameter μ. This assumption seems reasonable in this context because estimation of $\hat{\chi}$ is based on separate data that does not involve the consumption observations underlying $\hat{s}$.

discount factor and the coefficient of relative risk aversion—which in turn determine the relative importance of consumption smoothing and precautionary incentives.[10] Each violation we consider leads to a divergence between observed consumption and the consumption quantity predicted by the model. Formally, we consider perturbed models $F_n(\mu)$ under which $\varepsilon = \tilde{\varepsilon} + \mu\eta$, where the distribution of $\tilde{\varepsilon}$ does not depend on μ and η is a vector of constants that will differ depending on the alternative model at hand. We take $\mu_n = \frac{1}{\sqrt{n}}$. By proposition 2, the asymptotic bias is then $\mathbb{E}\left(\tilde{\theta}\right) = \Lambda\eta$. We estimate the model using the authors' original code and data, and then estimate Λ with its plug-in.[11]

Figure 3 plots the columns of the estimated Λ corresponding to the discount factor and the coefficient of relative risk aversion. The two plots are essentially inverse to one another. This reflects the fact that both a higher discount factor and a higher coefficient of relative risk aversion imply the same qualitative change in the consumption profile: lower consumption early in life and greater consumption later in life. A change in consumption at a particular age that leads to higher estimates of one parameter thus tends to be offset by a reduction in the other parameter in order to hold consumption at other ages constant. The two parameters are separately identified because they have different quantitative implications at different ages, depending on the relative importance of precautionary and lifecycle savings.

Figure 3 reveals useful qualitative lessons about the estimator. The plots suggest that we can divide the lifecycle into three periods. Up to the late 30s, saving is primarily precautionary, so risk aversion matters comparatively more than discounting and higher consumption is interpreted as evidence of low risk aversion. From the late 30s to the early 60s, incentives shift toward retirement savings, so discounting matters comparatively more than risk aversion and higher consumption is interpreted as evidence of a low discount factor. From the early 60s on, retirement savings continues to be the dominant motive, but now we are late enough in the lifecycle that high consumption signals the household has already accumulated substantial retirement wealth and thus is interpreted as evidence of a high discount factor. These divisions align well with the phases of precautionary and lifecycle savings that Gourinchas and Parker (2002) highlight in their figure 7.

Figure 3 also permits readers to form quantitative intuitions about the asymptotic bias in the estimator. Suppose, for example, that a reader believes that 26-year-olds overstate their consumption by 1 percent (0.01 log points). Then the reader believes that the estimated discount factor is biased (asymptotically) upwards by $0.0006 = 0.01 \times 0.06$ where 0.06 is roughly the sensitivity of the discount factor to the moment corresponding to log consumption at age 26.

[10] We fix the two retirement parameters at their estimated values for the purposes of our analysis.

[11] We are grateful to Pierre-Olivier Gourinchas for providing the original GAUSS code, first-stage parameters, and input data. We use the published parameter values as starting values. We compute sensitivity at the value $\hat{\theta}$ to which our run of the solver converges, and report this value as the baseline estimate in table 1 below. This value is similar, though not identical, to the published parameters.

A range of economically interesting assumptions η can be translated into implied asymptotic bias using the elements of figure 3. To illustrate, table 1 shows the first-order asymptotic bias associated with each of four specific perturbations. First, to allow for variable shopping intensity, we define the elements η_e to match the age-specific log price increments that Aguiar and Hurst (2007) estimate in column 1 of their table I. Second, to allow for work-related consumption expenses, we define η_e so that true consumption at each age is overstated by five percent of work-related expenses as calculated in Aguiar and Hurst's (2013) table 1 and figure 2a. Third, to allow for young consumers receiving family transfers, we choose η_e so that true average consumption prior to age 30 is one percent below average income (rather than above average income as the raw data suggest). Finally, to allow older consumers to make corresponding transfers to their children, we choose η_e so that consumption from ages 50 through 65 is overstated by an annual amount whose lifetime sum is equal to the cumulative gap between consumption and income over ages 26 through 29.

The first row of table 1 shows that if shopping intensity changes with age as in Aguiar and Hurst (2007), the estimated discount factor is overstated by 0.4 percentage points and the estimated coefficient of relative risk aversion is understated by roughly a third of its corrected value. The second row shows that if there are significant work-related expenses as in Aguiar and Hurst (2013), the estimated discount factor and coefficient of relative risk aversion are asymptotically biased in the opposite direction. The third row shows that if part of the measured consumption of young workers is funded by unobserved transfers, the discount factor is overstated by more than a percentage point and the coefficient of relative risk aversion is understated by half of its corrected value. The fourth row shows that allowing for older consumers to fund such transfers has a more modest effect in the opposite direction. The final row shows the net effect when we account for transfers both from the old and to the young.

Importantly, each of these alternatives can be contemplated based only on figure 3 and other basic information provided in Gourinchas and Parker (2002) (e.g., the average log consumption and income at each age). This illustrates the sense in which a plot like figure 3 can aid transparency by letting readers consider the effects of different forms of misspecification on the asymptotic behavior of the estimator, without direct access to the estimation code or data.

7.3 Automobile Demand

BLP (1995) use data on US automobiles from 1971 to 1990 to estimate a structural model of demand and pricing. The model yields estimates of markups and cross-price elasticities, which can in turn be used to evaluate changes such as trade restrictions (BLP 1999), mergers (Nevo 2000), and the introduction of a new good (Petrin 2002). We follow BLP (1995) in suppressing

the time dimension of the data in our notation.

The data $D = [S, P, X, Z]$ consist of a vector of endogenous market shares S; a vector of endogenous prices P; a matrix X of exogenous car characteristics such as size and mileage; and a matrix $Z = \begin{bmatrix} Z_d & Z_s \end{bmatrix}$ of instruments partitioned into those used to estimate the demand-side and supply-side equations respectively. An observation i is a vehicle model. The instruments Z are functions of X, with row Z_i containing functions of the number and characteristics X_{-i} of models other than i (including other car models produced by the same firm).[12]

The demand model is a random-coefficients logit in which the utility from purchasing a given vehicle model i depends on its characteristics X_i and an unobserved preference factor ξ_i. The marginal cost of producing vehicle model i likewise depends on its characteristics X_i and an unobserved cost factor ω_i. Consumers make purchase decisions to maximize utility. Multi-product firms set prices simultaneously to maximize profits. Equilibrium prices correspond to a Bertrand-Nash equilibrium.

Under the assumed model F_n,

$$S = s(X, \xi, \omega; \theta_0)$$

$$P = p(X, \xi, \omega; \theta_0),$$

where $\mathbb{E}(\xi_i | Z_{di}) = \mathbb{E}(\omega_i | Z_{si}) = 0$. The function $s(\cdot)$ maps primitives to market shares under the assumption of utility maximization. The function $p(\cdot)$ maps primitives to prices under the assumption of Nash equilibrium.

Because the functions $s(\cdot)$ and $p(\cdot)$ are known and invertible, it is possible to compute the errors $\hat{\xi}_i(\theta)$ and $\hat{\omega}_i(\theta)$ implied by given parameters and data. The estimator $\hat{\theta}$ solves (1) with moments

[12]The elements of Z_{di} are (i) a constant term (equal to one); (ii) horsepower per 10 pounds of weight; (iii) an indicator for standard air conditioning; (iv) mileage measured in ten times miles per dollar (miles per gallon divided by the average real retail price per gallon of gasoline in the respective year); (v) size (length times width); (vi) the sum of (i)-(v) across models other than i produced in the same year by the same firm as i; and (vii) the sum of (i)-(v) across models produced in the same year by rival firms. This yields 15 instruments, of which all except (i)-(v) are "excluded" in the sense that they do not also enter the utility function directly. We drop two of these instruments—the sums of (v) across same-firm and rival-firm models—because they are highly collinear with the others. This leaves 13 instruments (8 excluded) for estimation. The elements of Z_{si} are (i) a constant term; (ii) the log of horsepower per 10 pounds of weight; (iii) an indicator for standard air conditioning; (iv) the log of ten times mileage measured in miles per gallon; (v) the log of size; (vi) a time trend equal to the year of model i minus 1971; (vii) mileage measured in miles per dollar; (viii) the sum of (i)-(vi) across models other than i produced in the same year by the same firm as i; and (ix) the sum of (i)-(vi) across models produced in the same year by rival firms. This yields 19 instruments, of which all except (i)-(vi) are excluded. The inclusion of (vii) as an excluded instrument in Z_{si} is motivated by the assumption that marginal cost depends on miles per gallon but not on the retail gasoline price (which creates variation in miles per dollar conditional on miles per gallon). The sum of (vi) across rival firms' models is dropped due to collinearity, leaving 18 instruments (12 excluded) for estimation. We demean all instruments other than those involving the constant terms.

$$\hat{g}(\theta) = \frac{1}{n}\begin{bmatrix} \sum_i Z_{di}' \otimes \hat{\xi}_i(\theta) \\ \sum_i Z_{si}' \otimes \hat{\omega}_i(\theta) \end{bmatrix}.$$

The weight matrix is the inverse of the variance-covariance matrix of $\hat{g}\left(\hat{\theta}^{FS}\right)$, where $\hat{\theta}^{FS}$ denotes the first-stage estimator.

The demand and supply moment conditions $\mathbb{E}(\xi_i|Z_{di}) = 0$ and $\mathbb{E}(\omega_i|Z_{si}) = 0$ encode distinct economic assumptions. The demand-side condition $\mathbb{E}(\xi_i|Z_{di}) = 0$ requires that the unobserved component ξ_i of the utility from purchasing model i is mean-independent of the number and characteristics of cars other than i in a given year. The assumption is especially reasonable if the determinants of ξ_i are unknown until after product line decisions are made. The assumption could be violated if ξ_i depends on anticipated shocks to preferences that affect the number of models introduced or their characteristics. Draganska et al. (2009), Fan (2013), and Wollmann (2016) estimate models in which firms' choices of products and product characteristics depend on consumer demand.

The supply-side condition $\mathbb{E}(\omega_i|Z_{si}) = 0$ requires that the unobserved component ω_i of the marginal cost of producing model i is mean-independent of the number and characteristics of cars other than i. This assumption could be violated if a firm's product line affects the cost of producing a given model through economies of scope or scale. Levitt et al. (2013) show that learning-by-doing leads to large economies of scale in automobile production, though the effects they document accrue within rather than across models.[13]

We show how a reader can use sensitivity to assess the asymptotic bias in the estimated markup implied by violations of the exclusion restrictions. We estimate the model using BLP's (1995) data and our own implementation of the authors' estimator.[14] We consider a perturbed model $F_n(\mu)$ under which the instruments influence the structural errors, i.e.

$$\begin{bmatrix} \hat{\xi}_i(\theta_0) \\ \hat{\omega}_i(\theta_0) \end{bmatrix} = \begin{bmatrix} \tilde{\xi}_i \\ \tilde{\omega}_i \end{bmatrix} + \mu \begin{bmatrix} Z_{di}'\gamma_d \\ Z_{si}'\gamma_s \end{bmatrix}, \tag{6}$$

where the distribution of $\begin{bmatrix} \tilde{\xi}_i & \tilde{\omega}_i \end{bmatrix}'$ does not depend on μ. We consider the sequence of pertur-

[13] BLP (1995) also discuss the possibility of within-model increasing returns, finding some support for it in their reduced-form estimates (876).

[14] We obtained data and estimation code for BLP (1999) from an archived version of Jim Levinsohn's web page (https://web.archive.org/web/20041227055838/http://www-personal.umich.edu/~jamesl/verstuff/instructions.html, accessed July 16, 2014). We confirm using the summary statistics in BLP (1995) that the data are the same as those used in the BLP (1995) analysis. Since the algorithms in the two papers are almost identical, we follow the BLP (1999) code as a guide to implementing the estimation, and in particular follow the algorithm in this code for choosing which instruments to drop due to collinearity. We use the published BLP (1995) parameters as starting values and in computing importance sampling weights. We compute sensitivity at the parameter vector $\hat{\theta}$ we estimate, which is similar though not identical to the published estimates.

bations $\mu_n = \frac{1}{\sqrt{n}}$ and assume that the regularity conditions of assumption 1 are satisfied. Letting C denote the gradient of the markup, defined as the ratio of price minus marginal cost to price, with respect to θ at θ_0, remark 1 and remark 3 imply that the asymptotic bias in the markup is $C\Lambda\tilde{\Omega}_{ZZ}\gamma$, where $\tilde{\Omega}_{ZZ} = \begin{bmatrix} \mathbb{E}(Z_{di}Z_{di}') & 0 \\ 0 & \mathbb{E}(Z_{si}Z_{si}') \end{bmatrix}$ and $\gamma = \begin{bmatrix} \gamma_d & \gamma_s \end{bmatrix}'$.[15] We estimate C, Λ and $\tilde{\Omega}_{ZZ}$ with their respective plug-ins. The vector of constants γ encodes a reader's beliefs about the excludability of the instruments, with $\gamma = 0$ corresponding to BLP's (1995) assumptions.

Figure 4 plots the estimated value of $C\Lambda\tilde{\Omega}_{ZZ}K$, where K is a diagonal matrix whose diagonal elements are normalizing constants that allow us to interpret γ as the effect (in percent of the average price) of a one-standard-deviation change in each instrument on willingness-to-pay (for demand-side instruments Z_{di}) or marginal cost (for supply-side instruments Z_{si}). Elements of $C\Lambda\tilde{\Omega}_{ZZ}K$ corresponding to demand-side instruments are plotted on the left; elements corresponding to supply-side instruments are plotted on the right.[16]

Figure 4 delivers some qualitative lessons that are useful in thinking about BLP's (1995) estimator. It shows that the asymptotic bias in the average markup is very sensitive to whether the number of different vehicle models produced by the firm influences marginal costs directly, suggesting that firm-level economies of scope may be a particularly important threat to the validity of the estimates. More broadly, the plot shows that beliefs about the excludability of supply-side instruments really matter. This is consistent with a sense in the literature that the supply-side moments play a critical role in estimation.[17]

A reader can use figure 4 to assess the asymptotic bias associated with a range of specific alternatives. On the supply side, we suppose that, for a car with average marginal cost at the midpoint sample year, removing a different car from the firm's product line increases the marginal cost by one percent of the average price, say because of lost economies of scope. On the demand side, we assume that removing a car from a firm's product line decreases the average willingness to pay for the firm's other cars by one percent of the average price, say because buyers have a preference for buying a car from a manufacturer with a more complete line of cars. We also repeat both exercises for the effect of removing a car from *rival* firms' product lines, which could matter because of industry-wide economies of scope (on the supply side) or effects on consumer search

[15] Sensitivity is

$$\Lambda = -\left(\Omega_{Z\tilde{X}}'W\Omega_{Z\tilde{X}}\right)^{-1}\Omega_{Z\tilde{X}}'W,$$

where the pseudo-regressors are $\tilde{X}_i = \begin{bmatrix} \frac{\partial\hat{\xi}_i(\theta_0)}{\partial\theta} & \frac{\partial\hat{\omega}_i(\theta_0)}{\partial\theta} \end{bmatrix}'$.

[16] The online appendix provides a table showing the standard deviation of each instrument so that a reader can easily transform γ into native units. The online appendix also reports an analogue of figure 4 based on sample sensitivity.

[17] In the original article, BLP (1995) note that they had estimated the model with the demand moments alone and found that this led to "much larger estimated standard errors" (875). In subsequent work, the authors recall finding that "estimates that used only the demand system were too imprecise to be useful" (BLP 2004, 92).

behavior (on the demand side).

Table 2 shows that all of these beliefs imply meaningful first-order asymptotic bias in the estimated average markup. The first violation of the supply-side exclusion restrictions, for example, would mean that the estimated markup of 0.33 is biased (asymptotically) downward by 17 percentage points, implying a corrected estimate of 0.50. The violation of the demand-side exclusion restrictions has an effect of similar magnitude, biasing the markup downward by 13 percentage points. The online appendix compares these local estimates of sensitivity to a global analogue of sample sensitivity.

Importantly, all of the asymptotic bias calculations reported in table 2 can be read off of figure 4: the estimated biases correspond to the lengths (and signs) of their corresponding elements in the plot times the standard deviations of their associated instruments, which are reported in the online appendix. An implication is that a reader interested in any particular violation γ of the exclusion restrictions can approximate its effect by reading the appropriate elements of the plot. For example, a reader who thinks that a one-standard-deviation increase in fuel economy increases marginal cost by two percent can learn that this implies a positive asymptotic bias of $0.0018 = 0.0013 \times 0.6981 \times 2$ in the average markup. A reader could also combine multiple elements of figure 4 to approximate the effect of multiple violations of the exclusion restrictions—say, a direct effect of both number of cars and fuel economy on marginal cost.

8 Conclusions

We propose a method for increasing the transparency of structural estimates by permitting readers to quantify the effects of a wide range of violations of identifying assumptions on the asymptotic behavior of the estimator. We provide several formal interpretations of our proposed approach and we illustrate it with three substantive applications. In all three cases, we argue that readers of the original article would have benefited from the information we propose to present.

References

Aguiar, Mark and Erik Hurst. 2007. Life-cycle prices and production. *American Economic Review* 97(5): 1533-1559.

—. 2013. Deconstructing life cycle expenditure. *Journal of Political Economy* 121(3): 437-492.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. An evaluation of instrumental variable strategies for estimating the effects of Catholic schooling. *Journal of Human Resources* 40(4): 791-821.

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

Berger, David and Joseph Vavra. 2015. Consumption dynamics during recessions. *Econometrica* 83(1): 101-154.

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

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* 63(4): 841-890.

—. 1999. Voluntary export restraints on automobiles: Evaluating a trade policy. *American Economic Review* 89(3): 400-430.

—. 2004. Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of Political Economy* 112(1): 68-105.

Chen, Xiaohong, Elie Tamer, and Alexander Torgovitsky. 2011. Sensitivity analysis in semiparameteric likelihood models. Cowles Foundation Discussion Paper No. 1836.

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

Crawford, Gregory S., and Ali Yurukoglu. 2012. The welfare effects of bundling in multichannel television markets. *American Economic Review* 102(2): 643-685.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113(1): 1-45.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics* 127(1): 1-56.

Draganska, Michaela, Michael Mazzeo, and Katja Seim. 2009. Beyond plain vanilla: Modeling joint product assortment and pricing decisions. *Quantitative Marketing and Economics* 7(2): 105-146.

Einav, Liran, Amy Finkelstein, and Paul Schrimpf. 2015. The response of drug expenditures to

non-linear contract design: Evidence from Medicare Part D. *Quarterly Journal of Economics* 130(2): 841–899.

Fan, Ying. 2013. Ownership consolidation and product characteristics: A study of the US daily newspaper market. *American Economic Review* 103(5): 1598-1628.

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

Glad, Ingrid and Nils Lid Hjort. 2016. Model uncertainty first, not afterwards. *Statistical Science* 31(4): 490-494.

Goettler, Ronald L. and Brett R. Gordon. 2011. Does AMD spur Intel to innovate more? *Journal of Political Economy* 119(6): 1141-1200.

Gourieroux, Christian S., Alain Monfort, and Eric Renault. 1993. Indirect inference. *Journal of Applied Econometrics* 8: S85-S118.

Gourinchas, Pierre-Olivier and Jonathan A . Parker. 2002. Consumption over the life cycle. *Econometrica* 70(1): 47-89.

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

Heckman, James J. 2010. Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature* 48(2): 356-398.

Huber, Peter J. and Elvezio M. Ronchetti. 2009. *Robust statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Kaplan, Greg. 2012. Moving back home: Insurance against labor market risk. *Journal of Political Economy* 120(3): 446-512.

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

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

Kristensen, Dennis and Bernard Salanié. Forthcoming. Higher order properties of approximate estimators. *Journal of Econometrics*.

Leamer, Edward E. 1983. Let's take the con out of econometrics. *American Economic Review* 73(1): 31-43.

Levitt, Steven D., John A. List, and Chad Syverson. 2013. Toward an understanding of learning by doing: Evidence from an automobile assembly plant. *Journal of Political Economy* 121(4): 643-681.

Matzkin, Rosa L. 2013. Nonparametric identification in structural economic models. *Annual Review of Economics* 5: 457-486.

Morten, Melanie. 2016. Temporary migration and endogenous risk sharing in village India. NBER Working Paper No. 22159.

Müller, Ulrich K. 2012. Measuring prior sensitivity and prior informativeness in large Bayesian models. *Journal of Monetary Economics* 59(6): 581-597.

Nevo, Aviv. 2000. Mergers with differentiated products: The case of the ready-to-eat cereal industry. *RAND Journal of Economics* 31(3): 395-421.

Nevo, Aviv and Adam M. Rosen. 2012. Identification with imperfect instruments. *Review of Economics and Statistics* 94(3): 659-671.

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

Newey, Whitney K. and Daniel McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, edited by R. Engle and D. McFadden, 4: 2111-2245. Amsterdam: Elsevier, North-Holland.

Nikolov, Boris and Toni M. Whited. 2014. Agency conflicts and cash: Estimates from a dynamic model. *Journal of Finance* 69(5): 1883-1921.

Petrin, Amil. 2002. Quantifying the benefits of new products: The case of the minivan. *Journal of Political Economy* 110(4): 705-729.

Saltelli, Andrea, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. 2008. *Global sensitivity analysis: The primer*. West Sussex, UK: John Wiley & Sons Ltd.

Smith, Anthony A. 1993. Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics* 8: S63-S84.

Sobol, Ilya M. 1993. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiments* 1(4): 407-414.

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

Voena, Alessandra. 2015. Yours, mine and ours: Do divorce laws affect the intertemporal behavior of married couples? *American Economic Review* 105(8): 2295-2332.

Wollmann, Thomas. 2016. Trucks without bailouts: Equilibrium product characteristics for commercial vehicles. Working paper. Accessed at <http://faculty.chicagobooth.edu/thomas.wollmann/docs/Trucks_without_Bailouts_Wollmann.pdf> on March 16, 2017.

Table 1: Asymptotic bias of preference parameters in Gourinchas and Parker (2002) under particular local violations of identifying assumptions

	Bias in discount factor	Bias in coefficient of relative risk aversion
Consumption and leisure are nonseparable:		
Shopping intensity changes with age	0.0041	-0.2913
Exclude 5% of work-related expenses	-0.0073	0.3997
Consumption includes interhousehold transfers:		
Consumption at early ages includes transfers in	0.0107	-0.6022
Consumption at later ages includes transfers out	-0.0041	0.2673
Include both early and late transfers	0.0065	-0.3349
Baseline estimate	0.9574	0.6526

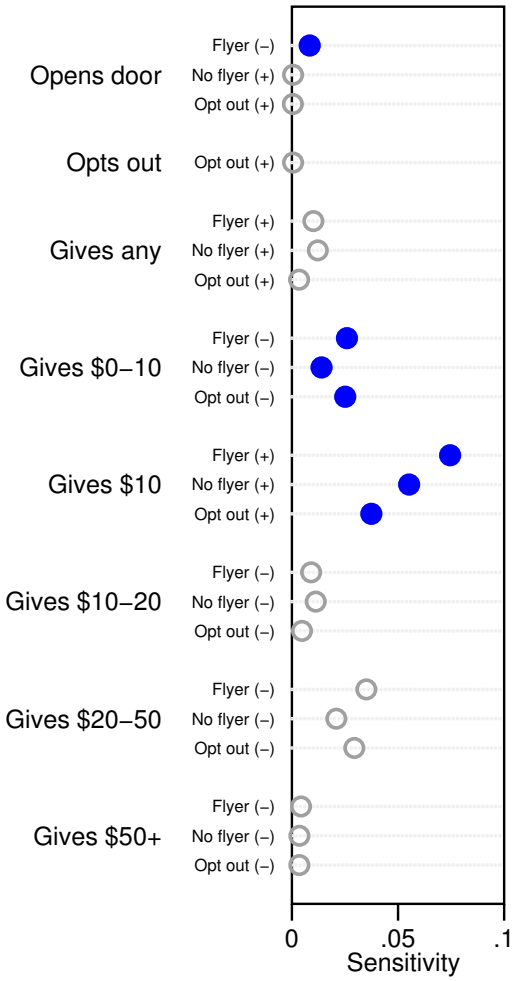
Note: The table reports the estimated first-order asymptotic bias in Gourinchas and Parker's (2002) published parameter values under various forms of misspecification, as implied by proposition 2. Our calculations use the plug-in estimator of sensitivity. We consider perturbations under which measured log consumption overstates true log consumption at each age e by an amount equal to $\eta_e/\sqrt{n}$. In the row labeled "shopping intensity changes with age," η_e is chosen to match the age-specific log price increment estimated in Aguiar and Hurst (2007, column 1 of table I). Aguiar and Hurst (2007) report these increments for ages 30 and above. We set increments for younger ages to zero. In the row labeled "exclude 5% of work-related expenses," η_e is chosen so that the true consumption at each age e is overstated by five percent of work-related expenses as calculated in Aguiar and Hurst (2013, table 1 and figure 2a). In the row labeled "consumption at early ages includes transfers in," η_e is chosen so that true average consumption prior to age 30 is one percent below average income. In the row labeled "consumption at later ages includes transfers out," η_e is chosen so that from age 50 through age 65 consumption is overstated by a constant annual amount whose lifetime sum is equal to the total gap between consumption and income over ages 26 through 29. In the row labeled "include both early and late transfers," η_e combines the early age and later age transfers.

Table 2: Asymptotic bias of average markup in BLP (1995) under particular local violations of the exclusion restrictions

	Bias in average markup
Violation of supply-side exclusion restrictions:	
Removing own car increases average marginal cost by 1% of average price	-0.1731 (0.0433)
Removing rival's car increases average marginal cost by 1% of average price	0.2095 (0.0689)
Violation of demand-side exclusion restrictions:	
Removing own car decreases average willingness to pay by 1% of average price	-0.1277 (0.0915)
Removing rival's car decreases average willingness to pay by 1% of average price	0.2515 (0.1285)
Baseline estimate	0.3272 (0.0392)

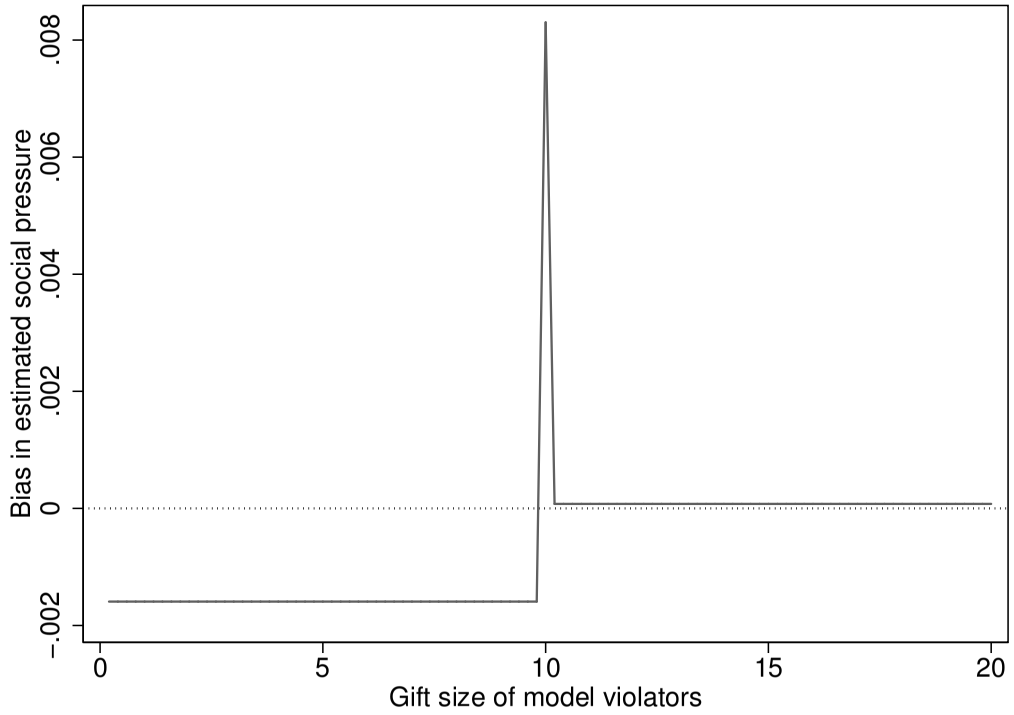
Note: The average markup is the average ratio of price minus marginal cost to price across all vehicles. The table reports the estimated first-order asymptotic bias in the parameter estimates from BLP's (1995) estimator under various forms of misspecification, as implied by proposition 3 under the setup in equation (6). Our calculations use the plug-in estimator of sensitivity. In the first two rows, we set $V_{di} = 0$ and $V_{si} = -0.01\left(\overline{P}/\overline{mc}\right) Num_i$, where Num_i is the number of cars produced by the [same firm / other firms] as car i in the respective year, $\overline{mc}$ is the sales-weighted mean marginal cost over all cars i in 1980, and $\overline{P}$ is the sales-weighted mean price over all cars i in 1980. In the second two rows, we set $V_{si} = 0$ and $V_{di} = 0.01\left(\overline{P}/K_\xi\right) Num_i$, where K_ξ is the derivative of willingness to pay with respect to ξ for a 1980 household with mean income. Standard errors are obtained from a non-parametric block bootstrap over sample years with 70 replicates. We hold the average price $\overline{P}$, the marginal cost $\overline{mc}$, and the derivative K_ξ constant across bootstrap replications.

Figure 1: Sensitivity of ECU social pressure cost in DellaVigna et al. (2012) to local violations of identifying assumptions



Notes: The plot shows one-hundredth of the absolute value of plug-in sensitivity of the social pressure cost of soliciting a donation for the East Carolina Hazard Center (ECU) with respect to the vector of estimation moments, with the sign of sensitivity in parentheses. While sensitivity is computed with respect to the complete set of estimation moments, the plot only shows those corresponding to the ECU treatment. Each moment is the observed probability of a response for the given treatment group. The leftmost axis labels in larger font describe the response; the axis labels in smaller font describe the treatment group. Filled circles correspond to moments that DellaVigna et al. (2012) highlight as important for the parameter.

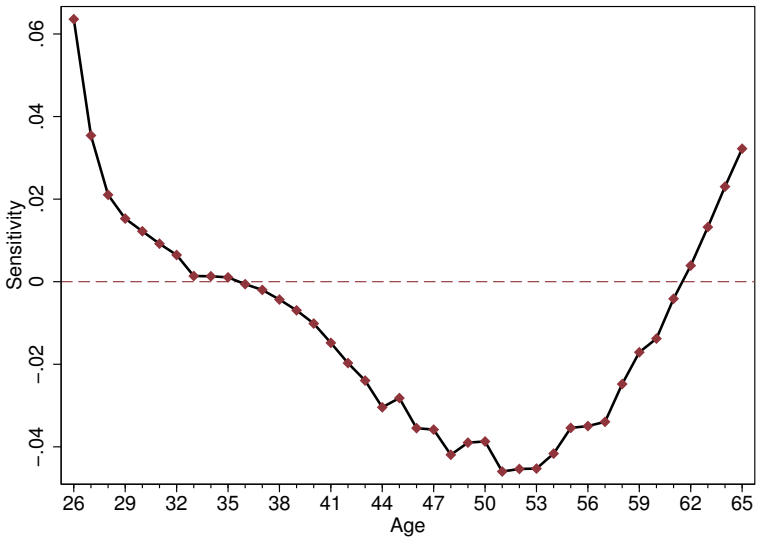
Figure 2: Sensitivity of ECU social pressure cost in DellaVigna et al. (2012) to exogenous gift levels



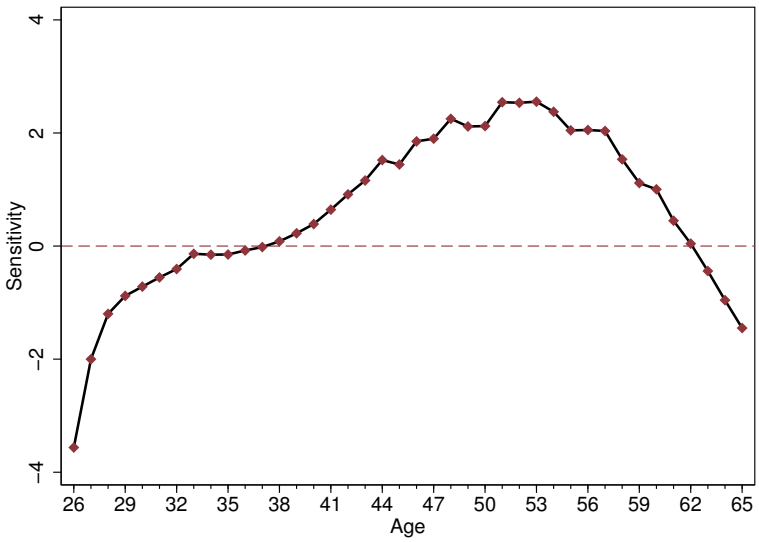
Notes: The plot shows the estimated first-order asymptotic bias in DellaVigna et al.'s (2012) published estimate of the per-dollar social pressure cost of not giving to the East Carolina Hazard Center (ECU) under various levels of misspecification, as implied by proposition 2. Our calculations use the plug-in estimator of sensitivity. We consider perturbations under which a share $\left(1 - \frac{0.01}{\sqrt{n}}\right)$ of households follow the paper's model and and a share $\frac{0.01}{\sqrt{n}}$ give with the same probabilities as their model-obeying counterparts but always give an amount $\tilde{d}$ conditional on giving. First-order asymptotic bias is computed for values of $\tilde{d}$ in \$0.20 increments from \$0 to \$20 and interpolated between these increments. Values of $\tilde{d}$ are shown on the x axis.

Figure 3: Sensitivity of select parameters in Gourinchas and Parker (2002) to local violations of identifying assumptions

Panel A: Discount factor

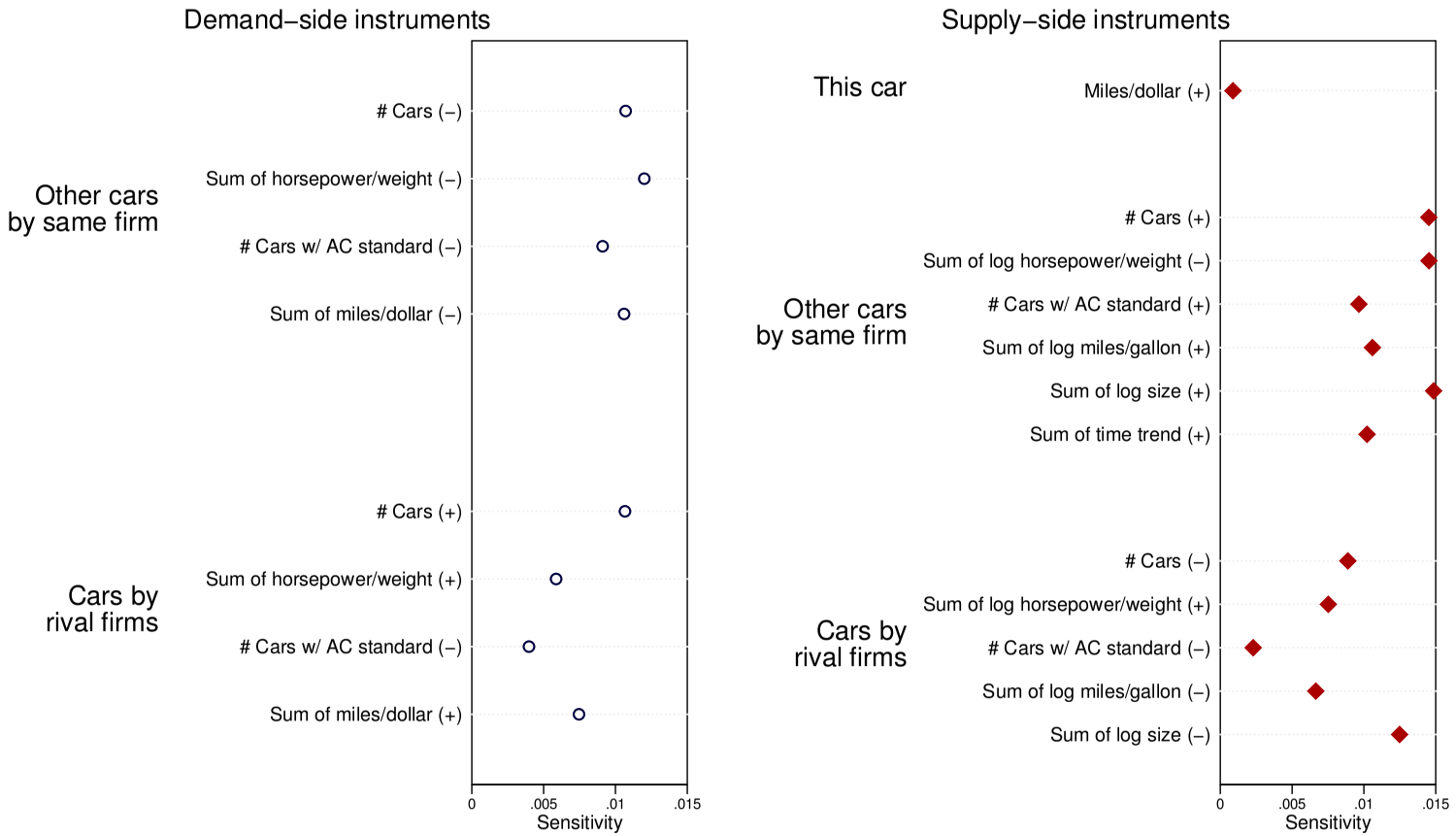


Panel B: Coefficent of relative risk aversion



Notes: Each plot shows the plug-in sensitivity of the parameter named in the plot title with respect to the full vector of estimation moments, which are the mean adjusted log of consumption levels at each age.

Figure 4: Sensitivity of average markup in BLP (1995) to local violations of the exclusion restrictions



Notes: The plot shows the absolute value of the plug-in for $C\Lambda\tilde{\Omega}_{ZZ}K$, where C is the gradient of the average markup with respect to model parameters, Λ is sensitivity of parameters to estimation moments, $\tilde{\Omega} = \begin{bmatrix} \mathbb{E}(Z_{di}Z_{di}') & 0 \\ 0 & \mathbb{E}(Z_{si}Z_{si}') \end{bmatrix}$, and K is a diagonal matrix of normalizing constants. The sign of $C\Lambda\tilde{\Omega}_{ZZ}K$ is shown in parentheses. For the demand-side instruments on the left, the diagonal elements of K are chosen so that the plotted values can be interpreted as sensitivity of the markup to beliefs about the effect of a one-standard-deviation increase in each instrument on the willingness-to-pay of a household with mean income in 1980, expressed as a percent of the sales-weighted mean price over all cars i in 1980. For the supply-side instruments on the right, the diagonal elements of K are chosen so that the plotted values can be interpreted as sensitivity of the markup to beliefs about the effect of a one-standard-deviation increase in each instrument on the marginal cost of a car with the sales-weighted average marginal cost in 1980, expressed as a percent of the sales-weighted mean price over all cars i in 1980. While sensitivity is computed with respect to the complete set of estimation moments, the plot only shows those corresponding to the excluded instruments (those that do not enter the utility or marginal cost equations directly).

A Relationship to Alternative Measures of Sensitivity to Moments

A.1 Dropping Moments

One common method for assessing the relevance of particular moments is to re-estimate the model parameters after dropping the corresponding moment condition from the function $\hat{g}(\theta)$ (see, e.g., Altonji et al. 2005). The following result specifies how this procedure is related to sensitivity Λ.

Corollary 1. *Consider the setup of proposition 1, and suppose that under the local perturbation $\{\mu_n\}_{n=1}^{\infty}$ only one moment j is potentially misspecified ($\mathbb{E}(\tilde{g}_k) = 0$ for $k \neq j$). Let $\hat{\theta}^j$ be the estimator that results from excluding the j^{th} moment condition and suppose that this estimator satisfies our maintained assumptions for $\hat{\theta}$. Then, under $F_n(\mu_n)$, the difference between the first-order asymptotic biases of $\left(\hat{\theta}^j - \theta_0\right)$ and $\left(\hat{\theta} - \theta_0\right)$ is $\Lambda_{\cdot j}\mathbb{E}\left(\tilde{g}_j\right)$, for $\Lambda_{\cdot j}$ the j^{th} column of Λ.*

Proof. Applying proposition 1, under $F_n(\mu_n)$, $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ converges in distribution to a random variable with mean $\Lambda_{\cdot j}\mathbb{E}\left(\tilde{g}_j\right)$, and $\sqrt{n}\left(\hat{\theta}^j - \theta_0\right)$ converges in distribution to a random variable with mean zero. □

Dropping moments does not yield an analogue of Λ. Rather, when a given moment j is suspect (and the other moments are not), re-estimating after dropping the moment gives an asymptotically unbiased estimate of $\Lambda_{\cdot j}\mathbb{E}\left(\tilde{g}_j\right)$, the product of the sensitivity of the original estimator to moment j and the degree of misspecification of moment j.

Dropping moments need not be informative about what moments "drive" a parameter in the sense that changing the realized value of the moment would affect the realized estimate. Consider, for example, an over-identified model for which all elements of $\hat{g}\left(\hat{\theta}\right)$ happen to be exactly zero. Then dropping any particular moment leaves the parameter estimate unchanged, but changing its realized value will affect the parameter estimate so long as its sample sensitivity is not zero.

A.2 Effect of Parameters on Moments

Another common method for assessing the importance of moments is to ask (say, via simulation) how the population values of the moments change when we vary a particular parameter of interest (see, e.g., Goettler and Gordon 2011; Kaplan 2012; Berger and Vavra 2015; and Morten 2016).

This approach yields an estimate of minus one times a right inverse of our sensitivity measure. The large-sample effect of a small change in the parameters θ on the moments is given by G. Recalling that $\Lambda = -\left(G'WG\right)^{-1}G'W$, we have $-\Lambda G = I$, so that Λ is a left inverse of $-G$. When G is square, $\Lambda = (-G)^{-1}$. When $\hat{\theta}$ is a CMD estimator, and $\hat{g}(\theta) = \hat{s} - s(\theta)$, we have

$-G = \frac{\partial}{\partial \theta} s(\theta_0)$, so Λ is minus one times a left inverse of the matrix we obtain by perturbing the parameters and looking at the resulting changes in the model's predictions $s(\theta)$.

The matrix G is not a measure of the sensitivity of an estimator to misspecification. Indeed, G is not a property of the estimator at all, but rather a (local) property of the model. A moment can respond to a change in the value of a parameter even if that moment plays no role in estimation at all. This is true, for example, for an over-identified MDE in which we set the elements of $\hat{W}$ corresponding to a particular moment equal to zero.

B Proofs for Results in Main Text

B.1 Proof of Proposition 1

Because $\theta_0 \in interior(\Theta)$ and $\hat{g}(\theta)$ is continuously differentiable in θ, the following first-order condition must be satisfied with probability approaching one as $n \to \infty$:

$$\hat{G}\left(\hat{\theta}\right)' \hat{W} \hat{g}\left(\hat{\theta}\right) = 0.$$

By the mean value theorem,

$$\hat{g}\left(\hat{\theta}\right) = \hat{g}(\theta_0) + \hat{G}\left(\overline{\theta}\right)\left(\hat{\theta} - \theta_0\right),$$

for some $\overline{\theta} \in \left(\theta_0, \hat{\theta}\right)$ which may vary across rows. Substituting this expression into the first-order condition yields

$$\hat{G}\left(\hat{\theta}\right)' \hat{W} \hat{g}(\theta_0) + \hat{G}\left(\hat{\theta}\right)' \hat{W} \hat{G}\left(\overline{\theta}\right)\left(\hat{\theta} - \theta_0\right) = 0.$$

Rearranging, we have

$$\left(\hat{\theta} - \theta_0\right) = \hat{L}\hat{g}(\theta_0),$$

where $\hat{L} = -\left(\hat{G}\left(\hat{\theta}\right)' \hat{W} \hat{G}\left(\overline{\theta}\right)\right)^{-1} \hat{G}\left(\hat{\theta}\right)' \hat{W}$.

We know that $\hat{\theta} \xrightarrow{p} \theta_0$ under $F_n(\mu_n)$, so $\overline{\theta} \xrightarrow{p} \theta_0$. This plus uniform convergence of $\hat{G}(\theta)$ to $G(\theta)$ implies that under $F_n(\mu_n)$, $\hat{G}\left(\hat{\theta}\right)$ and $\hat{G}\left(\overline{\theta}\right)$ both converge in probability to G. Recalling that $\Lambda = -\left(G'WG\right)^{-1} G'W$, the above, along with $\hat{W} \xrightarrow{p} W$, implies $\hat{L} \xrightarrow{p} \Lambda$.

Then

$$\begin{aligned}
\sqrt{n}\left[\left(\hat{\theta} - \theta_0\right) - \Lambda \hat{g}(\theta_0)\right] &= \sqrt{n}\left[\hat{L}\hat{g}(\theta_0) - \Lambda \hat{g}(\theta_0)\right] \\
&= \left(\hat{L} - \Lambda\right)\sqrt{n}\hat{g}(\theta_0),
\end{aligned}$$

which converges in probability to zero by Slutsky's theorem (using the fact that $\sqrt{n}\hat{g}(\theta_0)$ con-

verges in distribution). Therefore, under $F_n(\mu_n)$, $\sqrt{n}\left(\hat{\theta}-\theta_0, \Lambda\hat{g}(\theta_0)\right)$ converges in distribution to a random vector $\left(\tilde{\theta}, \Lambda\tilde{g}\right)$ with $\Pr\left\{\tilde{\theta}=\Lambda\tilde{g}\right\}=1$. This implies in particular that $\mathbb{E}\left(\tilde{\theta}\right)=\Lambda\mathbb{E}(\tilde{g})$.

B.2 Proof of Proposition 2

We begin by stating and proving an additional lemma, from which proposition 2 then follows.

Lemma 1. *Consider a sequence $\{\mu_n\}_{n=1}^{\infty}$. Suppose that under $F_n(\mu_n)$*

$$\hat{g}(\theta)=\hat{a}(\theta)+\hat{b},$$

where the distribution of $\hat{a}(\theta)$ is the same under $F_n(0)$ and $F_n(\mu_n)$ for every n, and $\sqrt{n}\hat{b}$ converges in probability. Also, $\hat{W}\xrightarrow{p}W$ under $F_n(\mu_n)$.[18] *Then $\{\mu_n\}_{n=1}^{\infty}$ is a local perturbation.*

Proof. Uniform convergence of $\hat{G}(\theta)$ to $G(\theta)$ in probability under $F_n(\mu_n)$ follows from the fact that $\hat{b}$ does not depend on θ and that the distribution of $\hat{a}(\theta)$ is unaffected by μ. Convergence in distribution of $\sqrt{n}\hat{g}(\theta_0)$ follows from the fact that $\sqrt{n}\hat{a}(\theta_0)$ converges in distribution and $\sqrt{n}\hat{b}$ converges in probability. That $\hat{\theta}\xrightarrow{p}\theta_0$ then follows from the observation that $\hat{g}(\theta)'\hat{W}\hat{g}(\theta)$ converges uniformly in probability to $g(\theta)'Wg(\theta)$. □

Turning now to proposition 1, that $\{\mu_n\}_{n=1}^{\infty}$ is a local perturbation follows from lemma 1 with $\hat{a}(\theta)=\tilde{s}-s(\theta)$ and $\hat{b}=\mu_n\hat{\eta}$. The expression for $\mathbb{E}\left(\tilde{\theta}\right)$ then follows by proposition 1.

B.3 Proof of Proposition 3

To prove this result, we again state and prove an additional lemma, which then implies the proposition.

Lemma 2. *Consider a sequence $\{\mu_n\}_{n=1}^{\infty}$ with $\mu_n=\frac{\mu^*}{\sqrt{n}}$ for a constant μ^*. Suppose that assumption 1 holds, and that under $F_n(\mu)$ we have $\hat{\zeta}_i(\theta_0)=\tilde{\zeta}_i+\mu V_i$, where the distribution of $\left(\tilde{\zeta}_i, X_i, V_i\right)$ does not depend on μ. Then $\{\mu_n\}_{n=1}^{\infty}$ is a local perturbation.*

Proof. By assumption 1 part (ii) we know that (ζ_i, X_i, V_i) has density $f(\zeta_i, X_i, V_i)$ with respect to ν under $F(0)$. Thus, the density $f(\zeta_i, X_i, V_i|\mu)$ is given by $f(\zeta_i-\mu V_i, X_i, V_i)$. By assumption 1 part

[18] This is true in particular if $\hat{W}$ either does not depend on the data or is equal to $w\left(\hat{\theta}^{FS}\right)$, where $w(\cdot)$ is a continuous function and $\hat{\theta}^{FS}$ is a first-stage estimator that solves (1) for $\hat{W}$ equal to a positive semi-definite matrix W^{FS} not dependent on the data. In the latter case, the fact that $\hat{g}(\theta)'W^{FS}\hat{g}(\theta)$ converges uniformly to $g(\theta)'W^{FS}g(\theta)$ implies that we have $\hat{\theta}^{FS}\xrightarrow{p}\theta_0$ by theorem 2.1 of Newey and McFadden (1994). Thus, $\hat{W}\xrightarrow{p}W$ by the continuous mapping theorem.

(iii), $\sqrt{f(\zeta_i - \mu V_i, X_i, V_i)}$ is continuously differentiable in ζ_i, which implies that

$$\frac{\partial}{\partial \mu}\sqrt{f(\zeta_i - \mu V_i, X_i, V_i)} = -\frac{1}{2}\frac{V_i' \frac{\partial}{\partial \zeta_i} f(\zeta_i - \mu V_i, X_i, V_i)}{\sqrt{f(\zeta_i - \mu V_i, X_i, V_i)}}$$

is continuous in μ for all $(\zeta_i - \mu \cdot V_i, X_i, V_i)$. By assumption 1 part (iv) we know that

$$0 < \int \left(\frac{V_i' \frac{\partial}{\partial \zeta_i} f(\zeta_i, X_i, V_i)}{f(\zeta_i, X_i, V_i)} \right)^2 f(\zeta_i, X_i, V_i)\, d\nu < \infty,$$

but using the linear structure of the model we see that this is equal to the information matrix for μ

$$I_\mu = \int \left(\frac{V_i' \frac{\partial}{\partial \zeta_i} f(\zeta_i - \mu V_i, X_i, V_i)}{f(\zeta_i - \mu V_i, X_i, V_i)} \right)^2 f(\zeta_i - \mu V_i, X_i, V_i)\, d\nu,$$

for all μ. Thus, the information matrix for estimating μ is continuous in μ, finite, and non-zero.

Given these facts, lemma 7.6 of van der Vaart (1998) implies that the family of distributions $F(\mu)$ is differentiable in quadratic mean in a neighborhood of zero. Thus, if we take $\mu_n = \frac{\mu^*}{\sqrt{n}}$, then by theorem 7.2 of van der Vaart (1998) we have that under $F_n(0)$,

$$\log \frac{dF_n(\mu_n)}{dF_n(0)} = \frac{1}{\sqrt{n}} \sum_i \mu^* \frac{V_i' \frac{\partial}{\partial \zeta_i} f(\zeta_i, X_i, V_i)}{f(\zeta_i, X_i, V_i)} - \frac{1}{2}(\mu^*)^2 I_\mu + o_p(1).$$

Moreover, the Cauchy-Schwarz inequality, assumption 1 parts (iv) and (v), and the central limit theorem imply that under $F_n(0)$,

$$\begin{pmatrix} \sqrt{n}\hat{g}(\theta_0) \\ \log \frac{dF_n(\mu_n)}{dF_n(0)} \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ -\frac{1}{2}(\mu^*)^2 I_\mu \end{pmatrix}, \begin{pmatrix} \Omega & \mu^* \Xi \\ \mu^* \Xi & (\mu^*)^2 I_\mu \end{pmatrix} \right),$$

for Ξ the asymptotic covariance of $\sqrt{n}\hat{g}(\theta_0)$ and $\frac{1}{\sqrt{n}}\sum_i \left(V_i' \frac{\partial}{\partial \zeta_i} f(\zeta_i, X_i, V_i) \right) / f(\zeta_i, X_i, V_i)$. However, by LeCam's first lemma (lemma 6.4 in van der Vaart 1998), this implies that the sequences $F_n(0)$ and $F_n(\mu_n)$ are contiguous. Moreover, by LeCam's third lemma (example 6.7 of van der Vaart 1998),

$$\sqrt{n}\hat{g}(\theta_0) \xrightarrow{d} N(\mu^* \Xi, \Omega)$$

under $F_n(\mu_n)$. Furthermore, contiguity immediately implies that the other conditions for a local perturbation are satisfied, since any object which converges in probability under $F_n(0)$ must, by the definition of contiguity, converge in probability to the same limit under $F_n(\mu_n)$. □

Returning to proposition 3, that $\{\mu_n\}_{n=1}^{\infty}$ is a local perturbation follows from lemma 2. The expression for $\mathbb{E}\left(\tilde{\theta}\right)$ then follows by proposition 1.

B.4 Proof of Proposition 4

Proof. Since μ_n is a local perturbation, $\hat{\theta} \xrightarrow{p} \theta_0$ under $F_n(\mu_n)$. Thus, since we have assumed that $\hat{g}(\theta)$ and $\hat{G}(\theta)$ converge uniformly to limits $g(\theta)$ and $G(\theta)$,

$$\left(\hat{g}\left(\hat{\theta}\right), \hat{G}\left(\hat{\theta}\right), \hat{W}\right) \xrightarrow{p} \left(g(\theta_0), G(\theta_0), W\right).$$

However, we have also assumed that $\sup_{\theta \in \mathscr{B}_\theta} \| \frac{\partial}{\partial \theta_p} \hat{G}(\theta) \|$ is asymptotically bounded for $1 \leq p \leq P$, which, by the consistency of $\hat{\theta}$, implies that $\left\{ \frac{\partial}{\partial \theta_p} \hat{G}\left(\hat{\theta}\right) \right\}_{p=1}^{P}$ is asymptotically bounded as well. Thus, since $g(\theta_0) = 0$, we see that $\hat{A} \xrightarrow{p} 0$. Finally, since we have assumed that $G'WG$ has full rank, the continuous mapping theorem implies that $\hat{\Lambda}_S \xrightarrow{p} \Lambda$. □