NBER WORKING PAPER SERIES

MEASURING THE SENSITIVITY OF PARAMETER ESTIMATES TO ESTIMATION MOMENTS

Isaiah Andrews
Matthew Gentzkow
Jesse M. Shapiro

Working Paper 20673
http://www.nber.org/papers/w20673

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge MA 02138
November 2014

Previously circulated as "Measuring the Sensitivity of Parameter Estimates to Sample Statistics." 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, Toronto, and Yale for advice and suggestions, and to our dedicated research assistants for important contributions to this project. We thank the following 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; and Amil Petrin. This research was 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, and the National Science Foundation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Measuring the Sensitivity of Parameter Estimates to Estimation Moments
Isaiah Andrews, Matthew Gentzkow, and Jesse M. Shapiro
NBER Working Paper No. 20673
November 2014, Revised November 2016
JEL No. C1,C52

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 it can make the mapping from data to conclusions in structural estimation more transparent, and we show that it can be interpreted formally as a measure of sensitivity to violations of identifying assumptions. We illustrate with applications to published articles in several fields of economics.

Isaiah Andrews
Department of Economics
MIT
Building E52, Office 504
50 Memorial Drive
Cambridge MA 02139
and NBER
iandrews@mit.edu

Matthew Gentzkow
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305
and NBER
gentzkow@stanford.edu

Jesse M. Shapiro
Economics Department
Box B
Brown University
Providence, RI 02912
and NBER
jesse_shapiro_1@brown.edu

An online appendix is available at http://www.nber.org/data-appendix/w20673

1 Introduction

A recent conversation in economics about the relative merits of "structural" and "reduced-form" methods centers, in part, on the perception that "the often complex computational methods that are required to implement [structural estimation] make it less transparent" (Heckman 2010, p. 358), and that "it's hard to see precisely which features of the data drive the ultimate results" (Angrist and Pischke 2010, p. 21). Perhaps in response to this critique, it has become increasingly common for articles using structural methods to describe informally how the results relate to key moments of the data.

While these discussions are commonly framed in terms of "identification," Keane (2010) notes that they often do not speak to issues of identification in the formal sense.[1] 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: claims that a moment is the "main" or "primary" source of identification are hard to interpret.[2] It is a property of a model, not the way it is estimated: saying that a parameter is identified by a moment need not mean that changing that moment would affect a particular estimate, or even that the moment must be involved in estimation at all. Authors often acknowledge the imprecision of their statements by saying they discuss identification "loosely," "casually," or "heuristically."[3]

In this paper, we suggest a different way to look at the the relationship between an estimator and the moments of the data it depends on. We introduce a measure that describes the sensitivity of an estimator to small perturbations of individual moments, holding other relevant moments constant. We show that this measure can be computed at low cost even for complex models. We argue that it can shed light on which features of the data drive key conclusions, and so make structural

[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?" (p. 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" (p. 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]." DellaVigna et al. (2012) write: "Though the parameters are estimated jointly, it is possible to address the *main sources of identification* of individual parameters." (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..." Crawford and Yurukoglu (2012) write, "One may *casually* think of [a set of moments] as 'empirically identifying' [a set of parameters]." 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." (Emphasis added.)

estimation more transparent. We also show that it can be interpreted as a measure of sensitivity to violations of specific identifying assumptions, providing a way to translate heuristic discussion into precise statements about the credibility of a researcher's conclusions in the face of model misspecification.

Our analysis takes as given that a model is identified and focuses on 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. We think it is closer to the concept that some authors have in mind when discussing identification informally.

We consider the following setting. A sample of size n is drawn from a distribution $F(\cdot|\theta,\psi)$, where θ is a finite-dimensional parameter vector of interest with true value θ_0, and ψ is a possibly infinite-dimensional nuisance parameter with true value ψ_0. A researcher computes an estimator $\hat{\theta}$ that minimizes a criterion function $\hat{g}(\theta)'\hat{W}\hat{g}(\theta)$, where $\hat{g}(\theta)$ is a function of data and parameters and $\hat{W}$ is a data-dependent weight matrix. We assume standard regularity conditions such that $\hat{\theta}$ is consistent and asymptotically normal (as in Newey and McFadden 1994). This class of minimum distance estimators (MDEs) includes generalized method of moments (GMM), classical minimum distance (CMD), maximum likelihood (MLE), and their simulation-based analogues, and so encompasses most of the workhorse methods of structural point estimation. We focus on characterizing the drivers of $\hat{\theta}$, but we show that our approach extends naturally to the case where the outcome of interest is a function of $\hat{\theta}$, such as a counterfactual experiment or welfare calculation.

Our goal is to characterize the sensitivity of $\hat{\theta}$ to the estimation moments $\hat{g}(\cdot)$. We consider two ways to pose this question. One starts from the observed realization of the data and asks how $\hat{\theta}$ would change with small perturbations of the realized moment functions $\hat{g}(\cdot)$. The other asks how the population value of $\hat{\theta}$ would change if we perturb the population values of the moments $g(\theta) = \text{plim}\,\hat{g}(\theta)$. We show that both questions have the same answer in large samples. Our measure of sensitivity Λ is this limiting value. It can be written as $\Lambda = -\left(G'WG\right)^{-1}G'W$, where G is the Jacobian of $g(\cdot)$ 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.

We show that Λ can be interpreted formally as a measure of sensitivity to local model misspecification. The standard econometric assumption that $\hat{g}(\theta_0)$ converges to zero encodes a set of economic assumptions, such as exclusion restrictions or optimality conditions. Perturbing the population value of the moments is equivalent to allowing small violations of these assumptions. Our measure captures the effect of such violations on the estimates of interest. In the case of CMD or indirect inference, where the moments $\hat{g}(\theta) = \hat{s} - s(\theta)$ are differences between predicted quantities $s(\theta)$ and their empirical analogues $\hat{s}$, we show that it can be used to assess the bias resulting from either misspecification of $s(\theta)$ or mismeasurement of $\hat{s}$. In the case of nonlinear instrumental variables (IV), where the moments $\hat{g}(\theta) = Z'\hat{\zeta}(\theta)$ are the product of excluded instruments Z and structural error terms $\hat{\zeta}(\theta)$, we show that it can be used to assess the effect of omitted variables that introduce correlation between the errors and the instruments.

This latter result delivers a local analogue of the usual omitted variables bias formula that applies to nonlinear IV settings such as the demand model of Berry et al. (hereafter "BLP," 1995). It generalizes the findings of Conley et al. (2012) on the effect of local misspecification in a linear IV setting.

We argue that sensitivity can be a valuable addition to the applied economist's toolkit. Most simply, it provides a low-cost way to answer Angrist and Pischke's (2010) question—"which features of the data drive the ultimate results?"—in the context of computationally challenging models. Sensitivity is the limit of the partial derivative of the realized estimate with respect to the realized moments $\hat{g}\left(\hat{\theta}\right)$ in a particular sample. A moment "drives" an estimate in this sense to the extent that small changes in the moment lead to substantially different results.

Sensitivity also provides a formal rationale for why we want to answer Angrist and Pischke's (2010) question in the first place. If the statistical model is correct and the estimator is efficient, it is not obvious why we would care how the estimator translates data into conclusions. This knowledge becomes essential, however, once we entertain the possibility of misspecification. An estimator that depends on changes in individuals' behavior over time will be valid under different assumptions than an estimator that compares different individuals at a point in time. An estimator that depends only on values of an outcome local to a discontinuity may be valid under more credible assumptions than one that depends on data far from the discontinuity (Gelman and Imbens 2014). Sensitivity is the effect of changing the population moments on the limiting value of the estimator. It therefore tells us how the estimator will behave if population moments do not behave exactly as the researcher has assumed.

We illustrate the utility of our approach with three applications. The first is to BLP's (1995) model of automobile demand and pricing. The moments $\hat{g}(\theta)$ used to estimate the model are products of vehicle characteristics—used as excluded instruments—with demand- and supply-side errors. The moment conditions encode the economic assumption that a particular vehicle characteristic is orthogonal to unobserved variation in either utility or marginal cost. These assumptions could be violated if vehicle characteristics respond to unobserved demand shocks (as in Fan 2013 and Wollmann 2016) or if economies of scope lead marginal costs to depend on the lines of vehicles produced (as suggested by Levitt et al. 2013).

We apply our measure to assess sensitivity to such violations. We focus on one of the model's key outputs, the estimated average markup. This quantity is of direct interest as a measure of market power, and it is one of the features of the estimates that informs evaluations of policies such as trade restrictions (BLP 1999), mergers (Nevo 2000), and the introduction of new goods (Petrin 2002). We find that the sensitivity of the markup to the supply moments is larger than the sensitivity to the demand moments on average, with a single moment—the product of a car's marginal cost shock with the number of vehicle models produced by the same firm—playing a particularly important role. This suggests that economies of scope may be among the most significant threats to validity. We apply our misspecification results to evaluate quantitatively the bias that would be introduced by such economies of scope, as well as by correlation between demand errors and the composition of product lines. Both violations cause meaningful bias.

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 apply our measure to assess the sensitivity of the conclusions 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 substantially affect the results. For example, we show that varying shopping intensity as in Aguiar and Hurst (2007) would mean that the estimates substantially 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 DellaVigna et al.'s (2012) model of charitable giving. The authors use a field experiment in conjunction with with a structural model to distinguish 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 as a result. We apply our measure to assess the sensitivity of these conclusions to a key functional form assumption that causes households who are very sensitive to social pressure to bunch their donations at \$10. We show that even very small perturbations away from this assumption can change the estimated social pressure, but that the implied bias is small, and adjusting for it would tend to strengthen the qualitative conclusions.

In appendix A, we discuss two alternative approaches that have appeared in the literature. One is to ask how the parameter estimates change when the 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. Neither of these alternative methods delivers a measure of sensitivity to misspecification, and neither is a reliable guide to which moments "drive" a parameter in the sense that changing the realized value of the moment would change the realized estimate.

We emphasize two limitations to our approach. The first is that we consider sensitivity to local perturbations only. The sensitivity of the estimator to perturbations over a larger range could in principle look very different.

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 we estimate for specific forms of misspecification, but it does matter for qualitative conclusions about the relative importance of different moments. The problem is analogous to the classic problem of comparing the magnitude of regression coefficients, and it has no perfect solution. In the paper, we propose a default normalization that is analogous to standardizing a regressor to have a standard deviation of unity.

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 moment estimators under local misspecification (for example Newey 1985; Berkowitz et al. 2008; Guggenberger 2012; Conley et al. 2012; Nevo and Rosen 2012; Kitamura et al. 2013). Our results also relate the literature on sensitivity analysis (including Leamer 1983; Sobol 1993; Saltelli et al. 2008; Chen et al. 2015), though our focus is on local, rather than global, deviations from the assumed model.

Relative to the existing literature on local misspecification, our main contribution is the proposal to report the sensitivity matrix alongside structural estimates, both to complement heuristic discussions of sensitivity and to allow readers to assess the impact of interesting forms of misspecification. 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 the estimation moments, and to show that this derivative's limiting value is the sensitivity matrix.

The remainder of the paper is organized as follows. Section 2 defines sensitivity and characterizes its properties. Section 3 derives results for the special cases of CMD and IV. Section 4 considers estimation. Section 5 presents our applications and Section 6 concludes. Appendix A discusses some common alternatives, and the online appendix extends our main results along several dimensions.

2 Measure

We have observations $D_i \in \mathscr{D}$ for $i = 1, ..., n$, which comprise a sample $D \in \mathscr{D}^n$. A model implies that D_i follows $F(\cdot|\theta, \psi)$, where θ is a finite-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 $\hat{G}(\theta)$. Under F_n: (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 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).

Our measure of sensitivity is as follows.

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

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

We consider two interpretations of the sensitivity of the estimator $\hat{\theta}$ to small perturbations of the moments. One is the effect of perturbing the realized moment functions $\hat{g}(\cdot)$ on the realization of $\hat{\theta}$ in a particular sample. The other is the effect of perturbing the population values of the moments $g(\theta_0)$ on the asymptotic behavior of $\hat{\theta}$. We show that Λ is the limiting value of the answer to the first question, and that Λ is the answer to the second question posed either in terms of small fixed perturbations to $g(\theta_0)$ or perturbations that grow small with n.

Consider, first, how the estimate $\hat{\theta}$ in a particular sample changes when we perturb the realized moment functions $\hat{g}(\theta)$. Define a family of perturbed moments $\hat{g}(\theta, \delta)$, where the scalar δ controls the degree of perturbation and $\hat{g}(\theta, 0) = \hat{g}(\theta)$. Define the resulting estimator $\hat{\theta}(\delta)$ to solve

$$\min_{\theta \in \Theta} \hat{g}(\theta, \delta)' \hat{W} \hat{g}(\theta, \delta). \tag{2}$$

We assume that $\hat{g}(\theta,\delta)$ and $\hat{G}(\theta,\delta)=\frac{\partial}{\partial\theta}\hat{g}(\theta,\delta)$ are continuously differentiable in (θ,δ) on $\Theta\times\mathscr{B}_\delta$, for some ball $\mathscr{B}_\delta$ around zero. This is true, for example, if $\hat{g}(\theta,\delta)=\hat{g}(\theta)+\delta\cdot\eta$ for some vector η.

Define the *sample sensitivity* of $\hat{\theta}$ to $\hat{g}(\cdot)$ to be

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

where

$$\hat{A}=\left[\begin{array}{ccc}\hat{G}_1\left(\hat{\theta}\right)'\hat{W}\hat{g}\left(\hat{\theta}\right) & \ldots & \hat{G}_P\left(\hat{\theta}\right)'\hat{W}\hat{g}\left(\hat{\theta}\right)\end{array}\right],$$

and $\hat{G}_p\left(\hat{\theta}\right)$ is the partial derivative of $\hat{G}\left(\hat{\theta}\right)$ with respect to the p-th element of θ.

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

$$\frac{\partial}{\partial\delta}\hat{\theta}(\delta)=\hat{\Lambda}_S\frac{\partial}{\partial\delta}\hat{g}\left(\hat{\theta},0\right)$$

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.)

To relate $\hat{\Lambda}_S$ to Λ, we make an additional technical assumption.

Assumption 1. *For $1\leq p\leq P$ and $\mathscr{B}_\theta$ a ball around θ_0, $\sup_{\theta\in\mathscr{B}_\theta}\|\hat{G}_p(\theta)\|$ is asymptotically bounded.*[4]

This condition is satisfied if, for example, $\hat{G}_p(\theta)$ converges to a continuous function $G_p(\theta)$ uniformly on $\mathscr{B}_\theta$. Assumption 1 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 1. *Under assumption 1, $\hat{\Lambda}_S\xrightarrow{p}\Lambda$ under F_n as $n\to\infty$.*

Proof. See appendix. □

Consider, next, what happens to the asymptotic distribution of $\hat{\theta}$ as we perturb the population value of the moments $g(\theta_0)$. Define a family of perturbed distributions $F(\cdot|\theta,\psi,\mu)$, where μ controls the degree of perturbation and $F(\cdot|\theta,\psi,0)=F(\cdot|\theta,\psi)$. Let $F_n(\mu)\equiv\{\times_n F(\cdot|\theta_0,\psi_0,\mu)\}_n$. When $\mu\neq 0$, the model is misspecified in the sense that under $F_n(\mu)$, $\hat{g}(\theta_0)\stackrel{p}{\nrightarrow}0$. Proposition 2 below shows that Λ relates changes in the population values of the moments resulting from such misspecification to changes in the limiting value of the estimator.

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

Assumption 2. *For a ball $\mathscr{B}_\mu$ around zero, we have that under $F_n(\mu)$ for any $\mu \in \mathscr{B}_\mu$, (i) $\hat{g}(\theta)$ and $\hat{G}(\theta)$ converge uniformly in θ to functions $g(\theta,\mu)$ and $G(\theta,\mu)$ that are continuously differentiable in (θ,μ) on $\Theta \times \mathscr{B}_\mu$, and (ii) $\hat{W} \xrightarrow{p} W(\mu)$ for $W(\mu)$ continuously differentiable on $\mathscr{B}_\mu$.*

Proposition 2. *Under assumption 2, there exists a ball $\mathscr{B}_\mu^* \subset \mathscr{B}_\mu$ around zero such that for any $\mu \in \mathscr{B}_\mu^*$, $\hat{\theta}$ converges in probability under $F_n(\mu)$ to a continuously differentiable function $\theta(\mu)$, and*

$$\frac{\partial}{\partial \mu}\theta(0) = \Lambda \frac{\partial}{\partial \mu} g(\theta_0, 0).$$

Proof. See appendix. □

Under the fixed levels of misspecification μ contemplated in proposition 2, many conventional asymptotic approximations—in particular the usual standard errors—are no longer valid. For this reason, the literature on misspecification (e.g., Newey 1985; Conley et al. 2012) often allows the perturbation to shrink with the sample size. Under appropriate rate conditions, the asymptotic effect of the perturbation is on the same order as sampling uncertainty, and it is often possible to characterize the bias in the asymptotic distribution of the estimator and adjust inference for hypothesized misspecification of the data generating process.

We will 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 bias in the asymptotic distribution of the estimator.

Proposition 3. *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 $\mathrm{E}\left(\tilde{\theta}\right)$ is given by

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

Proof. See appendix. □

Propositions 1, 2, and 3 are our main formal results. Together they show that Λ describes the relationship between the parameter estimates and the moments in several distinct senses. It is the limit of the partial derivative of the realized estimate with respect to the realized moments in a particular sample. In this sense, it provides a direct answer to the question "which features of the data drive the results." A moment "drives" a parameter estimate to the extent that changing the moment would affect the estimate holding the value of other moments constant.

Sensitivity also captures the effect of a perturbation of the population moments on the asymptotic behavior of the estimator, whether we think of a small, fixed perturbation $\mu \neq 0$ affecting the probability limits, or a shrinking perturbation μ_n affecting the asymptotic distribution. In this sense, it can be interpreted as a measure of sensitivity to model misspecification. The tight relationship between sample sensitivity and sensitivity to misspecification implicit in our main results provides a reason to be interested in knowing which features of the data drive the results.

Several extensions of our results 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 $c\left(\hat{\theta}\right)$ is a continuously differentiable function not dependent on the data, with non-zero gradient $C = C\left(\theta_0\right) = \frac{\partial}{\partial\theta}c\left(\theta_0\right)$ at θ_0. Then under any local perturbation, the delta method implies that $\mathrm{E}\left(c\left(\tilde{\theta}\right)\right) = C\Lambda\mathrm{E}\left(\tilde{g}\right)$. We will refer to $C\Lambda$ as the sensitivity of $c\left(\hat{\theta}\right)$.[5]

Remark 2. We may be interested in the sensitivity of some elements of the parameter vector holding other elements constant. Decomposing θ into subvectors $\left(\theta_1, \theta_2\right)$, 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 first-order asymptotic bias of $\hat{\theta}_1$ under local perturbations when $\hat{\theta}_2$ is held fixed at $\theta_{2,0}$. We can calculate the conditional version of sample sensitivity in an analogous manner.

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.[6] 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. The online appendix also shows how to generalize our notion of sample sensitivity to allow perturbations that affect the weight matrix.

We conclude this section with two simple examples that illustrate the computation of Λ and the application of these results. We focus for simplicity on the application of proposition 3, though the

[5]Analogously, by the chain rule the sample sensitivity of $c\left(\hat{\theta}\right)$ is $C\left(\hat{\theta}\right)\hat{\Lambda}_S$.

[6]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).

analogous results for proposition 2 are transparent.[7] Here and in what follows we let Ω_{AB} denote $\operatorname{plim}\left[\frac{1}{n}A'B\right]$ for matrices A and B.

Example 1. (OLS) Suppose the data are $D = [Y, X]$ and under the assumed model F_n

$$Y = X\theta_0 + \varepsilon,$$

where $\mathrm{E}(\varepsilon|X) = 0$. The OLS regression of Y on X can be written as a GMM estimator with $\hat{g}(\theta) = \frac{1}{n}X'(Y - X\theta)$ and weight matrix $\hat{W} = I$. Sensitivity is $\Lambda = \Omega_{XX}^{-1}$. Because the model is linear, sample sensitivity $\hat{\Lambda}_S$ is just the sample analogue $\left(\frac{1}{n}X'X\right)^{-1}$.

Consider now a perturbed model $F_n(\mu)$ under which

$$Y = X\theta_0 + \mu V + \varepsilon,$$

where μ is a scalar and V is an (unobserved) omitted variable with $\Omega_{XV} \neq 0$. Applying proposition 3 with $\mu_n = \frac{\mu^*}{\sqrt{n}}$ for a fixed μ^*, we see that the resulting bias in the estimator is

$$\mathrm{E}\left(\tilde{\theta}\right) = \Omega_{XX}^{-1}\Omega_{XV}\mu^*.$$

This is of course the standard omitted variables bias formula—the product of the coefficients on the omitted variable (μ^*) and the coefficients from the regression of the omitted variable on the endogenous regressors ($\Omega_{XX}^{-1}\Omega_{XV}$). Through the lens of sensitivity, we see that a one unit change in the moments $\hat{g}(\theta_0)$ induces a bias of Ω_{XX}^{-1} in the estimator, and adding the omitted variable changes these moments by $\Omega_{XV}\mu^*$.

Example 2. (2SLS) Suppose the data are $D = [Y, X, Z]$ and the expression for Y under the assumed model is the same as in example 1 with $\mathrm{E}(\varepsilon|Z) = 0$ and $\mathrm{E}(\varepsilon|X) \neq 0$. The 2SLS estimator can be written as a GMM estimator with $\hat{g}(\theta) = \frac{1}{n}Z'(Y - X\theta)$ and $\hat{W} = \left(\frac{1}{n}Z'Z\right)^{-1}$. Sensitivity Λ in this case is $\Lambda = \left(\Omega_{ZX}'\Omega_{ZZ}^{-1}\Omega_{ZX}\right)^{-1}\Omega_{ZX}'\Omega_{ZZ}^{-1}$, and sample sensitivity $\hat{\Lambda}_S$ is the same expression replacing the population expectations with their sample analogues.

Conley et al. (2012) consider a perturbed model $F_n(\mu)$ in which ε is replaced by $\frac{1}{\sqrt{n}}Z\mu^* + \varepsilon$. Applying proposition 3, and noting that the asymptotic mean of $\sqrt{n}\hat{g}(\theta_0)$ under $F_n\left(\frac{\mu^*}{\sqrt{n}}\right)$ is $\mathrm{E}(\tilde{g}) = \Omega_{ZZ}\mu^*$, we see that the first-order asymptotic bias of the estimator is

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

This is the expression Conley et al. (2012) derive in section III.C. Note that if we consider the

[7] We do not verify that the conditions for a local perturbation are satisfied in each case, but this follows directly from the sufficient conditions given in lemma 1 below.

exactly identified case where $\hat{W} = I$ and replace $\frac{\mu^*}{\sqrt{n}}Z$ in the perturbation with a general omitted variable $\frac{\mu^*}{\sqrt{n}}V$, we obtain $\mathrm{E}\left(\tilde{\theta}\right) = \Omega_{ZX}^{-1}\Omega_{ZV}\mu^*$. We can think of this as a generalized omitted variables bias formula where the coefficient $\Omega_{XX}^{-1}\Omega_{XV}$ from the regression of V on X is replaced by the 2SLS coefficient $\Omega_{ZX}^{-1}\Omega_{ZV}$ from the regression of V on X instrumenting with Z.

3 Special Cases

Two special cases encompass the applications we present below and provide a template for many other cases of interest. Proposition 1 applies to both of these cases without further assumptions. In order to apply propositions 2 and 3, we must consider perturbations of the model that satisfy the relevant assumptions. We provide sufficient conditions for proposition 3 here, and sufficient conditions for proposition 2 in the online appendix.

The first case of interest is where $\hat{\theta}$ is a classical minimum distance estimator, with moments corresponding to differences between sample statistics $\hat{s}$ and corresponding predictions $s(\theta)$ under the model. Our definition of this case also includes estimation by indirect inference (Gourieroux et al. 1993; Smith 1993).

Definition. *$\hat{\theta}$ is a **classical minimum distance (CMD) estimator** if $\hat{g}(\theta) = \hat{s} - s(\theta)$, where $\mathrm{E}(\hat{s}) = s(\theta_0)$ and $s(\theta)$ 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 $s(\theta_0)$.

A natural category of perturbations to consider in this case is additive shifts of the moment functions due to either misspecification of $s(\theta)$ or measurement error in $\hat{s}$. The following lemma gives sufficient conditions for such shifts to be local perturbations as defined above.

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 $F_n(\mu_n)$.[8] Then $\{\mu_n\}_{n=1}^{\infty}$ is a local perturbation.

Proof. See appendix. □

[8] 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 is a continuous function and $\hat{\theta}^{FS}$ is a first-stage estimate 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.

Applying lemma 1 and proposition 3 yields a simple characterization of the first-order asymptotic bias of a misspecified CMD estimator.

Proposition 4. *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 $\mu_n = \frac{\mu^*}{\sqrt{n}}$, and suppose that $\hat{W} \xrightarrow{p} W$ under $F_n(\mu_n)$. Then taking $\mu_n = \frac{\mu^*}{\sqrt{n}}$, we have $\mathrm{E}\left(\tilde{\theta}\right) = \Lambda\eta\mu^*$.*

Proof. 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 $\mathrm{E}\left(\tilde{\theta}\right)$ then follows by proposition 3. □

The second case of interest is where $\hat{\theta}$ is estimated by non-linear instrumental variables, with moments formed by interacting the instruments with estimated structural errors.

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 $\mathrm{E}\left(\hat{\zeta}_i(\theta_0)|Z_i\right) = 0$ under F_n.*[9]

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 $\Omega_{Z\tilde{X}} = \mathrm{plim}\,\frac{1}{n} Z'\tilde{X}$ and $\tilde{X}$ are the "pseudo-regressors" $\partial\hat{\zeta}(\theta_0)/\partial\theta$. A natural perturbation to consider in this case is the introduction of an omitted variable V_i that causes the errors ζ to be correlated with the instruments Z. We provide sufficient conditions for this form of misspecification to be a local perturbation. These conditions apply more generally than nonlinear IV.

Assumption 3. *The observed data $D_i = [Y_i, X_i]$ consist of iid 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 v; (iii) $\sqrt{f(\zeta_i, X_i, V_i)}$ is continuously differentiable in ζ_i; (iv) we have*

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

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

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

[9] 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 $\hat{\zeta}_i(\theta)$.

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

The main substantive restriction imposed by assumption 3 is that the structural errors have full support and map one-to-one to the outcomes Y. 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.

Lemma 2. *Consider a sequence $\{\mu_n\}_{n=1}^{\infty}$ with $\mu_n = \frac{\mu^*}{\sqrt{n}}$ for a constant μ^*. Suppose that assumption 3 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. See appendix. □

Applying lemma 2 and proposition 3 allows us to characterize the effects of omitted variables in nonlinear IV estimators.

Proposition 5. *Suppose that $\hat{\theta}$ is an IV estimator satisfying assumption 3, 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 $\mu_n = \frac{\mu^*}{\sqrt{n}}$, we have $\mathrm{E}\left(\tilde{\theta}\right) = \Lambda\Omega_{ZV}\mu^*$.*

Proof. That $\{\mu_n\}_{n=1}^{\infty}$ is a local perturbation follows from lemma 2. The expression for $\mathrm{E}\left(\tilde{\theta}\right)$ then follows by proposition 3. □

We can think of proposition 5 as yielding a local analogue of the omitted variables bias formula for nonlinear models. Plugging in for the value of Λ, and focusing for simplicity on the exactly identified case where $W = I$, we can restate the conclusion of the proposition as

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

This has the same form as the omitted variables bias formula for 2SLS in example 2, except that the regressors X are replaced by the pseudo-regressors $\tilde{X}$.

4 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 6. *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, the result follows by the continuous mapping theorem. □

The online appendix develops an analogous result for the non-vanishing perturbations considered in proposition 2. It also extends proposition 1 to show that, under mild conditions, sample sensitivity $\hat{\Lambda}_S$ is consistent for Λ under any local perturbation $\{\mu_n\}_{n=1}^{\infty}$. In this sense, $\hat{\Lambda}_S$ can be interpreted as an alternative estimator of Λ.

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 one of our applications 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.

A final practical issue concerns the units of sensitivity. The elements of Λ have the same units as the coefficients from a regression of $\hat{\theta}$ on the realizations of $\hat{g}(\theta_0)$. Comparing the relative magnitudes of these coefficients is therefore only coherent if we can compare the different elements of $\hat{g}(\theta_0)$ in some meaningful way. This is an example of the age-old problem of assessing the relative importance of regressors in a regression model (Kim and Ferree 1981; Bring 1994; Gelman 2008). To establish a default, by analogy to the standardized regression coefficient we define the *standardized sensitivity* of $\hat{\theta}_p$ to be $\Lambda_{pj}\sqrt{\frac{\Omega_{jj}}{\Sigma_{pp}}}$ (recalling that Ω is the asymptotic variance of the moments and Σ is the asymptotic variance of the parameter estimate). We define standardized plug-in sensitivity and standardized sample sensitivity analogously. Standardized sensitivity is invariant to changing the units of $\hat{g}(\theta_0)$, say by converting an element from dollars to euros.[10] Whether this attribute is desirable or whether it is preferable to compare elements of Λ directly ultimately depends on the economic application.

5 Applications

5.1 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

[10]Formally, for strictly positive diagonal matrices A and B, the standardized sensitivity of $A\hat{\theta}$ to $B\hat{g}(\theta_0)$ with weight matrix $B^{-1}\hat{W}B^{-1}$ is equal to the standardized sensitivity of $\hat{\theta}$ to $\hat{g}(\theta_0)$ with weight matrix $\hat{W}$.

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 $Z = \left[\begin{array}{cc} Z_d & Z_s \end{array} \right]$ 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).[11]

Under the assumed model F_n

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

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

where ξ and ω are structural errors in utility and marginal cost, respectively, and $\mathrm{E}(\xi_i | Z_{di}) = \mathrm{E}(\omega_i | Z_{si}) = 0$. The functions $s(\cdot)$ and $p(\cdot)$ are known and invertible, so we can compute the errors $\hat{\xi}(\theta)$ and $\hat{\omega}(\theta)$ implied by given parameters and data. The estimator $\hat{\theta}$ solves (1) with moments

$$\hat{g}(\theta) = \frac{1}{n} \left[\begin{array}{c} \sum_i Z'_{di} \otimes \hat{\xi}_i(\theta) \\ \sum_i Z'_{si} \otimes \hat{\omega}_i(\theta) \end{array} \right].$$

The weight matrix is the inverse of the variance-covariance matrix of $\hat{g}\left(\hat{\hat{\theta}}\right)$ computed at first-stage estimates $\hat{\hat{\theta}}$.

The demand and supply moment conditions $\mathrm{E}(\xi_i | Z_{di}) = 0$ and $\mathrm{E}(\omega_i | Z_{si}) = 0$ encode distinct economic assumptions. The demand-side condition $\mathrm{E}(\xi_i | Z_{di}) = 0$ 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

[11]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.

determinants of ξ are unknown until after product line decisions are made. The assumption could be violated if ξ 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 with endogenous choice of product characteristics and find substantial responses of characteristics to consumer demand.

The supply-side condition $\mathrm{E}(\omega_i|Z_{si}) = 0$ 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.[12]

We apply our measure to assess the sensitivity of BLP's (1995) estimates to violations of the demand-side and supply-side assumptions respectively. We estimate the model using BLP's (1995) data and our own implementation of the authors' estimator.[13]

We consider a perturbed model $F_n(\mu)$ in which omitted variables lead to a correlation between the instruments and structural errors. Specifically, we assume that under $F_n(\mu)$,

$$\text{(3)} \qquad \begin{bmatrix} \hat{\xi}_i(\theta_0) \\ \hat{\omega}_i(\theta_0) \end{bmatrix} = \begin{bmatrix} \tilde{\xi}_i \\ \tilde{\omega}_i \end{bmatrix} + \mu \begin{bmatrix} V_{di} \\ V_{si} \end{bmatrix},$$

where $\begin{bmatrix} V_{di} & V_{si} \end{bmatrix}'$ is a vector of supply-side and demand-side omitted variables, the distribution of $\begin{bmatrix} \tilde{\xi}_i & \tilde{\omega}_i \end{bmatrix}'$ does not depend on μ, and $\mathrm{plim} \begin{bmatrix} \frac{1}{n}\sum_i Z_{di}' \otimes V_{di} \\ \frac{1}{n}\sum_i Z_{si}' \otimes V_{si} \end{bmatrix} = \Omega_{ZV}$. We consider the sequence of perturbations $\mu_n = \frac{1}{\sqrt{n}}$ and assume that the regularity conditions of assumption 3 are satisfied.

Sensitivity is

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

where the pseudo-regressors are $\tilde{X} = -\begin{bmatrix} \frac{\partial \xi(S,P,X;\theta_0)}{\partial \theta} & \frac{\partial \omega(S,P,X;\theta_0)}{\partial \theta} \end{bmatrix}$. By proposition 5, the first-order asymptotic bias in the estimator is $\mathrm{E}\left(\tilde{\theta}\right) = \Lambda \Omega_{ZV}$.

We focus on the sensitivity of the average markup, defined as the ratio of price minus marginal

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

[13]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. The online appendix reports the numerical values of standardized plug-in sensitivity for all parameter estimates.

cost to price, which is a key object of economic interest. Letting C denote the gradient of this markup at θ_0, we can apply remark 1 to obtain sensitivity $C\Lambda$.

Figure 1 plots the standardized plug-in sensitivity of the average markup. The left panel plots standardized sensitivity to the demand moments $\hat{g}_d(\theta)$; the right panel plots standardized sensitivity to the supply moments $\hat{g}_s(\theta)$. Sensitivity to supply moments is generally larger (in absolute value) than sensitivity to demand moments. The estimate is most sensitive to the product of unobserved marginal cost with the number of cars produced by the same firm as the car in question. This suggests qualitatively that firm-level economies of scope may be a particularly important threat to validity of the estimates. The online appendix presents analogous plots based on sample sensitivity and an alternative normalization of plug-in sensitivity.

Figure 2 shows for each parameter the mean absolute value of standardized sensitivity across all supply moments and all demand moments respectively. Across all parameters (including those from the demand side of the model), we find that the sensitivities to supply moments are larger, sometimes substantially so. Finding large sensitivities to the supply moments is consistent with a sense in the literature that the supply-side moments play a critical role in estimation. 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" (p. 875). In subsequent work, the authors recall finding that "estimates that used only the demand system were too imprecise to be useful" (BLP 2004, p. 92).

To translate these qualitative intuitions into quantitative statements, table 1 considers specific examples of omitted variables V. On the supply side, we assume 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 behavior (on the demand side). The table presents the first-order asymptotic bias implied by perturbation as well as bootstrap standard errors.

Table 1 shows that all of these omitted variables introduce meaningful first-order asymptotic bias in the estimated average markup. The first violation of the supply-side exclusion restriction, for example, would mean that the estimated markup of 0.33 is biased downward by 17 percentage points, implying a corrected estimate of 0.50. The violation of the demand-side exclusion restriction has an effect of similar magnitude, biasing the markup downward by 13 percentage points. We conclude that economically plausible violations of both supply and demand-side assumptions could lead to meaningful bias in the results.

5.2 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 the observed combination of a hump-shaped consumption profile with high marginal propensity to consume out of income shocks at young ages.

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

$$\hat{s}_a = s_a\left(\theta_0, \chi_0\right) + \varepsilon_a,$$

where $s_a\left(\theta, \chi\right)$ is the average log consumption predicted by the model and ε_a is a measurement error satisfying $\mathrm{E}\left(\varepsilon_a\right) = 0$ for all a. The estimator $\hat{\theta}$ solves (1) with moments

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

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 not dependent on the data.[14] This is then a CMD estimator as defined above.

The condition $\mathrm{E}\left(\hat{s}\right) = s\left(\theta_0, \chi_0\right)$ 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

[14] 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 $\hat{s}$.

(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 apply our measure to assess the way small violations of these key assumptions would affect the results. We estimate the model using the authors' original code and data.[15]

We focus on the sensitivity of the two key preference parameters—the discount factor and the coefficient of relative risk aversion—which in turn determine the relative importance of consumption smoothing and precautionary incentives.[16] 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 4, the first-order asymptotic bias is then $\mathrm{E}\left(\tilde{\theta}\right) = \Lambda\eta$.

Figure 3 presents standardized plug-in sensitivities for 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.

[15]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 2 below. This value is similar, though not identical, to the published parameters.

[16]We fix the two retirement parameters at their estimated values for the purposes of our analysis. The online appendix reports the numerical values of standardized sensitivity of the discount factor and the coefficient of relative risk aversion.

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.

Table 2 considers four specific perturbations. First, to allow for variable shopping intensity, we define the elements η_a to match the age-specific log price increments that Aguiar and Hurst (2007) estimate in their table I, column 1. Second, to allow for work-related consumption expenses, we define η_a so that true consumption at each age is overstated by 5 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_a 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_a so that consumption from age 50 through 65 is overstated by an annual amount whose lifetime sum is equal to the total gap between consumption and income over ages 26-29.

Table 2 shows the first-order asymptotic bias $\mathrm{E}\left(\tilde{\theta}\right)$ implied by each perturbation. The first row shows that if shopping intensity changes with age as in Aguiar and Hurst (2007), the estimated discount factor is over-stated by 0.4 percentage points and the estimated coefficient of relative risk aversion is under-stated by a third. The direction of these effects is consistent with what we would expect given figure 3: increasing the true consumption of older workers relative to what the expenditure data suggest tends to raise risk aversion and reduce discounting (though the effect after age 62 would work in the opposite direction). 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 biased in the opposite direction. This again matches the intuition from figure 3 since we are now reducing consumption for older workers (since they have more work-related expenses). 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. 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. Putting these results together, we conclude that reasonable alternatives imply meaningful first-order asymptotic bias in estimates of parameters of interest.

5.3 Charitable Giving

DellaVigna et al. (2012) use data from a field experiment to estimate a model of charitable giving. In the model, charitable giving may be motivated by altruism (caring about the aims of the charity or "warm glow") and social pressure. In the experiment, solicitors go door to door asking households to either donate to a charity or complete a survey. 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 the preference parameters and allow the authors to assess the welfare effects of solicitation. The main results are that social pressure is an important driver of giving and that the average visited household is made worse off by the solicitation.

From household-level data D, the authors construct a vector of statistics $\hat{s}$, including the share 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. The parameter vector θ includes determinants of the distribution of altruism and the social pressure cost of choosing not to give. The estimator $\hat{\theta}$ solves (1) with moments

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

where $s(\theta)$ is the predicted value of each statistic under the model, computed numerically by quadrature. The weight matrix $\hat{W}$ is equal to the diagonal of the variance-covariance matrix of the observed statistics $\hat{s}$. Under the assumed model F_n, $\mathrm{E}(\hat{s}) = s(\theta_0)$. This is a CMD estimator as defined above.

The model predictions $s(\theta_0)$ reflect a number of economic assumptions, including functional forms for the distribution of altruism, the utility function, and the social pressure cost as a function of the amount given. The authors present evidence that the results are not sensitive to alternatives to many of these assumptions. One assumption that they do not vary, however, is the form of the social pressure cost, assumed throughout to decrease linearly in the amount given up to a threshold $d^* = \$10$, after which there is no social pressure. The threshold value of \$10 is motivated as the sample median donation. The choice of this threshold is potentially significant because it means bunching of donations at \$10 will be taken to be evidence of social pressure.[17]

[17] DellaVigna et al. (2012) write: "The social pressure S^{ch} 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).

We apply our measure to assess how misspecification of the threshold affects estimates of the key social pressure parameter. We obtained the estimates of G and W needed to compute $\hat{\Lambda}$ directly from the authors.[18]

We consider perturbed models $F_n(\mu)$ under which fraction μ of households have a different threshold $d^* = d'$, while fraction $(1-\mu)$ of households use the assumed threshold $d^* = 10$. Let $s(\theta, d)$ be the expectation of $\hat{s}$ in a population with threshold d, so $s(\theta) = s(\theta, 10)$. Then under $F_n(\mu)$, $\hat{s} = \tilde{s} + \mu\hat{\eta}$, where the distribution of $\tilde{s}$ does not depend on μ and $\hat{\eta}$ converges in probability to the vector of constants $s(\theta_0, d') - s(\theta_0)$.[19] Taking $\mu_n = \frac{1}{\sqrt{n}}$ and noting that the probability limit of $\hat{W}$ is unchanged under $F_n(\mu_n)$, proposition 4 implies that the first-order asymptotic bias is $\mathrm{E}\left(\tilde{\theta}\right) = \Lambda\left[s(\theta, d') - s(\theta)\right]$.

Of those solicited in the experiment, some are asked to donate to the local La Rabida Children's Hospital and others are asked to donate to the East Carolina Hazard Center (ECU). The social preference parameter is allowed to be different for the two charities. We focus for simplicity on the estimates for ECU, and we assume that the perturbation to the threshold d^* only affects the utility function for giving to ECU. We present analogous results for the La Rabida social preference parameter in the online appendix.

Figure 4 presents the column of Λ corresponding to the social pressure cost S of giving \$0 to ECU. The estimated value of this parameter is \$1.44 with a standard error of \$0.78 (DellaVigna et al. 2012). We plot the sensitivities with respect to the elements of $\hat{g}(\theta_0)$ associated with the ECU treatments; sensitivities with respect to other elements of $\hat{g}(\theta_0)$ are shown in the online appendix. 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. The results line up well with the authors' expectations, reinterpreted as statements about sensitivity rather than identification. The share of people bunching at \$10 increases estimated social pressure. Donations of less than \$10 decrease it. The share of people opening the door in the flyer treatment also decreases it, 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, suggesting that misspecification of the threshold may be a significant concern.

Figure 5 shows the first-order bias associated with misspecification of the threshold at each possible value of d' ranging from \$0 to \$20. The value at $d' = \$10$ is zero by definition. The figure

[18] 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 population moment function $s(\theta)$ and confirm that our calculation of $s\left(\hat{\theta}\right)$ closely matches the published results.

[19] Define $\hat{\eta} = \frac{1}{\mu}(\hat{s} - \tilde{s})$. We have under $F_n(\mu)$: (i) $\mathrm{E}(\tilde{s}) = s(\theta_0)$ for all μ; (ii) $\mathrm{E}(\hat{s}) = \mu s(\theta_0, d') + (1-\mu)s(\theta_0)$; and so (iii) $\mathrm{E}(\hat{\eta}) = s(\theta_0, d') - s(\theta_0)$. Convergence in probability follows by the law of large numbers.

shows that that even a small change of d' away from this value implies that the estimated social pressure is biased downward, and that the sign of this bias is the same regardless of whether the threshold is higher or lower. This is intuitive: any $d' \neq \$10$ redistributes probability mass from giving exactly \$10 to giving either \$0-10 or \$10-20; figure 4 shows that either of these will reduce the estimated social pressure. The different magnitudes above and below \$10 reflect the different sensitivities to the \$0-10 and \$10-20 moments, and the slopes of the lines reflect the fact that as the threshold moves further from \$10 other model predictions $s(\theta, \chi)$ (such as the probability of opening the door) change as well. Any $d' \neq \$10$ would imply that the correct social pressure cost is not \$1.44 but about \$0.10-0.15 higher. We thus find that the estimated social pressure is quite sensitive to the assumed threshold, but that the magnitude of the bias is modest and if anything adjusting for it tends to strengthen the main qualitative conclusions.

6 Conclusions

We propose a new way to inspect the local relationship between an estimator and the moments of the data it depends on. The measure of sensitivity that we propose is essentially costless to estimate even in complex models. It has a precise interpretation in terms of both perturbations to sample realizations of the moments and to their population counterparts. And it provides a formal link between the descriptive question "what features of the data drive the estimates?" and the econometric question "how robust are the conclusions to deviations from the model's assumptions?"

What do we suggest that researchers do in practice? We think that heuristic discussions of the link between data and parameters, commonly framed as statements about identification, could be usefully reframed as statements about sensitivity. Authors could provide economic intuition about which moments are likely to matter most for a given parameter estimate, and back these intuitions up by providing estimates of Λ. This would have several advantages: replacing imprecise heuristic statements with precise ones, supporting intuitive claims with quantitative evidence, and providing a formal rationale for the discussions as a way to gauge sensitivity to model misspecification. Sensitivity could also be used to supplement existing discussions of robustness, with authors evaluating specific hypothetical violations of their assumptions as we do in the applications above.

We do not recommend that researchers abandon the analysis of identification. On the contrary, we think formal analysis of identification and sensitivity are complementary tools for understanding a model and assessing the credibility of results.

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.

Bring, Johan. 1994. How to standardize regression coefficients. *The American Statistician* 48(3): 209-213.

Chen, Xiaohong, Elie Tamer, and Alexander Torgovitsky. 2015. 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.

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 Eco-*

nomics 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.

Gelman, Andrew. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27(15): 2865-2873.

Gelman, Andrew and Guido Imbens. 2014. Why high-order polynomials should not be used in regression discontinuity designs. NBER Working Paper No. 20405.

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.

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*. New York: Wiley.

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.

Kim, Jae-On and G. Donald Ferree Jr. 1981. Standardization in causal analysis. *Sociological Methods & Research* 10(2): 187-210.

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

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: North-Holland.

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.

Wollmann, Thomas. 2016. Trucks without bailouts: Equilibrium product characteristics for commercial vehicles. Chicago-Booth working paper.

Table 1: Sensitivity of average markup in BLP (1995) to beliefs about misspecification

	Bias in average markup
Violation of supply-side exclusion restrictions:	
Removing own car increases average marginal cost	-0.1731
by 1% of average price	(0.0444)
Removing rival's car increases average marginal cost	0.2095
by 1% of average price	(0.0728)
Violation of demand-side exclusion restrictions:	
Removing own car decreases average willingness to pay	-0.1277
by 1% of average price	(0.0920)
Removing rival's car decreases average willingness to pay	0.2515
by 1% of average price	(0.1309)
Baseline estimate	0.3272
	(0.0907)

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 5 under the setup in equation (3). Our calculations use the plug-in estimator of sensitivity. In the first two rows, we set $V_{di} = 0$ and $V_{si} = -0.01(\overline{P}/\overline{mc})Num_i$, 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 $V_{di} = 0.01(\overline{P}/K_{\xi})Num_i$, 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.

Table 2: Sensitivity of preference parameters in Gourinchas and Parker (2002) to beliefs about misspecification

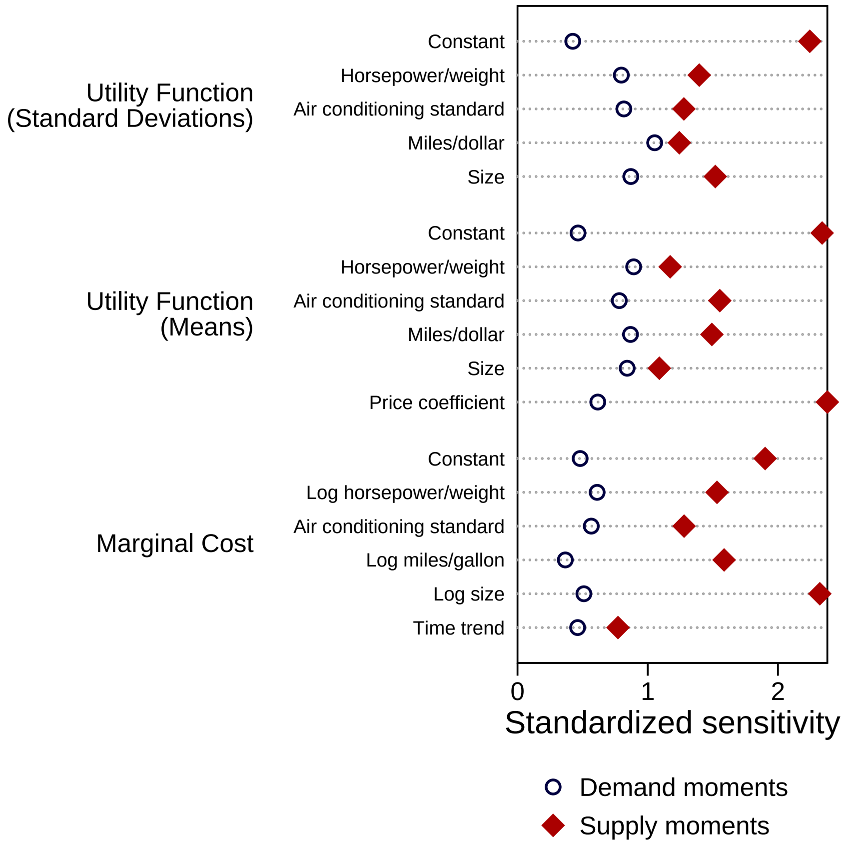
	Bias in discount factor	Bias in coefficient of relative risk aversion
Consumption and leisure are nonseparable:		
Shopping intensity changes with age	0.0041	-0.2913
Exclude 5% of work-related expenses	-0.0073	0.3997
Consumption includes interhousehold transfers:		
Consumption at early ages includes transfers 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 4. Our calculations use the plug-in estimator of sensitivity. We consider perturbations under which measured log consumption overstates true log consumption at each age a by an amount equal to $\eta_a/\sqrt{n}$. In the row labeled "shopping intensity changes with age," η_a is chosen to match the age-specific log price increment estimated in Aguiar and Hurst (2007, table I, column 1). 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_a is chosen so that the true consumption at each age a is overstated by 5% 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_a 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_a 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 26-29. In the row labeled "include both early and late transfers," η_a combines the early age and later age transfers.

Figure 1: Sensitivity of average markup in BLP (1995)

Notes: The plot shows the absolute value of standardized plug-in sensitivity of the implied average markup with respect to the estimation moments, with the sign of sensitivity in parentheses. The demand-side moments are the product of demand-side instruments Z_d with the structural error in utility $\hat{\xi}(\theta_0)$. The supply-side moments are the product of supply-side instruments Z_s with the structural error in marginal cost $\hat{\omega}(\theta_0)$. 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).

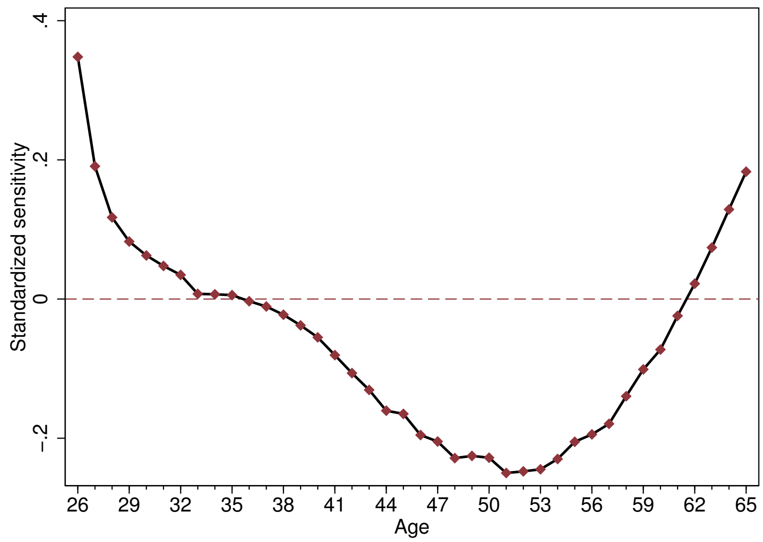
Figure 2: Sensitivity of parameters to supply and demand moments in BLP (1995)



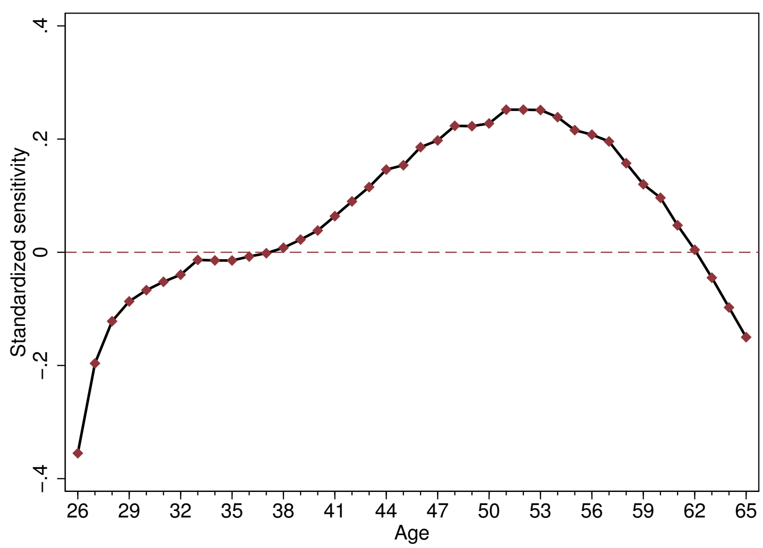
Notes: Rows correspond to parameters of the model. For each parameter, we plot the mean over the all supply and demand moments, respectively, of the absolute value of the standardized plug-in sensitivity. The first group of rows is for standard deviations of the random coefficients in the utility function, the second group is for the means of the random coefficients in the utility function, and the final group is for the parameters of the marginal cost equation.

Figure 3: Sensitivity of select parameters in Gourinchas and Parker (2002)

Panel A: Discount factor

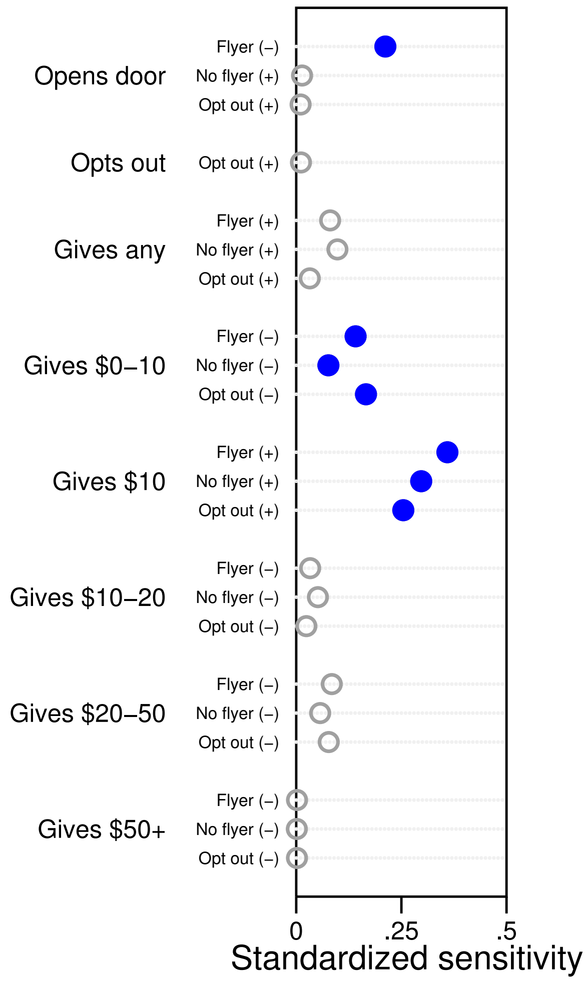


Panel B: Coefficent of relative risk aversion



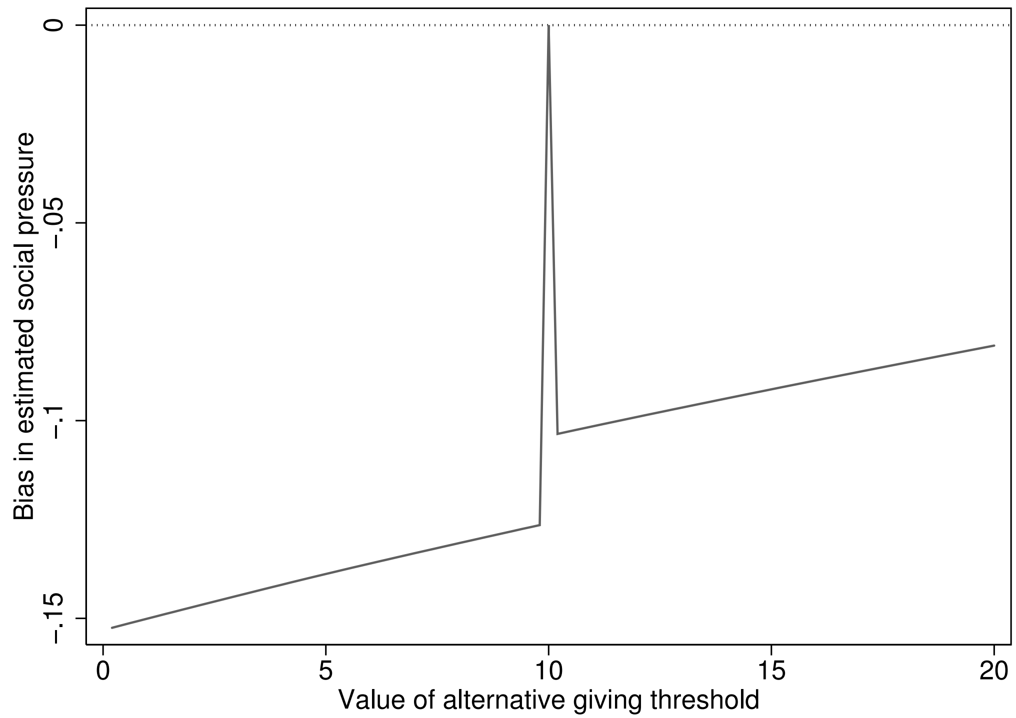
Notes: Each plot shows the standardized plug-in sensitivity of the parameter named in the plot title with respect to the full vector of estimation moments, which are the mean adjusted consumption levels at each age.

Figure 4: Sensitivity of ECU social pressure cost in DellaVigna et al. (2012)



Notes: The plot shows the absolute value of standardized 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 East Carolina Hazard Center 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 5: Sensitivity of DellaVigna et al.'s (2012) ECU social pressure estimate to beliefs about misspecification



Notes: The plot shows the estimated first-order asymptotic bias in DellaVigna et al.'s (2012) published estimate of the social cost of giving to the East Carolina Hazard Center (ECU) under various levels of misspecification, as implied by proposition 4. Our calculations use the plug-in estimator of sensitivity. We consider perturbations under which a share $(1-\mu/\sqrt{n})$ of households have the baseline social preference cost for giving to ECU, which decreases linearly in the amount given up to a threshold of \$10 and is zero thereafter, and a share $\mu/\sqrt{n}$ have a cost function with a different threshold d'. First-order asymptotic bias is computed for values of d' in \$0.20 increments from \$0 to \$20 and interpolated between these increments. Values of d' are shown on the x axis, and each point in the plot corresponds to a different misspecification experiment.

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 3, and suppose that under the local perturbation $\{\mu_n\}_{n=1}^{\infty}$ only one moment j is potentially misspecified ($\mathrm{E}(\tilde{g}_k) = 0$ 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 $\Lambda_{\cdot j}\mathrm{E}\left(\tilde{g}_j\right)$, for $\Lambda_{\cdot j}$ the j^{th} column of Λ.*

Proof. Applying proposition 3, under $F_n(\mu_n)$, $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ converges in distribution to a random variable with mean $\Lambda_{\cdot j}\mathrm{E}\left(\tilde{g}_j\right)$, 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 estimate of sensitivity to misspecification. Rather, when a given moment j is suspect (and the other moments are not), re-estimating after dropping the moment gives a first-order asymptotically unbiased estimate of $\Lambda_{\cdot j}\mathrm{E}\left(\tilde{g}_j\right)$, 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; Morten 2016; and Berger and Vavra 2015).

What this approach yields is 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

$-G = \nabla_\theta s(\theta_0)$, 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

Noting that $F_n = F_n(0)$ is a local perturbation, the proposition is a special case of the consistency of sample sensitivity under local perturbations, which is proved in the online appendix.

B.2 Proof of Proposition 2

By the differentiability of $g(\theta,\mu)$ in μ, we know that $g(\theta,\mu) \to g(\theta)$ pointwise in θ as $\mu \to 0$. Moreover, since $G(\theta,\mu)$ is continuous in $(\theta,\mu) \in \Theta \times \mathscr{B}_\mu$, for $\tilde{\mathscr{B}}_\mu \subset \mathscr{B}_\mu$ a closed ball around zero we know that $\sup_{(\theta,\mu)\in\Theta\times\tilde{\mathscr{B}}_\mu} \lambda_{\max}\left(G(\theta,\mu)' G(\theta,\mu)\right)$ is bounded, where $\lambda_{\max}(A)$ denotes the maximal eigenvalue of a matrix A. This implies that $g(\theta,\mu)$ is uniformly Lipschitz in θ for $\mu \in \tilde{\mathscr{B}}_\mu$, and thus that $g(\theta,\mu) \to g(\theta)$ uniformly in θ as $\mu \to 0$. Thus, $g(\theta,\mu)' W(\mu) g(\theta,\mu)$ converges uniformly to $g(\theta)' W g(\theta)$ as $\mu \to 0$. Since θ_0 is the unique solution to $\min_\theta g(\theta)' W g(\theta)$, this implies that, for any $\varepsilon > 0$, there exists $\mu(\varepsilon) > 0$ such that $\|\theta(\mu) - \theta_0\| < \varepsilon$ whenever $|\mu| < \mu(\varepsilon)$, where $\theta(\mu)$ is the unique solution to

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

Moreover, standard consistency arguments (e.g., theorem 2.1 in Newey and McFadden 1994) imply that $\hat{\theta} \xrightarrow{p} \theta(\mu)$ under $F_n(\mu)$.

Next, note that, for any μ such that $\theta(\mu)$ belongs to the interior of Θ, $\theta(\mu)$ satisfies the first order conditions (in θ)

$$f(\theta(\mu),\mu) = G(\theta(\mu),\mu)' W(\mu) g(\theta(\mu),\mu) = 0.$$

Note that

$$\frac{\partial}{\partial\theta} f(\theta(\mu),\mu) = G(\theta(\mu),\mu)' W(\mu) G(\theta(\mu),\mu) + A(\mu)$$

for

$$A(\mu) = \left[\begin{array}{ccc} G_1(\theta(\mu),\mu)' W(\mu) g(\theta(\mu),\mu) & \dots & G_P(\theta(\mu),\mu)' W(\mu) g(\theta(\mu),\mu) \end{array} \right]$$

with

$$G_p(\theta(\mu),\mu) = \frac{\partial}{\partial \theta_p} G(\theta(\mu),\mu).$$

Since we have assumed that $G'WG$ is non-singular and $\frac{\partial}{\partial\theta} f(\theta,\mu)$ is continuous in θ and μ, $\frac{\partial}{\partial\theta} f(\theta(\mu),\mu)$ has full rank for μ sufficiently close to zero. Thus, by the implicit function theorem, for μ in an open neighborhood of zero we can define a unique continuous function $\tilde{\theta}(\mu)$ such that $f\left(\tilde{\theta}(\mu),\mu\right) = 0$ and

$$\frac{\partial}{\partial \delta}\tilde{\theta}(\mu) = \begin{array}{l} -\left(G(\theta(\mu),\mu)' W(\mu) G(\theta(\mu),\mu) + A(\mu)\right)^{-1} \\ \times \left(G(\theta(\mu),\mu)' W(\mu) \frac{\partial}{\partial\mu} g(\theta(\mu),\mu) + B(\mu)\right) \end{array}$$

for

$$B(\mu) = G(\theta(\mu),\mu) \frac{\partial}{\partial \mu} W(\mu) g(\theta(\mu),\mu).$$

By the argument at the beginning of this proof, uniqueness implies that $\theta(\mu) = \tilde{\theta}(\mu)$ for μ sufficiently small. This means that $\tilde{\theta}(0) = \theta(0) = \theta_0$. Thus, since $g(\theta(0),0) = g(\theta_0) = 0$, both $A(0)$ and $B(0)$ are equal to zero, from which the conclusion follows immediately for $\mathscr{B}^*_\mu$ sufficiently small.

B.3 Proof of Proposition 3

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:

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

By the mean value theorem,

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

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

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

Rearranging, we have

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

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

We know that $\hat{\theta}\xrightarrow{p}\theta_0$ under $F_n\left(\mu_n\right)$, so $\overline{\theta}\xrightarrow{p}\theta_0$. This plus uniform convergence of $\hat{G}\left(\theta\right)$ to $G\left(\theta\right)$ implies that under $F_n\left(\mu_n\right)$, $\hat{G}\left(\hat{\theta}\right)$ 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

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

which converges in probability to zero by the Slutsky theorem (using the fact that $\sqrt{n}\hat{g}\left(\theta_0\right)$ converges in distribution). Therefore, under $F_n\left(\mu_n\right)$, $\sqrt{n}\left(\hat{\theta}-\theta_0,\Lambda\hat{g}\left(\theta_0\right)\right)$ converges in distribution to a random vector $\left(\tilde{\theta},\tilde{g}\right)$ with $\Pr\left\{\tilde{\theta}=\Lambda\tilde{g}\right\}=1$. This implies in particular that $\mathrm{E}\left(\tilde{\theta}\right)=\Lambda\mathrm{E}\left(\tilde{g}\right)$.

B.4 Proof of Lemma 1

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

B.5 Proof of Lemma 2

By assumption 3 part (ii) we know that $\left(\zeta_i,X_i,V_i\right)$ has density $f\left(\zeta_i,X_i,V_i\right)$ with respect to ν under $F\left(0\right)$. Thus, the density $f\left(\zeta_i,X_i,V_i|\mu\right)$ is given by $f\left(\zeta_i-\mu\cdot V_i,X_i,V_i\right)$. By assumption 3 part (iii), $\sqrt{f\left(\zeta_i-\mu\cdot V_i,X_i,V_i\right)}$ is continuously differentiable in ζ_i, which implies that

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

is continuous in μ for all $\left(\zeta_i-\mu\cdot V_i,X_i,V_i\right)$. By assumption 3 part (iv) we know that

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

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

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

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 $\mu_n = \frac{1}{\sqrt{n}}\mu^*$ for a fixed value μ^*, 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 3 parts (iv) and (v), and the central limit theorem imply that under $F_n(0)$,

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

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),

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

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