NBER WORKING PAPER SERIES

MEASURING THE SENSITIVITY OF PARAMETER ESTIMATES TO SAMPLE STATISTICS

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

Conversations with Kevin M. Murphy inspired and greatly improved this work. We are grateful also to Isaiah Andrews, Josh Angrist, Steve Berry, Alan Bester, Raj Chetty, Tim Conley, Ron Goettler, Brett Gordon, Phil Haile, Christian Hansen, Pat Kline, Mark Li, Magne Mogstad, Yaroslav Mukhin, Matt Taddy, E. Glen Weyl, and seminar audiences at Berkeley, University of Chicago, Harvard, University of Michigan, MIT, NBER, Northwestern, NYU, Princeton, Stanford, and Yale for advice and suggestions, and to our dedicated research assistants for important contributions to this project. Except where otherwise noted, all original code and data for our applications came from the paper's authors either via e-mail or from journal replication archives. We thank the following authors for their assistance in working with their code and data: Mariacristina De Nardi, Eric French, and John B. Jones; Stefano DellaVigna, John List, and Ulrike Malmendier; Ron Goettler and Brett Gordon; Pierre-Olivier Gourinchas and Jonathan Parker; Nathaniel Hendren; Chris Knittel and Konstantinos Metaxoglou; Michael Mazzeo; Boris Nikolov and Toni Whited; and Amil Petrin. This research was funded in part by the Initiative on Global Markets, the George J. Stigler Center for the Study of the Economy and the State, the Ewing Marion Kauffman Foundation, the Centel Foundation / Robert P. Reuss Faculty Research Fund, the Neubauer Family Foundation, and the Kathryn C. Gould Research Fund, all at the University of Chicago Booth School of Business, the Alfred P. Sloan Foundation, 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 Sample Statistics
Matthew Gentzkow and Jesse M. Shapiro
NBER Working Paper No. 20673
November 2014, Revised May 2015
JEL No. C1,C52

ABSTRACT

Empirical papers in economics often describe heuristically how their estimates depend on intuitive features of the data. We propose two quantitative measures of this relationship that can be computed at negligible cost even for complex models. We show that our measures can be informative about robustness to model misspecification, and can complement the discussions of identification that have become common in applied work. We illustrate our measures with applications to industrial organization, macroeconomics, public economics, and finance.

Matthew Gentzkow
University of Chicago
Booth School of Business
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
gentzkow@chicagobooth.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://faculty.chicagobooth.edu/matthew.gentzkow/research/trans_online.pdf

1 Introduction

An estimator is a mapping from data to parameters of interest. In many cases, it is possible to show analytically how the estimates change as the data vary along particular dimensions. In other cases, the estimator is sufficiently complicated that interrogating the mapping through brute-force computation, or through direct inspection of the economic and econometric assumptions, is prohibitively costly. In this paper, we suggest two measures of an estimator's relationship to specific features of the data that are easy to compute even for complex models. We then apply these measures to gain new insight into structural empirical models in industrial organization, macroeconomics, public economics, and finance.

Throughout the paper, we consider the following abstract setting. A researcher observes data from an unknown distribution $F(\cdot)$, and computes an estimator $\hat{\theta}$ of a finite number of economic parameters θ. The researcher also computes a vector of statistics $\hat{\gamma}$ that summarize some data features of interest. These may be the moments used in estimating $\hat{\theta}$ in a GMM or simulated moments procedure, descriptive statistics such as means or variances, or estimates of the parameters of an auxiliary model. The statistics $\hat{\theta}$ and $\hat{\gamma}$ are jointly asymptotically normal, with $\sqrt{n}\left(\hat{\theta}-\theta_0, \hat{\gamma}-\gamma_0\right) \xrightarrow{d} \left(\tilde{\theta}, \tilde{\gamma}\right) \sim N(0, \Sigma)$, where θ_0 is the true value of θ and γ_0 is the population value of $\hat{\gamma}$.

To take a concrete example, consider a life-cycle consumption model, where the parameters in θ are the discount factor and coefficient of relative risk aversion, and $\hat{\gamma}$ is a vector of moments of the joint distribution of consumption and income from a cross-section of consumers. There are two questions we might naturally ask about a specific estimator $\hat{\theta}$. First, which features of the data drive the estimates? Does the estimated discount factor, say, depend primarily on the mean consumption and income of consumers at different ages? Or does it also depend on higher order moments like the variance of consumption, or the covariance of consumption and income growth, for those of a given age? Second, how do the estimates change as we vary these features? If the discount factor depends only on mean consumption by age, for example, which consumption profiles will the model interpret as evidence for relatively higher or lower discount factors?

Motivated by the first question, we define the asymptotic *sufficiency* of $\hat{\gamma}$ for $\hat{\theta}$ to be the share of the variance in $\tilde{\theta}$ explained by $\tilde{\gamma}$ (i.e., $\mathrm{Var}\left(\mathrm{E}\left[\tilde{\theta}|\tilde{\gamma}\right]\right)/\mathrm{Var}\left(\tilde{\theta}\right)$). When this is high, $\hat{\gamma}$ captures most of the information in the data that is relevant for $\hat{\theta}$ in large samples, and the relationship between the two may provide useful intuition about the behavior of the estimator. When it is low, the relationship between $\hat{\gamma}$ and $\hat{\theta}$ may be less informative. When it is equal to one, $\tilde{\gamma}$ is a sufficient statistic for $\tilde{\theta}$.[1] This will be true, for example, when $\hat{\theta}$ is a moment estimator and $\hat{\gamma}$ is the vector

[1] That is, for any statistic $\hat{s}$ with $\sqrt{n}(\hat{s}-s_0) \xrightarrow{d} \tilde{s}$, the distribution of $\tilde{s}$ conditional on $\tilde{\gamma}$ and $\tilde{\theta}$ is equal to the distribution conditional on $\tilde{\gamma}$ alone. Like Chetty (2009), we extend the usual definition of "sufficient statistic" slightly: the standard case refers to sufficiency of a statistic for a parameter; we extend this in the natural way to encompass

of estimation moments evaluated at the true parameter value.

Motivated by the second question, we define the asymptotic *sensitivity* of $\hat{\theta}$ to $\hat{\gamma}$ to be the coefficient from a regression of $\tilde{\theta}$ on $\tilde{\gamma}$. Sensitivity measures which values of $\hat{\gamma}$ will be interpreted as evidence for higher or lower θ. We expect sensitivity to be most informative when sufficiency is close to one. When $\hat{\theta}$ is a deterministic function of $\hat{\gamma}$, sufficiency is one, and sensitivity corresponds to the partial derivative of $\hat{\theta}$ with respect to $\hat{\gamma}$ evaluated at the probability limit of $\hat{\gamma}$.

Both of our measures can be estimated at low cost even in computationally challenging models. In the common case where $\hat{\gamma}$ is the vector of moments used in estimating $\hat{\theta}$, sufficiency and sensitivity can be estimated at essentially no computational cost by manipulating the objects used to estimate asymptotic standard errors. In a large class of remaining cases, sufficiency and sensitivity can be estimated using easily computed empirical influence statistics, without any simulation or re-estimation of the model. The measures can also be trivially extended to the case where the economic quantity of interest is not a parameter itself but a function of underlying parameters: an elasticity, say, or a summary of a counterfactual policy simulation.

There are two related reasons why we think these measures may be useful additions to the applied economist's toolkit. First, knowing which data features drive a particular estimator is relevant to assessing the sensitivity of estimates to particular forms of misspecification. Suppose, for example, that our life-cycle consumption model assumes (i) the marginal utility of consumption does not vary by age, and (ii) cross-sectional variation in income is uncorrelated with preference shocks. If our sufficiency measure implies that $\hat{\theta}_1$ depends primarily on the average growth in consumption across ages ($\hat{\gamma}_1$), we might be relatively more concerned about violations of assumption (i). If instead $\hat{\theta}_1$ depends primarily on the covariance of income growth and consumption for those of a given age ($\hat{\gamma}_2$), we might be more concerned about assumption (ii). Moreover, knowing the sensitivity of $\hat{\theta}_1$ to the relevant moments in each case can help us say something about the bias we expect in $\hat{\theta}_1$. If the most plausible violations of (i) imply relatively greater marginal utility of consumption at older ages, for example, the resulting bias will have the same sign as the sensitivity of $\hat{\theta}_1$ to $\hat{\gamma}_1$.

We make these intuitions precise for forms of misspecification that are "local" in an appropriate sense. In a linear IV model, when $\hat{\theta}$ is the IV coefficient and $\hat{\gamma}$ is the product of instruments and model residuals, our sensitivity measure is precisely the expression derived by Conley et al. (2012) to adjust inference for local violations of the exogeneity assumption. Our characterization extends Conley et al.'s (2012) analysis to a much larger class of models, including nonlinear IV models such as Berry et al. (1995) and classical minimum distance. Our results also reproduce Gelman and Imbens (2014)'s analysis of the dependence of regression discontinuity estimators on outcomes at specific ranges of the forcing variable.

asymptotic sufficiency of one statistic for another statistic.

The second reason our measures may be useful is that they complement the discussions of identification that have become an increasingly important part of applied economics. Empirical papers now frequently devote sections to formal or informal proofs of identification in the spirit of Matzkin (2007; 2013) and Berry and Haile (2014). These proofs ideally do two things: they show that the economic quantities of interest can be recovered in a way that does not rely on arbitrary functional form or distributional assumptions, and they implicitly define an estimator that relates these quantities to specific features of the data in a transparent way.[2] Matzkin (2013), Angrist and Pischke (2010), Heckman (2010), and Pakes (2003), all emphasize various ways in which such transparency can enhance the credibility of empirical findings.

The problem is that the estimator implicitly defined by the discussion of identification is typically not the one researchers take to the data.[3] Many papers that offer informal proofs of nonparametric identification go on to use a parametric estimator that imposes strong functional form restrictions, and many that show that their model is identified by a small set of moments go on to use an estimator that depends on richer variation in the data. Knowing that the quantities of interest could *in principle* have been estimated with weaker assumptions or in a more transparent way is valuable for understanding the workings of the model. But it seems hard to see how this is relevant to the credibility of the *actual* estimates, unless the estimator being taken to the data bears at least some resemblance to the hypothetical one.

Sufficiency and sensitivity provide one low-cost way to evaluate whether the actual and hypothetical estimators are taking information from the data in similar ways. Suppose we show that θ is identified from a vector of data features γ through a function $\Phi(\gamma)$. For example, we might prove that the discount factor θ_1 in our life-cycle consumption model is identified from the average growth of consumption γ_1. If the estimator $\hat{\theta}_1$ we take to the data were in fact the hypothetical one defined by $\Phi(\cdot)$, we would observe (i) that sufficiency of $\hat{\gamma}_1$ for $\hat{\theta}_1$ is equal to one, and (ii) sensitivity of $\hat{\theta}_1$ to $\hat{\gamma}_1$ is equal to the partial derivative of $\Phi(\cdot)$ with respect to γ_1 evaluated at the population value. If this is approximately true for the actual estimator, we may conclude that the intuitions from the identification proof will be a good guide to the way the estimator takes information from the data, and we may be more confident using the discussion of identification as a guide to judging the credibility of the results. If sufficiency is very low or the sensitivities are very different from what the proof would suggest, we may conclude the proof of identification is less relevant to judging the credibility of the estimator.

[2] Matzkin (2013) writes: "Constructive identification methods indicate in a transparent way the connection between the [economic quantity] of interest and the distribution of observable variables. They provide a way to read off the distribution of the observable variables the [economic quantity] of interest. Methods for constructive identification directly lead to methods of estimation for the object of interest" (p. 461).

[3] For example, of the four structural papers published in the *American Economic Review* in 2013 that included formal or informal proofs of nonparametric or semiparametric identification, all but one take models with additional parametric restrictions to the data.

In the final sections of the paper, we present estimates of sensitivity and sufficiency for a number of empirical papers. We begin with an application to two models of intertemporal choice. Applying sensitivity to Gourinchas and Parker's (2002) model of life-cycle consumption and saving, we show how information on consumption at different ages drives inference about time and risk preference. In an application to De Nardi et al.'s (2010) model of post-retirement saving, we show how a parameter not present in Gourinchas and Parker's (2002) model is pinned down by data on the asset holdings of rich and poor households.

We turn next to an analysis of Berry et al.'s (1995) empirical model of the automobile market. We use sufficiency to quantify the importance of demand-side and supply-side estimation moments in driving the estimated markup, and we use sensitivity to gauge which instruments' exclusion restrictions would matter most if they were violated. We also show that estimates of a much simpler model—a logit with no unobserved heterogeneity—do a poor job of capturing the information in the data that pins down Berry et al.'s (1995) estimated parameters.

After these detailed applications we present shorter applications to Goettler and Gordon's (2011) study of competition between AMD and Intel, DellaVigna et al.'s (2012) model of charitable giving, and Nikolov and Whited's (2014) model of corporate investment. For each paper, we let $\hat{\gamma}$ be the vector of empirical moments used in estimation. In most cases, our analysis suggests the actual estimators take information from the data similarly to what the authors' identification argument would suggest, but we also find cases in which parameter estimates depend on information in the data that the authors did not highlight as important.

Our final applications are to Mazzeo's (2002) model of motel entry, Gentzkow's (2007) model of competition between print and online newspapers, and Hendren's (2013) model of the market for long-term care insurance. Because these papers use maximum likelihood estimators, we let $\hat{\gamma}$ be various descriptive statistics, rather than estimation moments. We find that there is often a tight link between the estimates of structural parameters and the corresponding descriptive statistics. For example, in the case of Mazzeo's (2002) model, we show that estimates of an analogous linear regression model capture more than 80 percent of the information in the data that is used to estimate key parameters. This finding suggests a way in which sensitivity and sufficiency can be used to build up linear intuitions for the inner workings of nonlinear models.

An important limitation of our formal approach is that, because we focus on properties of the asymptotic distribution, the notions of sufficiency and sensitivity that we consider are intrinsically local. The approximations that we work with have the same mechanics and hence the same limitations as those commonly used to compute asymptotic standard errors. Generalizing our approach to more global exploration of model properties is conceptually straightforward but may be computationally expensive. In our concluding section, we provide some guidance on how a researcher might minimize computational costs in practice.

A second limitation is that the units of sensitivity are contingent on the units of γ. We suggest a normalization in section 4.3 below that serves as a useful default for many practical applications but acknowledge that the appropriate scaling of sensitivity may be application-specific.

The main contributions of the paper are to suggest new tools for applied researchers, to note their properties, and to demonstrate their applicability. All of the econometric statements we make follow in a straightforward way from well-known results. As we focus on properties of the estimator $\hat{\theta}$ rather than of the underlying model F, our approach is closely related to the study of the robustness of estimators (Huber and Ronchetti 2009).

The recent methodological conversation about "structural" vs. "reduced-form" or "program evaluation" methods centers on a perceived tradeoff between the realism of an empirical model's economic assumptions and the transparency of its mapping from data to parameters.[4] Our measures make this tradeoff shallower by permitting a precise characterization of the dependence of a structural estimate on intuitive features of the data.[5] Our measures also facilitate the analysis of sensitivity to misspecification (Leamer 1983), including misspecification of exclusion restrictions (Rosenbaum and Rubin 1983; Conley et al. 2012; Nevo and Rosen 2012).[6]

Our work is also closely related to the large literature on sensitivity analysis for scientific models (Sobol 1993; Saltelli et al. 2008).[7] In a Bayesian context, our measures of sensitivity and sufficiency may be thought of as analogous to the measures of prior sensitivity and prior informativeness developed by Müller (2012) for studying the importance of the prior.

The remainder of the paper is organized as follows. Section 2 defines our measures, section 3 discusses their properties and interpretation, and section 4 shows how to estimate them. Sections 5 and 6 apply the measures to several empirical papers. Section 7 concludes. An appendix relates our approach to some alternatives.

[4] Heckman (2010) writes that "The often complex computational methods that are required to implement [structural estimation] make it less transparent" (p. 358). Angrist and Pischke (2010) write that "in [Nevo's (2000)] framework, it's hard to see precisely which features of the data drive the ultimate results" (p. 21).

[5] Because our sensitivity measure correctly identifies cases in which only a subset of empirical moments is needed to answer a question of interest, sensitivity analysis may also be seen as a complement to the "sufficient statistics" approach of Chetty (2009), Einav et al. (2010), and Jaffe and Weyl (2013). Our measure relates in a similar way to indirect inference methods (Gourieroux et al. 1993; Smith 1993), which can be used to estimate structural parameters from an intentionally limited set of descriptive statistics (see, e.g., Martin and Yurukoglu 2014). As we illustrate below, our approach allows a researcher to relate estimated structural parameters to descriptive statistics even if those statistics are not used directly in estimation.

[6] Vasnev (2006) defines a test statistic for the sensitivity of a focus parameter to the value of a nuisance parameter which may be thought of a controlling a particular type of misspecification. Hansen (2008) discusses some approaches to controlling the effects of misspecification in generalized method of moments estimation.

[7] Linear regression of model outputs on model inputs is a standard tool for model interrogation in the physical sciences. We show that the asymptotic properties of common estimators used in economics make it possible to perform such an analysis without repeatedly re-estimating or simulating the model, thus sparing substantial computational expense.

2 Measures of Sufficiency and Sensitivity

2.1 Definitions

A researcher observes a random sample of size n from a distribution $F(\cdot|\theta)$ on sample space $\mathscr{X}$, and computes (i) a $(P \times 1)$ estimator $\hat{\theta}$ of θ; and (ii) a $(J \times 1)$ vector of auxiliary statistics $\hat{\gamma}$, both of which depend on the data. The true value of θ is θ_0, and the population value of γ is γ_0.

We assume that under $F(\cdot|\theta_0)$,

$$\text{(1)} \qquad \sqrt{n}\begin{pmatrix} \hat{\theta}-\theta_0 \\ \hat{\gamma}-\gamma_0 \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tilde{\theta} \\ \tilde{\gamma} \end{pmatrix} \sim N(0,\Sigma),$$

for some finite Σ, and that the submatrix $\Sigma_{\gamma\gamma}$ of Σ corresponding to the variance of $\tilde{\gamma}$ is nonsingular.

From equation (1) it follows that the conditional expectation of $\tilde{\theta}$ given $\tilde{\gamma}$ is linear. Letting $\Sigma_{\theta\gamma}$ denote the submatrix of Σ corresponding to the covariance of $\tilde{\theta}$ and $\tilde{\gamma}$, we have:

$$\text{(2)} \qquad \mathrm{E}\left(\tilde{\theta}|\tilde{\gamma}\right) = \Sigma_{\theta\gamma}\Sigma_{\gamma\gamma}^{-1}\tilde{\gamma}.$$

Definition. *The* **sufficiency** *of $\hat{\gamma}$ for an element $\hat{\theta}_p$ of $\hat{\theta}$ is*

$$\Delta_p = \frac{\mathrm{Var}\left(\mathrm{E}\left(\tilde{\theta}_p|\tilde{\gamma}\right)\right)}{\mathrm{Var}\left(\tilde{\theta}_p\right)} = \frac{\left(\Sigma_{\theta\gamma}\Sigma_{\gamma\gamma}^{-1}\Sigma_{\theta\gamma}'\right)_{pp}}{\left(\Sigma_{\theta\theta}\right)_{pp}}.$$

We let Δ denote the column vector of Δ_p. We say that $\hat{\gamma}$ is **sufficient for** $\hat{\theta}_p$ *if $\Delta_p = 1$ and that $\hat{\gamma}$ is* **sufficient for** $\hat{\theta}$ *if $\Delta = 1$.*

Definition. *The* **sensitivity** *of $\hat{\theta}$ to $\hat{\gamma}$ is*

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

Sufficiency $\Delta_p \in [0,1]$ is the probability limit of the R^2 of a regression of realizations of $\tilde{\theta}_p$ on realizations of $\tilde{\gamma}$, as the number of realizations grows large. It captures the extent to which, in large samples, knowledge of $\hat{\gamma}$ is sufficient to predict the value of the estimator.

Sensitivity Λ is the coefficient from a regression of $\tilde{\theta}$ on $\tilde{\gamma}$. An element Λ_{pj} of Λ is the effect of changing the realization of a particular $\tilde{\gamma}_j$ on the expected value of a particular $\tilde{\theta}_p$, holding constant the other elements of $\tilde{\gamma}$.[8] We expect sensitivity to be of most interest when $\Delta = 1$ or $\Delta \approx 1$.

[8] It follows from equation (2) that Λ_{pj}^2 is the partial derivative of the variance of $\mathrm{E}\left(\tilde{\theta}_p|\tilde{\gamma}\right)$ with respect to the variance of $\tilde{\gamma}_j$. In this sense, Λ captures not only the impact of $\tilde{\gamma}$ on $\tilde{\theta}$, but also the impact of *uncertainty* about γ_0 on uncertainty about θ_0.

2.2 Useful Properties

$\hat{\theta}$ is a functional whose domain is typically the space of empirical distributions on $\mathscr{X}$. Sufficiency allows us to identify situations in which we can approximate $\hat{\theta}$ as a low-dimensional function that depends on the data only via a vector of interpretable statistics $\hat{\gamma}$. When this is the case, we can use sensitivity to interrogate the estimator's local behavior. The following proposition makes this way of thinking about our measures explicit.

Proposition 1. *$\hat{\gamma}$ is sufficient for $\hat{\theta}_p$ if and only if there exists a continuously differentiable function $h(\cdot;\theta_0)$ with non-zero gradient at γ_0 such that $\hat{\theta}_p = h(\hat{\gamma};\theta_0) + o_P\left(\frac{1}{\sqrt{n}}\right)$. If such an $h(\cdot)$ exists, its partial derivative at γ_0 is $\Lambda_{p\cdot}$.*

Proof. If $\hat{\gamma}$ is sufficient for $\hat{\theta}_p$, let $h(\hat{\gamma};\theta_0) = \theta_{0p} + \Lambda_{p\cdot}(\hat{\gamma} - \gamma_0)$. It follows from standard limit results that $\sqrt{n}\left[\hat{\theta}_p - h(\hat{\gamma};\theta_0)\right] \xrightarrow{p} 0$. Conversely, if there exists an $h(\cdot;\theta_0)$ satisfying the given conditions, we know that $\sqrt{n}\left(h(\hat{\gamma};\theta_0) - h(\gamma_0;\theta_0), \hat{\gamma} - \gamma_0\right)$ converges in distribution to $\left(\tilde{\theta}_p, \tilde{\gamma}\right)$. It follows from the delta method that $\hat{\gamma}$ is sufficient for $\hat{\theta}_p$ and the partial derivative of h at γ_0 is equal to $\Lambda_{p\cdot}$. □

Note that the function $h(\cdot;\theta_0)$ does not depend on the realized data except through $\hat{\gamma}$, but that it does depend on the true value of θ.

The relationship $h(\cdot;\theta_0)$ between $\hat{\gamma}$ and $\hat{\theta}$ remains valid for a class of local perturbations of the data-generating process. Denote by $F_n(\cdot|\theta) = \times_n F(\cdot|\theta)$ the joint distribution of the data for sample size n. This is a sequence of distribution functions with associated sample spaces $\mathscr{X}^n$. Consider some alternative sequence of distributions $F_n^*(\cdot|\theta)$, also defined on $\mathscr{X}^n$. Then $F_n(\cdot|\theta)$ and $F_n^*(\cdot|\theta)$ are *mutually contiguous* if for any sequence of statistics $T_n : \mathscr{X}^n \to \mathbb{R}^d$ (where d is an arbitrary natural number), $T_n \xrightarrow{p} 0$ under F_n if and only if $T_n \xrightarrow{p} 0$ under F_n^*.[9] Then:

Proposition 2. *Suppose that $F_n(\cdot|\theta_0)$ and $F_n^*(\cdot|\theta_0)$ are mutually contiguous, and that $\sqrt{n}\left(\hat{\theta} - \theta_0, \hat{\gamma} - \gamma_0\right)$ converges in distribution to a random variable $\left(\tilde{\theta}', \tilde{\gamma}'\right)$ under $F_n^*(\cdot|\theta_0)$. Then if $\hat{\gamma}$ is sufficient for $\hat{\theta}$ under $F_n(\cdot|\theta_0)$, we have $\tilde{\theta}' = \Lambda\tilde{\gamma}'$ almost surely.*

Proof. If $\hat{\gamma}$ is sufficient for $\hat{\theta}$, we know that $\sqrt{n}\left[\left(\hat{\theta} - \theta_0\right) - \Lambda(\hat{\gamma} - \gamma_0)\right] \xrightarrow{p} 0$ under $F_n(\cdot|\theta_0)$ which implies $\tilde{\theta} = \Lambda\tilde{\gamma}$ almost surely. By contiguity, $\sqrt{n}\left[\left(\hat{\theta} - \theta_0\right) - \Lambda(\hat{\gamma} - \gamma_0)\right] \xrightarrow{p} 0$ under $F_n^*(\cdot|\theta_0)$ and so $\tilde{\theta}^{'} = \Lambda\tilde{\gamma}^{'}$ almost surely. □

Contiguity is a standard way to define local perturbations of models in asymptotic statistics (Van der Vaart 1998).[10] Contiguity is equivalent to there being no sequence of tests which can

[9] There are several equivalent definitions of contiguity. The usual primitive definition is: for any sequence of measurable sets A_n, $F_n(A_n|\theta) \to 0$ if and only if $F_n^*(A_n|\theta) \to 0$. The equivalence to the definition in the text is given by Le Cam's first lemma (Van der Vaart 1998).

[10] Kitamura et al. (2013) discuss robust estimation under related forms of local model misspecification.

perfectly distinguish the two models asymptotically. We show in section 3.1 below that the class of mutually contiguous perturbations encompasses many intuitive cases, including small violations of exogeneity assumptions as in Conley et al. (2012).

It is easy to extend our measures to cases in which the quantity of interest is a function of the underlying parameters, such as a welfare calculation, an elasticity, or a counterfactual prediction simulated from the model. It is also easy to consider sensitivity to a function of $\hat{\gamma}$, such as an average of related moments.

Remark 1. If $c\left(\hat{\theta}\right)$ is a continuously differentiable function, not dependent on the data and with non-zero gradient C at θ_0, the delta method implies that the sensitivity of $c\left(\hat{\theta}\right)$ to $\hat{\gamma}$ is equal to $C\Lambda$, where Λ is the sensitivity of $\hat{\theta}$ to $\hat{\gamma}$. Similarly, if $a\left(\hat{\gamma}\right)$ is a continuously differentiable function, not dependent on the data and with non-zero gradient A at γ_0, the delta method implies that the sensitivity of $\hat{\theta}$ to $a\left(\hat{\gamma}\right)$ is equal to $\Sigma_{\theta\gamma}A'\left(A\Sigma_{\gamma\gamma}A'\right)^{-1}$.

The applications we will discuss are all examples of minimum distance estimators (MDE), a class that includes generalized method of moments (GMM), maximum likelihood (MLE), and classical minimum distance (CMD), as well as simulated analogues such as simulated minimum distance (SMD) and simulated method of moments (SMM). Formally:

Definition. *$\hat{\theta}$ is a* **minimum distance estimator** *(MDE) if we can write*

$$\hat{\theta} \quad = \quad \underset{\theta\in\Theta}{\arg\min}\, \hat{g}\left(\theta\right)'\hat{W}_g\hat{g}\left(\theta\right), \tag{3}$$

where $\hat{g}\left(\theta\right)$ is a function of parameters and data, $\sqrt{n}\hat{g}\left(\theta_0\right)\xrightarrow{d}N\left(0,\Omega_{gg}\right)$, $\hat{W}_g\xrightarrow{p}W_g$, and $\hat{W}_g$ and W_g are positive semi-definite. We assume standard regularity conditions such that $\hat{\theta}$ is consistent and asymptotically normal with variance $\left(G'W_gG\right)^{-1}G'W_g\Omega_{gg}W_gG\left(G'W_gG\right)^{-1}$, where G is the Jacobian of an appropriate limit of $\hat{g}$ evaluated at θ_0.[11]

When $\hat{\theta}$ is an MDE and $\hat{\gamma}=\hat{g}\left(\theta_0\right)$, we say that Λ is *sensitivity to moments*. In this case it is straightforward to derive an expression for Λ.

Remark 2. If $\hat{\theta}$ is an MDE and Λ is sensitivity to moments, then $\Lambda=-\left(G'W_gG\right)^{-1}G'W_g$ and $\Delta=1$.

[11] We allow some flexibility in the definition of G here so that our definition of MDE includes both cases where $\hat{g}\left(\theta\right)$ is a smooth function of parameters (as in GMM or CMD), and cases where $\hat{g}\left(\theta\right)$ is not smooth (as in SMM). For the precise regularity conditions and definition of G, see Newey and McFadden (1994) Theorem 3.2 for the smooth case and Theorem 7.2 for the non-smooth case.

2.3 Examples

In this section, we study two "pen and paper" examples that illustrate the intuitions our measures deliver in well-understood cases.

Example 1. (OLS) $\hat{\theta} = \left[\begin{array}{cc} \hat{\alpha} & \hat{\beta} \end{array}\right]'$ is the constant term and coefficient from an OLS regression of Y on a scalar X. We assume standard conditions for the consistency and asymptotic normality of $\hat{\theta}$. Define $\hat{\gamma} = \left[\begin{array}{cccc} \hat{\mu}_y & \hat{\sigma}_{xy} & \hat{\sigma}_x^2 & \hat{\mu}_x \end{array}\right]'$, where $\hat{\mu}_y$ is the sample mean of Y, $\hat{\sigma}_{xy}$ is the sample covariance of X and Y, $\hat{\sigma}_x^2$ is the sample variance of X, and $\hat{\mu}_x$ is the sample mean of X. We can write $\hat{\alpha} = \hat{\mu}_y - \hat{\beta}\hat{\mu}_x$ and $\hat{\beta} = \hat{\sigma}_{xy}/\hat{\sigma}_x^2$, so by proposition 1, $\Delta = 1$ and we can solve for Λ by evaluating the partial derivatives of the estimates at the population values of $\hat{\gamma}$. Focusing on the first two columns of Λ, which give sensitivity to $\hat{\mu}_y$ and $\hat{\sigma}_{xy}$ respectively, we have:

$$\Lambda = \begin{bmatrix} 1 & -\frac{\mu_x}{\sigma_x^2} & \ldots \\ 0 & \frac{1}{\sigma_x^2} & \ldots \end{bmatrix},$$

where μ_x and σ_x^2 are the population mean and variance of X. Consistent with intuition, we find that when the mean of X is zero, the constant $\hat{\alpha}$ is sensitive to $\hat{\mu}_y$ but not $\hat{\sigma}_{xy}$, and $\hat{\beta}$ is sensitive to $\hat{\sigma}_{xy}$ but not $\hat{\mu}_y$. When the mean of X is not zero, $\hat{\alpha}$ is also sensitive to $\hat{\sigma}_{xy}$ because this affects the sample average of $\hat{\beta}X$. It is straightforward to generalize the example to multivariate regression.[12]

Example 2. (Regression Discontinuity) Consider a regression discontinuity (RD) model estimated using OLS regression of Y on K-th degree polynomials in the forcing variable X above and below the discontinuity:

$$Y_i = \mathbf{1}_{X_i<0} \sum_{k=0}^{K} X_i^k \beta_k^- + \mathbf{1}_{X_i \geq 0} \sum_{k=0}^{K} X_i^k \beta_k^+ + \varepsilon_i,$$

where we normalize X so the discontinuity is at $X = 0$ and we limit the sample to observations within some fixed bandwidth of $X = 0$. The RD estimator of the treatment effect is $\hat{\tau} = \hat{\beta}_0^+ - \hat{\beta}_0^-$.

Gelman and Imbens (2014) analyze the weights that different estimators place on outcomes far from the discontinuity.[13] They use their analysis to criticize the common practice of combining

[12] Let independent variables be $X = [X_1, ..., X_J]$ and let $\hat{\gamma} = \frac{1}{n}\left[\begin{array}{ccccccc} Y'X & X_1'X_1 & X_1'X_2 & \cdots & X_j'X_k & \cdots & X_J'X_J \end{array}\right]'$ with $j \in \{1, ..., J\}$ and $k \in \{j, ..., J\}$ (so there are no redundant elements). Then $\Delta = 1$ and

$$\Lambda = \left[\begin{array}{cc} \Omega_{XX}^{-1} & \cdots \end{array}\right],$$

where Ω_{XX} is the probability limit of $\frac{1}{n}X'X$. The submatrix of Λ equal to Ω_{XX}^{-1} gives the sensitivity of $\hat{\beta}$ to $\frac{1}{n}Y'X$.

[13] Gelman and Imbens (2014) suggest that similar analysis may be useful more broadly: "Most, if not all, estimators for average treatment effects... can be written as the difference between two weighted averages... In those cases it is useful to inspect the weights in the weighted average expression... to assess whether some units receive excessive weight in the estimators" (p. 6). Using this example as a template, our sensitivity measure provides a low-cost way to compute analogues of these weights for arbitrary models, including those where deriving them algebraically would be difficult or infeasible.

high-order polynomials with large bandwidths. Their analysis of weights is closely related to our measure of sensitivity. To see this, let $\hat{\theta} = \left[\begin{array}{cc} \hat{\beta}_0^- & \hat{\beta}_0^+ \end{array} \right]'$ and suppose that X takes on discrete values, so that $X_i \in \{x_1, ..., x_J\}$. Finally, let $\hat{\gamma} = \left[\begin{array}{ccc} \overline{Y}_1 & ... & \overline{Y}_J \end{array} \right]'$, where $\overline{Y}_j$ is the mean of Y for observations such that $X_i = x_j$. Then the element of Λ corresponding to the sensitivity of $\hat{\beta}_0^+$ to $\overline{Y}_j$ is

$$\Lambda_j^+ = s_j \Omega_{XX+}^{-1} \left[\begin{array}{c} 1 \\ x_j \\ \vdots \\ x_j^K \end{array} \right],$$

where s_j is a row vector with first element equal to the share of positive X values in bin j and other elements equal to zero, Ω_{XX+} is the probability limit of $\frac{1}{N_+} \sum_{i:X_i \geq 0} \left[\begin{array}{cccc} 1 & X_i & ... & X_i^K \end{array} \right]' \left[\begin{array}{cccc} 1 & X_i & ... & X_i^K \end{array} \right]$, and N_+ is the number of observations with $X_i \geq 0$. This is analogous to the expression for the weights derived in section 2.1 of Gelman and Imbens (2014).[14]

3 Interpretation

3.1 Robustness to Model Misspecification

Empirical models in economics often depend on a large number of assumptions. Some may be justified on economic grounds, while others are "whimsical assumptions" made purely for convenience or tractability (Leamer 1983). Knowing how estimates depend on specific data features provides insight into the relative importance of these assumptions, and can guide a reader who holds a prior over possible violations toward more accurate inference.

Conley et al. (2012) develop this logic for linear IV models where the assumed orthogonality of the instruments and the error term may not hold. They index violations of orthogonality by a vector that we will call m, and consider the generalized model given by:

$$Y = X\beta + Zm + \varepsilon, \tag{4}$$

where X are endogenous variables, Z are instruments, ε are unobservables, and ε is orthogonal to Z. The standard setup corresponds to $m = 0$. They consider a decision maker whose beliefs about m can be described by a proper prior distribution, which is "local to zero" in that it becomes concentrated around zero at rate $\sqrt{n}$. They formalize this by letting $m = \eta / \sqrt{n}$, where η is drawn by nature from a known prior distribution P in a first stage. Under these assumptions, they derive

[14] The differences are that (i) we have Ω_{XX+}^{-1} in place of its sample analogue and (ii) we multiply by the scaling factor s_j (since Gelman and Imbens define weights for each observation).

an expression that relates the asymptotic distribution of the 2SLS estimator (integrating over the uncertainty in η) to the prior distribution P.[15]

Our sensitivity measure allows us to extend this approach to a much larger class of models. The key requirement is that the potential misspecification affects a sufficient vector of sample statistics $\hat{\gamma}$ in a known way. A natural class of applications will be those in which $\hat{\theta}$ is a moment estimator, $\hat{\gamma}$ is the vector of estimation moments, and small deviations from identifying assumptions translate into small violations of the moment conditions. The linear IV model is in this class because it can be expressed as a special case of GMM.

Towards a general analysis, let $F_n(\cdot|\theta) = \times_n F(\cdot|\theta)$ now be the model *assumed* by the researcher. The data are in fact drawn from a distribution F_n^*, which may be different from F_n. We define the set of possible alternatives by a family $\{F^*(\cdot|\theta,m)\}_{m\in\mathbb{R}^d}$ of distributions on $\mathscr{X}$ with dominating measure μ and densities $\{f^*(\cdot|\theta,m)\}_{m\in\mathbb{R}^d}$. We normalize the index m so that $F(\cdot|\theta) = F^*(\cdot|\theta,0)$, we let $F_n^*(\cdot|\theta,m_n) = \times_n F^*(\cdot|\theta,m_n)$, and we assume that deviations are "local to zero" in the sense that $m_n = \eta/\sqrt{n}$ for some $\eta \in \mathbb{R}^d$. In all results and examples in this section, we assume sufficient regularity conditions on the densities so that $F_n^*\left(\cdot|\theta_0,\frac{\eta}{\sqrt{n}}\right)$ and $F_n(\cdot|\theta_0)$ are mutually contiguous for any η.[16] We abbreviate $F_n^*\left(\cdot|\theta_0,\frac{\eta}{\sqrt{n}}\right)$ as F_n^* from here on.

Suppose now that we have $\hat{\gamma}$ sufficient for $\hat{\theta}$ under F_n. From proposition 1, we know that the limiting distribution of $\sqrt{n}\left(\hat{\theta}-\theta_0\right)$ under F_n^* will be the same as the limiting distribution of $\sqrt{n}\Lambda(\hat{\gamma}-\gamma_0)$. Therefore, if we know how varying η affects the asymptotic distribution of the moments $\hat{\gamma}$, we can use sensitivity Λ to translate this into effects on the asymptotic distribution of $\hat{\theta}$, and adjust inference accordingly.

A natural application is to the case where $\hat{\theta}$ is an MDE with $\hat{\gamma} = \hat{g}(\theta_0)$. Suppose that for all η in some open neighborhood of zero, $\sqrt{n}\left(\hat{\theta}-\theta_0,\hat{g}(\theta_0)\right)$ converges in distribution to a well-defined random variable under F_n^*. We then have the following result:

Proposition 3. *Suppose that, under F_n^*, $\sqrt{n}\hat{g}(\theta_0) \xrightarrow{d} \tilde{\gamma} + L\eta$, where L is a matrix of constants and $\tilde{\gamma} \sim N\left(0,\Sigma_{\gamma\gamma}\right)$ is the limit under the assumed model $F_n(\cdot|\theta_0)$. If η is drawn independently from the prior $P = N(\mu_\eta,\Sigma_\eta)$, we have that under F_n^*:*

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left(\Lambda L\mu_\eta, \Sigma_{\theta\theta} + \Lambda L\Sigma_\eta L'\Lambda'\right),$$

[15] Conley et al. (2012) describe this frequentist approach as a "large sample approximation" to inference under model uncertainty. They also show how to do a full Bayesian analysis, which requires specifying priors not only over m but also over all other parameters of the model. Guggenberger (2012) also analyzes misspecified linear IV models under the assumption that violations of the orthogonality condition become small at rate $\sqrt{n}$.

[16] Abbreviate $f^*(\cdot|\theta,m)$ by f_m and abbreviate its partial derivative with respect to m, if it exists, by $\dot{f}_m$. Following Van der Vaart (1998) theorem 7.2 and lemma 7.6, and Le Cam's lemmas, a sufficient condition for mutual contiguity is that (i) $\sqrt{f_m}$ is continuously differentiable in m at $m=0$ for every $X \in \mathscr{X}$, and (ii) the elements of the information matrix $I_m = \int\left(\dot{f}_m/f_m\right)\left(\dot{f}_m'/f_m\right) f_m d\mu$ are well-defined and continuous in m.

where $\Lambda = -\left(G'W_gG\right)^{-1}G'W_g$ is sensitivity as defined above.

Proof. From proposition 2, we know that the asymptotic distribution of $\sqrt{n}\left(\hat{\theta}-\theta_0\right)$ under F_n^* must be the same as the asymptotic distribution of $\sqrt{n}\Lambda\hat{g}\left(\theta_0\right)$. The result then follows because $\Lambda L\eta \sim N\left(\Lambda L\mu_\eta, \Lambda L\Sigma_\eta L'\Lambda'\right)$, $\Lambda\tilde{\gamma} \sim N\left(0,\Sigma_{\theta\theta}\right)$, and $\tilde{\gamma}$ and η are independent. □

According to this result, if we are willing to state a prior over misspecification that maps easily into violations of moment conditions $\hat{\gamma} = \hat{g}\left(\theta_0\right)$, we can use our sensitivity measure to adjust inference for θ accordingly. We now illustrate several cases in which the hypothesis of proposition 3 is satisfied.

Example 3. (Linear IV) Under standard assumptions, we can represent 2SLS as an MDE with $\hat{g}\left(\theta\right) = \frac{1}{n}Z'\left(Y-X\theta\right)$ and weight matrix $W_g = \left(Z'Z\right)^{-1}$. Let $F^*\left(\cdot|\theta,m\right)$ be the misspecified model of equation (4). Then under F_n^* we have $\sqrt{n}\hat{g}\left(\theta_0\right) \xrightarrow{d} \tilde{\gamma}+L\eta$ where $\tilde{\gamma} \sim N\left(0,\Sigma_{\gamma\gamma}\right)$ and $L = \Omega_{ZZ} \equiv \text{plim}\left(\frac{1}{n}Z'Z\right)$. Noting that $\Lambda = \left(\Omega_{ZX}'\Omega_{ZZ}^{-1}\Omega_{ZX}\right)^{-1}\Omega_{ZX}'\Omega_{ZZ}^{-1}$ where $\Omega_{ZX} = \text{plim}\left(\frac{1}{n}Z'X\right)$, proposition 3 implies

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left(A\mu_\eta, V_{2SLS}+A\Sigma_\eta A'\right)$$
$$A = \left(\Omega_{ZX}'\Omega_{ZZ}^{-1}\Omega_{ZX}\right)^{-1}\Omega_{ZX}'$$

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

Example 4. (General IV) Suppose that $\hat{\theta}$ is an MDE with moment conditions $\hat{g}\left(\theta\right) = \frac{1}{n}Z'\xi\left(\theta\right)$, where $\xi\left(\theta\right)$ is a residual that can be computed from the data given θ, and under the assumed model we have $\text{E}\left(\xi\left(\theta_0\right)|Z\right) = 0$. This includes nonlinear IV as well as simulated moment estimators such as Berry et al. (1995). Suppose that we entertain small violations of the moment conditions, so that under alternative model $F^*\left(\cdot|\theta,m\right)$, we have $\xi\left(\theta\right) = \xi^*\left(\theta\right)+Zm$. (Since $m=0$ in the assumed model, we must have $\text{E}\left(\xi^*\left(\theta_0\right)|Z\right) = 0$.) Then under F_n^* we again have $\sqrt{n}\hat{g}\left(\theta_0\right) \xrightarrow{d} \tilde{\gamma}+L\eta$ where $\tilde{\gamma} \sim N\left(0,\Sigma_{\gamma\gamma}\right)$ and $L = \Omega_{ZZ}$, and so proposition 3 implies

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left(A\mu_\eta, \Sigma_{\theta\theta}+A\Sigma_\eta A'\right)$$

$$A = -\left(G'W_gG\right)^{-1}G'W_g\Omega_{ZZ}.$$

Conley et al. (2012) thus generalizes to general IV models, with the asymptotic variance $\Sigma_{\theta\theta}$, Jacobian G, and weight matrix W_g replacing their linear model analogues.

Example 5. (Classical Minimum Distance) Suppose that $\hat{\theta}$ is a classical minimum distance estimator—that is, an MDE with $\hat{g}\left(\theta\right) = s\left(\theta\right)-\hat{s}$, where $\hat{s}$ is a vector of data moments independent of θ and $s\left(\theta\right)$ is a vector of corresponding model analogues independent of the data. For example, in our hypothetical life-cycle consumption model, $\hat{s}$ might be the average growth of consumption and the covariance of consumption with income growth, and $s(\theta)$ the expected value of these moments under the model. Suppose that utility shocks in the assumed model are independent of both income and age, but that we entertain small violations of these assumptions such that under alternative model $F^{*}(\cdot|\theta,m)$, we have $s(\theta) = s^{*}(\theta) + m$. Then under F_n^{*} proposition 3 applies with L equal to the identity matrix.

3.2 Identification

Figure 1 shows the dramatic increase in the number of articles published in top economic journals containing a claim that some estimator is "identified by" some feature of the data. In 2013, the *American Economic Review* published 15 empirical papers that include structural models; of these, 11 contain a section or subsection with "identification" in the title, while two others provide similar discussion without breaking it out into a separate subsection.[17] Consistent with figure 1, these discussions typically relate specific variation or data features to the identification of specific parameters.[18]

We can interpret these discussions as efforts to build a constructive argument for identification in the sense defined by Matzkin (2007; 2013). Consider the general class of models $\tilde{F}(\cdot|\zeta)$, where $\zeta \in Z$ is a possibly infinite-dimensional vector of functions and/or distributions of unobservables. Two primitives ζ and ζ' are *observationally equivalent* if $\tilde{F}(\cdot|\zeta) = \tilde{F}(\cdot|\zeta')$. An *economic quantity* c (e.g., a vector of elasticities) is a functional of ζ,[19] and is *identified* if for any observationally equivalent ζ and ζ', $c(\zeta) = c(\zeta')$. Because any feature $\gamma \in \mathbb{R}^J$ of the population distribution of the data can be written as a functional $\gamma(\zeta)$, it is natural to say that c is *identified by* γ if for any primitives ζ and $\zeta' \in Z$ such that $\gamma(\zeta) = \gamma(\zeta')$, we have $c(\zeta) = c(\zeta')$.

If economic quantity c is identified by data features γ, there must be some function Φ such that $c = \Phi(\gamma)$. A natural estimator for c is then $\hat{c} = \Phi(\hat{\gamma})$. As Matzkin (2013) stresses, providing a constructive proof of identification and then estimating the quantities of interest from the mapping Φ offers two key advantages: the estimates do not depend on arbitrary functional form or distributional assumptions, and the transparent mapping from data to estimates makes it easier for the reader to assess the credibility of the results.

[17] The online appendix lists these articles and shows how we classify them.

[18] For example, Barseghyan et al. (2013) write "given three or more deductible options, it is exogenous variation in premiums for a fixed [claim probability] that allows us to pin down [the coefficient of absolute risk aversion] and [the probability distortion function]" (p. 2511). Fan (2013) writes that "Identification of [the diminishing utility parameter] comes from the variation in the number of newspapers in a county" (p. 1610). Kawai and Watanabe (2013) write "we use the systematic difference between the predicted vote share and the actual vote share to partially identify the fraction of strategic voters" (p. 643).

[19] Matzkin (2007; 2013) refers to such quantities as "features" of a model, and uses a definition that includes the case where the object of interest is a distribution function or other high-dimensional object. We use different language to avoid confusion with the term "features of the data" which we use as a synonym for sample statistics above.

We can think about the parametric model $F(\cdot|\theta)$, $\theta \in \mathbb{R}^P$ defined in section 2 as a restriction that $\zeta \in Z' \subset Z$ for some subset Z' that can be mapped one-to-one to the parameter space $\mathbb{R}^P$. We can then write the quantity of interest as a function $c(\theta)$ of the model parameters, and write the estimator defined in section 2 as $c\left(\hat{\theta}\right)$.

While it has become common for applied researchers to discuss identification, it is also common for researchers to use estimators other than Φ. For example, it is common for authors to provide an argument for nonparametric identification of a general model $\tilde{F}(\cdot|\zeta)$, then estimate a parametric model $F(\cdot|\theta)$ due to data limitations or other practical concerns.[20]

Sufficiency and sensitivity can help assess whether the actual estimator $c\left(\hat{\theta}\right)$ resembles the hypothetical estimator $\Phi(\hat{\gamma})$ defined by the proof. If $c\left(\hat{\theta}\right)$ is equivalent to $\Phi(\hat{\gamma})$, then under suitable regularity conditions proposition 1 implies that $\hat{\gamma}$ is sufficient for $c\left(\hat{\theta}\right)$, and the sensitivity of $c\left(\hat{\theta}\right)$ to $\hat{\gamma}$ is equal to the partial derivative of Φ at γ_0. If these properties are at least approximately satisfied, then we may have some confidence that the attractive properties of the hypothetical estimator carry over to the actual one. If not, we might doubt the relevance of the discussion of identification for the properties of the estimator.

The following toy example provides a concrete illustration.

Example 6. (Standard Deviation of a Random Variable) Suppose the economic quantity of interest c is the population standard deviation σ_X of a scalar random variable X. A researcher shows that c is nonparametrically identified, with the nonparametric estimator $\hat{c}$ equal to the sample standard deviation $\hat{\sigma}_X$. The model she actually takes to the data makes the additional assumption that X is exponentially distributed. The actual estimator $c\left(\hat{\theta}\right)$ is the MLE for this parametric model.

How similar is the behavior of this parametric estimator to the hypothetical nonparametric one? Not very, as it turns out. The MLE of σ_X is in fact the sample *mean* of X, not the sample standard deviation. This estimator is efficient when the true model is in fact exponential, in which case the population mean and standard deviation are equal. The estimator is plainly very sensitive to model misspecification.

Suppose that we could not interrogate the nature of the MLE analytically, but that we used sufficiency and sensitivity to determine how it relates to the hypothetical nonparametric estimator. We would see immediately that the sufficiency of $\hat{\sigma}_X$ is not one as we would have expected, but substantially less (for example, if X is exponential, sufficiency of $\hat{\sigma}_X$ is $\frac{1}{2}$; if X is normal, sufficiency of $\hat{\sigma}_X$ is 0). If we explored further and expanded $\hat{\gamma}$ to include the sample mean, we would see that sensitivity is one for the mean and zero for the standard deviation, making clear how the maximum likelihood estimator actually takes information from the data.

[20] Einav et al. (2013) write, "Our actual data depart from the ideal data.... We thus make additional parametric assumptions to aid us in identification" (p. 201). Kawai and Watanabe (2013) write, "While our identification argument does not rely on the particular functional form [of preferences], our estimation does impose these functional forms" (p. 639).

Now suppose the researcher had instead assumed that X is normally distributed. The MLE is now the sample standard deviation. Sufficiency and sensitivity will reveal that the estimator corresponds to the nonparametric ideal.

This example is of course trivial, but it serves as a useful metaphor for issues that arise in applied research. There are always many possible sources of identification, and seemingly innocuous functional form assumptions can add additional sources that are not always apparent. Trying to intuit how an estimator works based on contemplation of modeling assumptions may be difficult; sufficiency and sensitivity provide a way to lower this cost. In the applications in sections 5 and 6, we will show that authors' discussions of identification line up closely with the patterns of sufficiency and sensitivity in some cases, and diverge in others.

3.3 Descriptive Statistics

It is common for researchers to discuss the relationship between structural parameters and model-free "descriptive statistics."[21] In principle, indirect inference makes it possible to base estimation solely on such descriptive statistics (Gourieroux et al. 1993; Smith 1993). In practice, either for econometric or computational reasons, researchers often choose not to base estimation directly on descriptive statistics. In such cases the link between parameter estimates and descriptive statistics is typically not made precise.

Sensitivity and sufficiency quantify the relationship between descriptive statistics and parameter estimates. Suppose that $\hat{\theta}_p$ estimates a structural parameter of interest and that $\hat{\gamma}_j$ is a descriptive statistic—say, a regression coefficient—that seems intuitively useful for estimating parameter p. If sufficiency is high and sensitivity matches intuition, the descriptive analysis may be a good guide to understanding how the full model takes information from the data. If sufficiency is low, or sensitivity has an unexpected sign, this is less likely to be true. In the applications in sections 5 and 6, we will show cases where descriptive statistics successfully "emulate" structural estimates, and others where they do not.

Whenever sufficiency is less than one, interpretation of sensitivity poses challenges analogous to the problem of omitted variables bias in standard regression analysis. Because some features of the data important for $\hat{\theta}$ are omitted from $\hat{\gamma}$, sensitivity values will be determined in part by the correlation of the included features with the omitted ones. The interpretation of sensitivity must then rely on priors about the importance of such correlation.

[21] Einav et al. (2013), for example, relate the identification of the moral hazard parameters in their model to a preceding difference-in-difference analysis in a section called "Descriptive Evidence of Moral Hazard" (p. 192). Lim (2013) relates the identification of key parameters in her model to evidence contained in a data plot that is "not dependent on any particular modeling decision or estimated parameter values of the model" (p. 1378).

4 Estimation of Sensitivity and Sufficiency

In this section we show that it is easy to estimate sufficiency and sensitivity even for computationally difficult models. We focus on the case in which $\hat{\theta}$ is an MDE, an encompassing class that includes GMM and MLE. We do not explicitly discuss the case in which the magnitude of interest is a counterfactual or the statistics of interest are transformations of $\hat{\gamma}$, but we note that the results below extend immediately to those cases following remark 1.

We generally assume that the researcher has in hand consistent estimators for objects such as the weight matrix W_g and Jacobian G of an MDE. Note that if such an estimator is consistent under the assumed model F_n, it will remain consistent under the contiguous perturbations F_n^* considered in section 3.1.

We focus here on computing point estimates of sensitivity and sufficiency. If one wishes to also compute asymptotic confidence intervals for these measures, it is typically straightforward to do so using a bootstrap. To illustrate, the online appendix reports confidence intervals for the sufficiency values in our application to Berry et al. (1995) (section 5.2 below) and for the sufficiency and sensitivity values in our application to Mazzeo (2002) (section 6.2 below).

4.1 Sensitivity to Moments

If Λ is sensitivity to moments, we know from remark 2 that $\Delta = 1$ and $\Lambda = -\left(G^{'}W_gG\right)^{-1}G^{'}W_g$. By assumption the researcher possesses $\hat{W}_g$, a consistent estimate of W_g. A consistent estimate $\hat{G}$ of G is typically in hand to estimate the asymptotic variance of $\hat{\theta}$.[22] Therefore in typical applications estimating Λ imposes no additional computational burden beyond the estimation of the asymptotic variance.

Remark 3. If $\hat{\theta}$ is an MDE and $\hat{G}$ is a consistent estimate of G then $\hat{\Lambda} = -\left(\hat{G}^{'}\hat{W}_g\hat{G}\right)^{-1}\hat{G}^{'}\hat{W}_g$ is a consistent estimate of sensitivity to moments. If the researcher has computed a plug-in estimator of $\text{Var}\left(\tilde{\theta}\right)$, then computing $\hat{\Lambda}$ requires only matrix algebra and no additional simulation or estimation.

4.2 Sensitivity to Descriptive Statistics

If Λ is not sensitivity to moments, then the most convenient way to estimate Λ depends on how $\hat{\gamma}$ is defined. We assume throughout that $\hat{\gamma}$ is also an MDE, which means it could include first or second moments, smooth functions of estimation moments, regression coefficients, and many other candidate statistics.

Let $\hat{m}(\gamma)$, M, and W_m denote the analogues of $\hat{g}(\theta)$, G, and W_g respectively that are used to estimate $\hat{\gamma}$. We assume conditions so that $\hat{g}(\theta)$ and $\hat{m}(\gamma)$ can be "stacked" to form an MDE $\left(\hat{\theta}, \hat{\gamma}\right)$,

[22] In CMD or SMD where $\hat{g}(\theta) = \hat{\pi} - h(\theta)$, $H = -G$ where H is the Jacobian of $h()$ at the true value θ_0.

in particular that $\hat{g}(\theta_0)$ and $\hat{m}(\gamma_0)$ are jointly asymptotically normal with variance Ω. We let Ω_{gg}, Ω_{mm}, and Ω_{gm} denote the sub-matrices of Ω corresponding to the asymptotic variance of $\hat{g}(\theta_0)$, the asymptotic variance of $\hat{m}(\gamma_0)$, and the asymptotic covariance of $\hat{g}(\theta_0)$ and $\hat{m}(\gamma_0)$ respectively.

Under these assumptions, it is straightforward to show that

$$\Sigma_{\theta\gamma} = \left(G'W_gG\right)^{-1}G'W_g\Omega_{gm}W_mM\left(M'W_mM\right)^{-1}.$$

Standard plug-in estimators $\hat{\Sigma}_{\theta\theta}$ and $\hat{\Sigma}_{\gamma\gamma}$ are typically available for $\Sigma_{\theta\theta}$ and $\Sigma_{\gamma\gamma}$. If we can construct an estimator $\hat{\Sigma}_{\theta\gamma}$ for $\Sigma_{\theta\gamma}$, we can form consistent estimators $\hat{\Lambda} = \hat{\Sigma}_{\theta\gamma}\hat{\Sigma}_{\gamma\gamma}^{-1}$ and $\hat{\Delta}_p = \left(\hat{\Lambda}\hat{\Sigma}_{\gamma\gamma}\hat{\Lambda}'\right)_{pp} / \left(\hat{\Sigma}_{\theta\theta}\right)_{pp}$ for Λ and the elements of Δ.

Of the components of $\Sigma_{\theta\gamma}$, W_g and W_m are consistently estimated by $\hat{W}_g$ and $\hat{W}_m$ which are in hand from estimation, and G and M are consistently estimated by the sample analogues $\hat{G} = G\left(\hat{\theta}\right)$ and $\hat{M} = M(\hat{\gamma})$. All that remains is to estimate Ω_{gm}. In cases such as CMD or SMD, it is common to use a bootstrap to estimate Ω_{gg}; in such cases the same bootstrap can typically be used to estimate Ω_{gm}.

Remark 4. If $\hat{\theta}$ and $\hat{\gamma}$ are MDEs and the researcher has computed plug-in estimators of $\mathrm{Var}\left(\tilde{\theta}\right)$ and $\mathrm{Var}(\tilde{\gamma})$ then computing a consistent estimate $\hat{\Lambda}$ requires only computing a consistent estimate $\hat{\Omega}_{gm}$ of the asymptotic covariance of the moment conditions.

An important special case is when $\hat{\theta}$ and $\hat{\gamma}$ are both estimated via GMM, a case that includes MLE (Hansen 1982). Then $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta)$ and $\hat{m}(\gamma) = \frac{1}{n}\sum_{i=1}^{n} m(z_i, \gamma)$ for i.i.d. data z_i and functions $g(z, \theta)$ and $m(z, \gamma)$ satisfying $\mathrm{E}(g(z, \theta_0)) = \mathrm{E}(m(z, \gamma_0)) = 0$. In this case a consistent estimator for Ω_{gm} is $\hat{\Omega}_{gm} = \frac{1}{n}\sum_{i=1}^{n} g\left(z_i, \hat{\theta}\right) m(z_i, \hat{\gamma})'$.[23]

An alternative representation of the estimator for $\hat{\Lambda}$ is useful for building intuition in this case.

Definition. Let $\tilde{g}_i = -\left(\hat{G}'\hat{W}_g\hat{G}\right)^{-1}\hat{G}'\hat{W}_g g\left(z_i, \hat{\theta}\right)$ *and define* $\tilde{m}_i$ *analogously. These* $(P \times 1)$ *and* $(J \times 1)$ *vectors are the* **influence** *of observation i on* $\hat{\theta}$ *and* $\hat{\gamma}$ *respectively (Hampel et al. 1986; Ronchetti and Trojani 2001).*

Intuitively, through the first-order condition $\tilde{g}_i$ tells us how much (and in what direction) observation i affects $\hat{\theta}$. The same property holds for $\tilde{m}_i$. Then by regressing $\tilde{g}_i$ on $\tilde{m}_i$ we recover how the influence of an observation on $\hat{\gamma}$ relates to its influence on $\hat{\theta}$, and hence how $\hat{\gamma}$ and $\hat{\theta}$ are related under the data-generating process:

Proposition 4. *The transposed coefficient matrix* $\hat{\Lambda} = (\tilde{g}'\tilde{m})(\tilde{m}'\tilde{m})^{-1}$ *from a regression of* $\tilde{g}_i'$ *on* $\tilde{m}_i'$ *is a consistent estimator of the sensitivity* Λ *of* $\hat{\theta}$ *to* $\hat{\gamma}$*. The* R^2 *associated with the regression of the p-th element of* $\tilde{g}_i'$ *on* $\tilde{m}_i'$ *is a consistent estimator of* Δ_p.

[23] In the case of dependent data with a group structure, the estimator $\hat{\Omega}_{gm}$ could be replaced with an appropriate group-level analogue.

Proof. Let $\tilde{g}$ and $\tilde{m}$ denote the matrices whose rows are $\tilde{g}_i'$ and $\tilde{m}_i'$, respectively. The first statement follows from the continuous mapping theorem and the definition of sensitivity after noting that $\hat{\Lambda} = (\tilde{g}'\tilde{m})(\tilde{m}'\tilde{m})^{-1}$, $\frac{1}{n}\tilde{g}'\tilde{m} \xrightarrow{p} \Sigma_{\theta\gamma}$ and $\frac{1}{n}\tilde{m}'\tilde{m} \xrightarrow{p} \Sigma_{\gamma\gamma}$. The second statement follows from the continuous mapping theorem and definition of sufficiency after noting that:

$$R^2 = \frac{\left(\hat{\Lambda} \cdot \left(\frac{1}{n}\tilde{m}'\tilde{m}\right) \cdot \hat{\Lambda}'\right)_{pp}}{\left(\frac{1}{n}\tilde{g}'\tilde{g}\right)_{pp}}$$

and that $\frac{1}{n}\tilde{g}'\tilde{g} \xrightarrow{p} \Sigma_{\theta\theta}$. □

Example 7. (Sensitivity of MLE to Sample Mean) Suppose the data are $z_i \in \mathbb{R}^D$, with elements z_{di}, the parameter of interest θ is a scalar, and $\hat{\theta}$ is an MLE with likelihood function $f(z_i|\theta)$:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \ln f(z_i|\theta).$$

Suppose we wish to assess sensitivity to the means of the elements of z_i, so we define $\hat{\gamma} \equiv \bar{z} \equiv \frac{1}{n}\sum_{i=1}^{n} z_i$.

We can interpret $\hat{\theta}$ as a GMM estimator with moment functions $g(z_i|\theta) = \partial \ln f(z_i|\theta)/\partial\theta$, weight matrix $W_g = I$, and Jacobian $G(\theta) = E\left(\partial^2 \ln f(z_i|\theta)/\partial\theta^2\right)$. We can interpret $\hat{\gamma}$ as a GMM estimator with moment functions $m(z_i|\gamma) = z_i - \gamma$, weight matrix $W_m = I$, and Jacobian $M(\gamma) = -I$. We can consistently estimate Λ with the coefficients from a regression of the (scaled) score of observation i:

$$\tilde{g}_i = -\frac{1}{\hat{G}} \cdot \left.\frac{\partial \ln f(z_i|\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}$$

on the deviation from the mean of observation i:

$$\tilde{m}_i = (z_i - \bar{z}).$$

Intuitively, $\hat{\theta}$ is more sensitive to the mean of a particular variable z_{di} when observations with high values of z_{di} have high values of the score (holding the other elements of z_i constant). This approach is easily extended to look at the sensitivity of $\hat{\theta}$ to higher-order moments of the data.

4.3 Units of Measurement

We have noted that Λ has an interpretation as the probability limit of coefficients from a regression of $\hat{\theta}$ on $\hat{\gamma}$. As with any regression coefficients, the elements of Λ depend on the units of measurement of the regressors $\hat{\gamma}$. Determining which element of $\hat{\gamma}$ is most "important" for a given $\hat{\theta}_p$ therefore requires judgment. The problem of assessing the relative importance of regressors is age-old and no solution is satisfactory in all situations (Kim and Ferree 1981; Bring 1994; Gelman 2008). But it is helpful to have a default. For this we propose the analogue of the standardized regression coefficient:

Definition. *The* **standardized sensitivity** *of $\hat{\theta}_p$ to $\hat{\gamma}_j$ is*

$$\tilde{\Lambda}_{pj} = \Lambda_{pj}\sqrt{\frac{\mathrm{Var}\left(\tilde{\gamma}_j\right)}{\mathrm{Var}\left(\tilde{\theta}_p\right)}}.$$

Standardized sensitivity measures how much a one-standard-deviation change in the realization of $\tilde{\gamma}_j$ affects the expected value of $\tilde{\theta}_p$, fixing other elements of $\tilde{\gamma}$, in units of the standard deviation of $\tilde{\theta}_p$. If the elements of $\hat{\gamma}$ are asymptotically independent (i.e., if $\Sigma_{\gamma\gamma}$ is diagonal) then the matrix $\tilde{\Lambda}$ of standardized sensitivities is the correlation matrix of $\tilde{\theta}$ with $\tilde{\gamma}$.

An attractive property of standardized sensitivity is that it is invariant to changes in units. Formally, for vectors a, c and strictly positive diagonal matrices B, D the standardized sensitivity of $a + B\hat{\theta}$ to $c + D\hat{\gamma}$ is equal to the standardized sensitivity of $\hat{\theta}$ to $\hat{\gamma}$. This means that, for example, if we switched from measuring an element of $\hat{\gamma}$ in dollars to measuring it in euros, our conclusions about the relative importance of different moments would be unchanged.[24]

Comparisons in units of standard deviations will not always be appropriate or necessary. If two statistics are in comparable economic units, it may be meaningful to compare their unstandardized sensitivities directly. Nevertheless, abstracting from any particular context it seems attractive to have a unitless measure as a default, and we will report estimates of standardized sensitivity in all of our applications.

5 Main Applications

5.1 Life-cycle Consumption and Savings

Gourinchas and Parker (2002)

Our first application is to Gourinchas and Parker's (2002) model of life-cycle consumption. Gourinchas and Parker (2002) model the behavior of a consumer with a time-separable constant-relative-risk-aversion felicity function and a stochastic income process. The parameters of the income process are estimated in a first step and are taken as given in a second step, in which preference parameters are estimated from moments corresponding to mean consumption at different ages (adjusted for family size and business cycle shocks).

[24] There are other transformations of Λ, such as the matrix of partial correlations of $\tilde{\theta}_p$ with $\tilde{\gamma}_j$ (conditional on $\tilde{\gamma}_{\sim j}$), that would also exhibit this invariance property.

Figure 2 presents sensitivity estimates for the second-step model's two key preference parameters: the discount factor and the coefficient of relative risk aversion.[25] The plot reveals three periods of life with different implications for the parameter estimates. In the first period, roughly ages 26-36, and in the third period, roughly ages 62-65, higher consumption implies a higher discount factor and a lower coefficient of relative risk aversion. In the second period, roughly ages 37-61, higher consumption implies a lower discount factor and a higher coefficient of relative risk aversion.

A stylized economic intuition is as follows. The consumer saves for retirement and for precautionary reasons. The strength of retirement saving motives is governed by the discount factor, and the strength of precautionary motives by the coefficient of relative risk aversion. Both a higher discount factor and a higher coefficient of relative risk aversion predict more delay of consumption, i.e., lower consumption early in life and greater consumption later in life. The two parameters are separately identified because of their different quantitative implications.

In the first period of life, saving is primarily precautionary, so risk aversion matters comparatively more than discounting, and higher consumption is interpreted as evidence of low risk aversion. In the second period, saving is primarily for retirement, so discounting matters comparatively more, and higher consumption is interpreted as evidence of impatience. In the third period, retirement looms and income uncertainty has essentially vanished, so high consumption is evidence that the household has already accumulated substantial retirement wealth, i.e., that the household is patient.

The fact that the two plots are essentially inverse to one another arises because both a higher discount factor and a higher coefficient of relative risk aversion imply the same qualitative change in the consumption profile. Therefore a change in consumption at a given age that implies a high discount factor must be offset by a lower coefficient of relative risk aversion in order to hold consumption at other ages constant.

De Nardi et al. (2010)

De Nardi et al. (2010) model consumption and saving by retired, nonworking households with uninsurable mortality and medical expense risk. Households have a time-separable constant-relative-risk aversion felicity function and a consumption floor guaranteed by the government. The parameters of the mortality and medical expense processes are estimated in a first step and are taken as given in a second step, in which the discount factor, coefficient of relative risk aversion,

[25] The baseline specification in Gourinchas and Parker (2002) has two additional parameters, which govern a reduced-form retirement consumption function. We fix these 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.

and consumption floor are estimated using SMM from moments corresponding to median assets for different cohorts, ages, and permanent income levels. We use the results in remark 1 to compute the sensitivity of second-step parameters to the means of related groups of estimation moments.[26]

The first two plots in figure 3 present the sensitivity of the consumption floor and the coefficient of relative risk aversion to the mean of the asset holdings by income quintile. The consumption floor is sensitive primarily to the savings of households in the lowest income quintile: the less these households save, the greater is the inferred consumption floor. The coefficient of relative risk aversion rises with the savings of the rich and falls with the savings of the poor. This pattern matches closely the intuition in De Nardi et al. (2010):

> The coefficient of relative risk aversion is identified by differences in saving rates across the income distribution, in combination with the consumption floor. Low-income households are relatively more protected by the consumption floor and will thus have lower [variance of consumption growth] and hence weaker precautionary motives. The parameter helps the model explain why individuals with high permanent income typically display less asset decumulation (p. 59).

The third plot in figure 3 presents the sensitivity of the discount factor to the mean of the asset holding moments by age. As expected, the estimator interprets large asset holdings at younger ages as evidence of patience.

5.2 Automobile Demand

Our second application is to Berry et al.'s (1995) model of automobile demand. We follow Berry et al. (1995) closely, using their data and SMM procedure, with moments from both the demand and supply sides of the model.[27]

The estimation moments are derived from two sets of identifying assumptions. On the demand side, the model assumes that the expected unobserved quality ξ_j of car j is zero conditional on instruments z_j^d. Berry et al. (1995) construct z_j^d from a set of demand-side variables: a constant, a dummy for whether car j has air conditioning, and car j's horsepower-per-weight, miles-per-dollar of gasoline, and size. For each variable, z_j^d includes: (i) the value of the variable for car j; (ii) the sum of the variable across other cars produced by the firm that produces car j; (iii) the sum of

[26] The online appendix reports the standardized sensitivity of second-step parameters to the full set of (untransformed) estimation moments.

[27] We extract automobile data and guide our implementation using the GAUSS code for Berry et al. (1999), downloaded from the Internet Archive's April 2005 web capture of James Levinsohn's (now defunct) website at the University of Michigan. Table 1 from Berry et al. (1995) and table 2 from Berry et al. (1999) imply that the two use the same dataset. We used code from Petrin (2002), Dubé et al. (2012), and Knittel and Metaxoglou (2014) as additional references.

the variable across cars produced by rival firms. The demand-side moments are the product of ξ_j (computed as a residual inverted from market shares) with each element of z_j^d.

On the supply side, the model assumes that the expected unobserved cost component ω_j of car j is zero conditional on instruments z_j^s. Berry et al. (1995) construct z_j^s in the same way as z_j^d, but using instead a set of supply-side variables: a constant, a dummy for whether car j has air conditioning, a time trend, and the logarithms of car j's horsepower-per-weight, miles-per-gallon of gasoline, and size. In addition, z_j^s includes an excluded demand variable, miles-per-dollar of gasoline for car j (but not the sums of this variable across other cars). The supply-side moments are the product of ω_j (computed as a residual inverted from estimated marginal costs) with each element of z_j^s.

We first assess the relative importance of the demand and supply moments. For each parameter, we compute the sufficiency Δ_p of all demand moments together (that is, we set $\hat{\gamma}$ equal to the vector of demand moments) and of all supply moments together.[28] These results are presented in figure 4. A natural hypothesis is that the cost side parameters of the model depend on the supply moments while the utility parameters depend on the demand moments. This is partly true: the supply moments are indeed nearly sufficient for the cost parameters, but the utility parameters depend on both demand and supply moments, and some key parameters such as the price coefficient are primarily driven by the latter.[29]

We can also apply our method to analyze the relative importance of specific instruments in driving a key economic outcome of the model: the estimated markups. We define the counterfactual $c\left(\hat{\theta}\right)$ of interest to be the average estimated markup across all cars.[30] We define $\hat{\gamma}$ to be the complete set of estimation moments $\left(\begin{array}{cc} z_j^{d\prime}\xi_j & z_j^{s\prime}\omega_j \end{array}\right)$, but plot sensitivity only for moments involving the "excluded" instruments—i.e., those that do not enter the utility or cost j directly.[31]

These results are presented in figure 5. We find that markups are overall more sensitive to the supply moments than to the demand moments. The supply-side instruments that play the largest role are the number of other products produced by the same firm (i.e., the sum of the constant term across other products produced by the same firm), the gas mileage of these cars, and the number of products produced by rival firms. To build intuition, recall that the model is estimated using data from 1971-1990, a period that saw the large-scale entry of Japanese cars and a shift toward greater

[28] The sufficiency of demand moments and the sufficiency of supply moments need not sum to one because the two sets of moments are correlated.

[29] The online appendix reports the bootstrap confidence intervals of sufficiency Δ_p of the demand and supply moments for each parameter, using the percentiles of 70 block bootstrap replicates. The bootstrap results suggest that the relative importance of the supply moments for the marginal cost parameters is robust to sampling error.

[30] The markup is tightly related to the own-price elasticity, and, therefore, to the price coefficient. We show in the online appendix that the pattern of sensitivities of the average own-price elasticity and the price coefficient are similar to those we present here.

[31] The online appendix reports the complete standardized sensitivity matrix $\tilde{\Lambda}$, with values for all parameters and moments.

fuel economy. A possible interpretation, therefore, is that the model uses the changes in prices induced by the increased competitiveness and product differentiation as a key source of exogenous variation.

Finally, we apply our method to ask to what extent the relationship of moments to estimated elasticities in the full BLP model is well approximated by the relationship of moments to estimated elasticities in the aggregate logit version of the model. The latter model, which has no random coefficients, can be estimated by two-stage least squares. Berry et al. (1995) present estimates from this logit model as a point of departure. We define the counterfactuals $c\left(\hat{\theta}\right)$ to be mean elasticities of demand with respect to price and product attributes implied by the estimated parameters of the full BLP model, and transformed statistics $a\left(\hat{\gamma}\right)$ to be the same mean elasticities implied instead by the estimated parameters of the logit model. We compute sensitivity of each elasticity in the full model to the elasticities implied by the logit model.

These results are presented in figure 6. We find that sufficiency Δ_p is low for all elasticities, ranging from 0.02 for the air conditioning dummy to 0.13 for miles-per-dollar. Moreover, there is no systematic pattern in which the estimated demand elasticity to a particular attribute in the full model is primarily related to the estimated demand elasticity for that same attribute in the logit model. This suggests, consistent with the discussion in Berry et al. (1995), that carrying forward simple intuitions from the logit model is not a useful way to understand the full model.

6 Other Applications

6.1 Sensitivity to Moments

In this subsection we apply our measure of the sensitivity to moments to several empirical papers that use MDEs. In each case we obtain plug-in estimators $\hat{G}$, $\hat{W}_g$, and $\hat{\Omega}_{gg}$ either directly from the authors or from replication files posted by the authors. For papers that estimate multiple specifications we use the baseline or main specification reported in the paper.

We present our findings as plots of standardized sensitivity of all moments for each of a set of key parameters. In the online appendix we report the complete standardized sensitivity matrix $\tilde{\Lambda}$ for each paper. Each plot indicates the key moments that the authors highlight as important for the identification of the given parameter.

Goettler and Gordon (2011)

Goettler and Gordon (2011) model innovation in the market for a durable good. In the model, each of a set of firms maintains a position on a quality ladder. The chance of moving up the ladder is greater the more the firm invests in R&D and the further the firm is from the technological frontier. Marginal costs are increasing in product quality. Consumers value quality and treat the firms' products as vertically and horizontally differentiated. Both firms and consumers are forward-looking.

The model is estimated on data from the market for computer microprocessors. The main research question is whether the market leader, Intel, innovates more or less than it would in a counterfactual world without its main competitor, AMD. Seven parameters are estimated from 15 empirical moments using SMD.

Figure 7 presents results for three key parameters. We follow Goettler and Gordon (2011) in dividing moments into "demand-side" and "supply-side" groups.

The first two parameters that we consider are demand parameters: the price coefficient, which reflects the disutility of higher prices, and the quality coefficient, which reflects the utility from higher quality. Regarding these parameters Goettler and Gordon (2011) write:

> The demand-side parameters ([the price coefficient], [the quality coefficient], [the Intel fixed effect], and [the AMD fixed effect]) are primarily identified by the pricing moments, the Intel share equation moments, and the mean ownership quality relative to the frontier quality. The pricing moments respond sharply to changes in any of these four parameters. The market share equation is primarily sensitive to [the quality coefficient] and [the Intel fixed effect minus the AMD fixed effect]. The mean [upgrading moment] decreases if consumers upgrade more quickly and is akin to an outside share equation that identifies the levels of [the Intel fixed effect and the AMD fixed effect] (p. 1161).

In figure 7, we find that the price coefficient is primarily sensitive to the average prices of Intel and AMD. This is intuitive because Goettler and Gordon (2011) have a direct measure of marginal cost. Given the assumption of dynamically optimal pricing, the higher is the observed price, the less price-sensitive consumers are estimated to be. The quality coefficient is primarily sensitive to the potential upgrade gains, a measure of the difference between the average CPU quality of the computer stock and the frontier quality available. Again, this is intuitive: the more sensitive consumers are to quality, the more often consumers will upgrade their PCs and the smaller will be the gap between average and frontier quality.

The third parameter that we consider is the innovation spillover, a measure of the extent to which innovation is easier the further the firm lies inside the technological frontier. Goettler and Gordon (2011) write:

> The supply-side parameters ([Intel's innovation efficiency], [AMD's innovation efficiency], and [the innovation spillover]), which govern the investment process, are primarily identified by observed innovation rates, quality differences, and investment

> levels. The investment efficiencies are chosen such that the observed investment levels (per unit revenue) yield innovation at the observed rates. The [innovation spillover parameter] is chosen to match the mean difference in quality across firms: a high spillover keeps the qualities similar (p. 1161).

We find that the innovation spillover is very responsive to the mean quality difference as expected. However, it responds slightly more to the average Intel price, and in general is very responsive to demand moments.

DellaVigna et al. (2012)

DellaVigna et al. (2012) model a household's charitable giving. In the model, a household may give to charity either out of altruism or because of social pressure. DellaVigna et al. (2012) conduct a field experiment in which they solicit charitable donations door-to-door. In some treatments they alert the household in advance that they will be coming to solicit. Households' response to this warning provides evidence on the motivations for giving and allows DellaVigna et al. (2012) to assess the welfare effects of charitable solicitations.

The model is estimated using 70 moments corresponding to the empirical frequencies of opening the door and giving different amounts of money in different treatment conditions. The model has 15 parameters estimated via CMD, using quadrature to approximate the expected value of the empirical moments as a function of the parameters.

Figure 8 presents results for two parameters. For each parameter, we show the standardized sensitivity to all moments, indicating key moments highlighted by the authors in red.

The first parameter, the baseline probability of being home, has a very simple relationship to the empirical moments. DellaVigna et al. (2012) explain that:

> The baseline probabilities of answering the door ... are identified by the observed probabilities of opening the door in treatments without flyer (p. 37).

Our plot bears out this discussion, showing that the empirical probabilities of being home in no-flyer conditions are the most important drivers of this parameter.

The second parameter, the social cost of giving less than $10 to the East Carolina Hazard Center (ECU), has a richer economic structure. DellaVigna et al. (2012) write:

> Finally, the social pressure ... is identified from two main sources of variation: home presence in the flyer treatment ... and the distribution of small giving (the higher the social pressure, the more likely is small giving and in particular bunching at [the threshold of $10]) (p. 38).

The authors define the social cost of giving \$$X$ as $S \times \max\{10 - X, 0\}$, where S is a parameter. We report sensitivity values for the cost of giving \$0, which is $10S$. The sensitivity values closely match the authors' discussion: Giving at the \$10 threshold increases the inferred level of social pressure, as does failing to open the door when warned in advance by a flyer. (The only exception is that giving less than \$10 is found to decrease rather than increase the estimated level of social pressure, perhaps because this level of giving does not allow the household to avoid feeling socially pressured.)

Nikolov and Whited (2014)

Nikolov and Whited (2014) model an infinitely lived firm whose manager makes decisions in discrete time about both the level of real investment and the extent of external financing. Capital is subject to depreciation and a convex adjustment cost. External financing imposes a real cost on the firm. The manager has an equity stake, a profit stake, and an ability to "tunnel" resources that are held as cash. The profit stake and the ability to tunnel lead to a divergence between the manager's interests and those of the shareholders.

The model has eight estimated parameters, corresponding to features of the production and investment technology, the external financial environment, and the manager's incentives. These parameters are estimated via SMM based on empirical moments that contain information on investment, financing, and compensation in a sample of firms.

Figure 9 presents standardized sensitivity of three select parameters. We follow Nikolov and Whited (2014) in dividing the moments loosely into "real" moments related to the investment decision, "financial" moments related to cash vs. external finance, and "incentives" moments related to managerial compensation and incentives.

The first parameter we study is the rate of depreciation of capital. Nikolov and Whited (2014) report that this parameter is identified by the mean rate of investment:

> The first two non financial or "real" moments are the first and second central moments of the rate of investment ... The first moment identifies the capital depreciation rate (p. 1899).

The economic logic here is that in a deterministic steady-state, the rate of investment is equal to the rate of depreciation of capital. The sensitivity values for the depreciation parameter bear out this intuition: the mean rate of investment is by far the most important moment in determining the estimated depreciation rate.

The second parameter that we study is the profit-sharing parameter, which corresponds to the fraction of after-tax operating earnings that accrue to the manager. Nikolov and Whited (2014) report that this parameter is identified principally by the average bonus paid to the firm's CEO:

> Finally, we discuss the identification of the profit-sharing parameter. ... First, without our data on ownership and compensation, we would have to infer the value of this parameter solely from firm decisions. In this case, a high value of [the profit-sharing parameter] implies low average profitability because the manager acts as if the firm is more profitable than it actually is and makes distorted investment decisions. However, many other parameters affect average profitability, so this moment alone cannot help identify [the profit-sharing parameter]. Fortunately, this parameter corresponds directly to one moment from our compensation data: the average bonus (p. 1900).

The authors also note that Tobin's q is useful in identifying this parameter. The sensitivity measure agrees with the authors' discussion. By far the most important driver of the estimated profit-sharing parameter is the average bonus. The average profit level is also relevant, and has the sign predicted by the model.

The third and final parameter that we study is the tunneling parameter, which corresponds to the fraction of the current stock and flow of cash that the manager consumes privately. Nikolov and Whited (2014) write:

> Not surprisingly, the moment that is most important for identifying resource diversion is the mean of Tobin's q: the greater resource diversion, the lower is q (p. 1900).

The sensitivity plot shows that greater Tobin's q does correspond to a lower inferred tunneling. Other moments also play an important role, however. Both lower investment and greater average profits imply greater tunneling. A possible explanation is that lower investment and greater profits imply a greater flow of resources, so for a fixed distribution to shareholders, managerial resource diversion must adjust to enforce the accounting identity that determines distributions to shareholders.

6.2 Sensitivity to Descriptive Statistics

Our final applications are cases in which the economic model is estimated via MLE. Formally, an MLE is an MDE in which the moments are first-order conditions. In the applications below these first-order conditions do not have a clear economic interpretation. We therefore define $\hat{\gamma}$ to be a set of descriptive statistics, typically those presented by the authors to provide a summary of key features of the data. We compute standardized sensitivity of key parameters or counterfactuals using the empirical influence components as described in section 4. (Recall that these calculations do not require re-estimation of the model.) Unlike in the case of sensitivity to moments, sufficiency need not be equal to one here.

Mazzeo (2002)

Mazzeo (2002) models entry into motel markets along US interstate highways. In the variant of Mazzeo's model that we consider, anonymous potential entrants to a local market make sequential decisions either not to enter the market, to enter as low quality, or to enter as high quality. Following the entry decision, firms realize payoffs that depend on observable market characteristics, the number of firms of each type, and a normally distributed profit shock that is specific to each firm type and local market and is unknown to the econometrician. Mazzeo (2002) estimates the model by MLE using data on the number and quality of motels along rural interstate highways in the United States.

Figure 10 reports the sensitivity of Mazzeo's (2002) estimates of the effect of market characteristics on firm profits.[32] Here we let $\hat{\gamma}$ be the coefficients from regressions of the number of low- and high-quality firms on observable market characteristics. Intuitively, we would expect the structural parameter governing the effect of a given characteristic on profitability to be tightly related to that characteristic's effect on the number of firms. We find that this is indeed the case, and that the regression coefficients are almost sufficient for the structural parameters. In all cases, knowing the regression coefficients would allow us to predict more than 80 percent of the variation in the structural parameter under the asymptotic distribution.

Gentzkow (2007)

Gentzkow (2007) uses survey data from a cross-section of individuals to estimate demand for print and online newspapers in Washington DC. A central goal of Gentzkow's (2007) paper is to estimate the extent to which online editions of papers crowd out readership of the associated print editions, which in turn depends on a key parameter governing the extent of print-online substitutability. We focus here on the substitutability of the print and online editions of the Washington Post.

Gentzkow (2007) exploits two features of the data to distinguish correlated tastes from true substitutability: (i) a set of variables—such as a measure of Internet access at work—that plausibly shift the utility of online papers but do not affect the utility of print papers; and (ii) a coarse form of panel data—separate measures of consumption in the last day and last seven days—that identifies stable individual preferences in a manner analogous to fixed or random effects in a linear model.

To capture these two features of the data, we define $\hat{\gamma}$ to consist of two components: (i) the coefficient from a 2SLS regression of last-five-weekday print readership on last-five-weekday online readership, instrumenting for the latter with the set of excluded variables such as Internet access at work; and (ii) the coefficient from an OLS regression of last-one-day print readership on last-one-

[32] The online appendix presents the corresponding numerical estimates of standardized sensitivity $\tilde{\Lambda}$ and sufficiency Δ_p, along with bootstrap-based confidence intervals and alternative point estimates computed from the bootstrap replicates.

day online readership controlling flexibly for readership of both editions in the last five weekdays. Each of these auxiliary models includes the standard set of demographic controls from Gentzkow (2007).

We define the counterfactual $c\left(\hat{\theta}\right)$ to be the change in readership of the Post print edition that would occur if the Post online edition were removed from the choice set (Gentzkow 2007, table 10).

The results are presented in figure 11. Sufficiency is 0.64, suggesting that these two features of the data capture much but not all of the variation that drives the counterfactual. Sensitivity is negative for both elements of $\hat{\gamma}$ as expected, reflecting the fact that a more positive relationship between print and online consumption implies less substitutability and thus a smaller gain of print readership. Finally, the results show that sensitivity to the panel variation is much larger than sensitivity to the IV variation, implying that the former is the more important driver of the estimated counterfactual.

Hendren (2013)

Hendren (2013) uses data on insurance eligibility and self-reported beliefs about the likelihood of different types of "loss" events (e.g., becoming disabled) to recover the distribution of underlying beliefs and rationalize why some groups are routinely denied insurance coverage. We focus here on Hendren's (2013) model of the market for long-term care insurance.

In Hendren's (2013) data, many respondents give "focal" responses of 0, 0.5, or 1 to survey elicitations of probabilistic beliefs. To allow for the possibility that these focal responses are not the respondents' actual beliefs, Hendren's (2013) model assumes that with some probability each respondent is a "focal point respondent" whose response is 0, 0.5, or 1, depending on which of three intervals her true beliefs falls into. The width of the intervals is controlled by a parameter called the "focal point window." For a given distribution of true beliefs, higher values of the focal point window make responses of 0 or 1 more likely relative to responses of 0.5.

Figure 12 reports the standardized sensitivity of the fraction focal point respondents and the focal point window. Here we let $\hat{\gamma}$ be shares of different responses to the survey elicitation, which together have sufficiency over 90 percent for the two parameters that we study. The results confirm our expectations based on Hendren's (2013) discussion.[33] The fraction focal respondents is highly sensitive to (and increasing in) the fraction of responses that are $\{0,1\}$ or 0.5. Fixing the share of responses equal to 0.5, increasing the share in $\{0,1\}$ increases the estimated focal window.

We can also compute sensitivity of the minimum pooled price ratio, formally a counterfactual $c\left(\hat{\theta}\right)$ that determines, as a function of model parameters, the range of preferences for which

[33] Hendren (2013) writes that "the fraction of focal point respondents... and the focal point window... are identified from the distribution of focal points and the loss probability at each focal point" (p. 1752).

insurance markets cannot exist. To study this counterfactual we define $\hat{\gamma}$ to be a vector of three descriptive statistics: the fraction of respondents who report a high (at or above 0.5) probability of needing long-term care and eventually need care, the fraction of respondents who report a high (at or above 0.5) probability of needing long-term care but do not eventually need care, and the fraction of respondents who report a low (below 0.5) probability of needing long-term care but do eventually need care. We find that the minimum pooled price ratio is increasing in the first of these and decreasing in the latter two, consistent with the intuition that insurance markets are more likely to unravel when respondents have more private information.[34]

7 Conclusions

We develop measures of the relationship between a parameter estimate and a set of given features of the data. The measures are easy to compute in common applications. Our measure of sensitivity has an interpretation as a measure of sensitivity to model misspecification, and our measures can be useful in complementing discussions of identification in empirical work.

An important limitation of our approach is that our measures are local, in the sense that they rely on the same asymptotic mechanics as commonly used formulae for standard errors. Conceptually, global exploration in a sample of a given size is straightforward. Consider the following exercise: (i) simulate or otherwise obtain data with dispersed values of $\hat{\gamma}$; (ii) estimate $\hat{\theta}$ on each dataset; and (iii) regress $\hat{\theta}$ on $\hat{\gamma}$ across these datasets. Such a procedure delivers global measures of sufficiency and sensitivity analogous to the local ones that we work with in this paper.

We focus on the local measures precisely because repeated simulation and estimation is often costly. We can, however, suggest approaches to minimizing this computational burden. First, for estimators whose cost of execution scales well with the size of the dataset, a researcher might use small-scale simulations to obtain the global measures and compare them to the local ones. If the two are similar, this adds confidence to the use of the local measures for sensitivity analysis. This is analogous to the common practice of using sampling experiments to validate inference from asymptotic standard errors.

Second, for cases where simulation from the data-generating process is cheaper than estimation, a researcher might simulate data from several possible values of θ and compute $\hat{\gamma}$ on the simulated data. Then, by regressing θ on $\hat{\gamma}$, one obtains a global analogue of sufficiency and sensitivity that does not require repeated model estimation. Developing the formal properties of more

[34] We present additional details in the online appendix. Note that the three descriptive statistics we study have low sufficiency (0.18) for the minimum pooled price ratio, indicating that these statistics do not capture most of the information in the data used to estimate the minimum pooled price ratio, and that our estimates of sensitivity should be interpreted with caution.

global approaches, and determining strategies for minimizing their computational costs, seem to be interesting areas for future work.

References

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.

Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua C. Teitelbaum. 2013. The nature of risk preferences: Evidence from insurance choices. *American Economic Review* 103(6): 2499-2529.

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

Berry, Steven and Philip A. Haile. 2014. Identification in differentiated products markets using market level data. *Econometrica* 82(5): 1749-1797.

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.

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

Chetty, Raj. 2009. Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. *Annual Review of Economics* 1: 451-488.

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

De Nardi, Mariacristina, Eric French, and John B. Jones. 2010. Why do the elderly save? The role of medical expenses. *Journal of Political Economy* 118(1): 39-75.

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.

Dubé, Jean-Pierre, Jeremy T. Fox, and Che-Lin Su. 2012. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica* 80(5): 2231-2267.

Einav, Liran, Amy Finkelstein, and Mark R. Cullen. 2010. Estimating welfare in insurance markets using variation in prices. *Quarterly Journal of Economics* 125(3): 877-921.

Einav, Liran, Amy Finkelstein, Stephen P. Ryan, Paul Schrimpf, and Mark R. Cullen. 2013. Selection on moral hazard in health insurance. *American Economic Review* 103(1): 178-219.

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. *Statis-*

tics 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. 2007. Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review* 97(3): 713-744.

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: 387-421.

Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 1986. *Robust statistics: The approach based on influence functions.* New York: John Wiley & Sons, Inc.

Hansen, Lars P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50(4): 1029-1054.

Hansen, Lars P. 2008. Generalized method of moments estimation. In S. Durlauf and L. Blume (eds.), *The New Palgrave Dictionary of Economics, Second Edition*. London, UK: Palgrave Macmillan.

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

Hendren, Nathaniel. 2013. Private information and insurance rejections. *Econometrica* 81(5): 1713-1762.

Huber, Peter J. and Elvezio M. Ronchetti. 2009. *Robust Statistics*. New York: Wiley.

Jaffe, Sonia and E. Glen Weyl. 2013. The first-order approach to merger analysis. *American Economic Journal: Microeconomics* 5(4): 188-218.

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

Kawai, Kei and Yasutora Watanabe. 2013. Inferring strategic voting. *American Economic Review* 103(2): 624-662.

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.

Knittel, Christopher R. and Konstantinos Metaxoglou. 2014. Estimation of random-coefficient demand models: Two empiricists' perspective. *Review of Economics and Statistics* 96(1): 34-59.

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

Lim, Claire S. H. 2013. Preferences and incentives of appointed and elected public officials: Evidence from state trial court judges. *American Economic Review* 103(4): 1360-1397.

Martin, Gregory J. and Ali Yurukoglu. 2014. Bias in cable news: Real effects and polarization. NBER Working Paper No. 20798.

Matzkin, Rosa L. 2007. Nonparametric identification. In Ch. 73 of J. Heckman and E. Leamer (eds.), *Handbook of Econometrics* 6B: 5307-5368. Amsterdam: Elsevier.

—. 2013. Nonparametric identification in structural economic models. *Annual Review of Economics* 5: 457-486.

Mazzeo, Michael J. 2002. Product choice and oligopoly market structure. *RAND Journal of Economics* 33(2): 221-242.

Morten, Melanie. 2013. Temporary migration and endogenous risk sharing in village India. Stanford mimeo.

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. and Daniel McFadden. 1994. Large sample estimation and hypothesis testing. In Ch. 36 of R. Engle and D. McFadden (eds.), *Handbook of Econometrics* 4: 2111-2245. Amsterdam: North-Holland.

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

Pakes, Ariel. 2003. Common sense and simplicity in empirical industrial organization. *Review of Industrial Organization* 23(3/4): 193-215.

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

Ronchetti, Elvezio and Fabio Trojani. 2001. Robust inference with GMM estimators. *Journal of Econometrics* 101(1): 37-69.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B (Methodological)* 45(2): 212-218.

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.

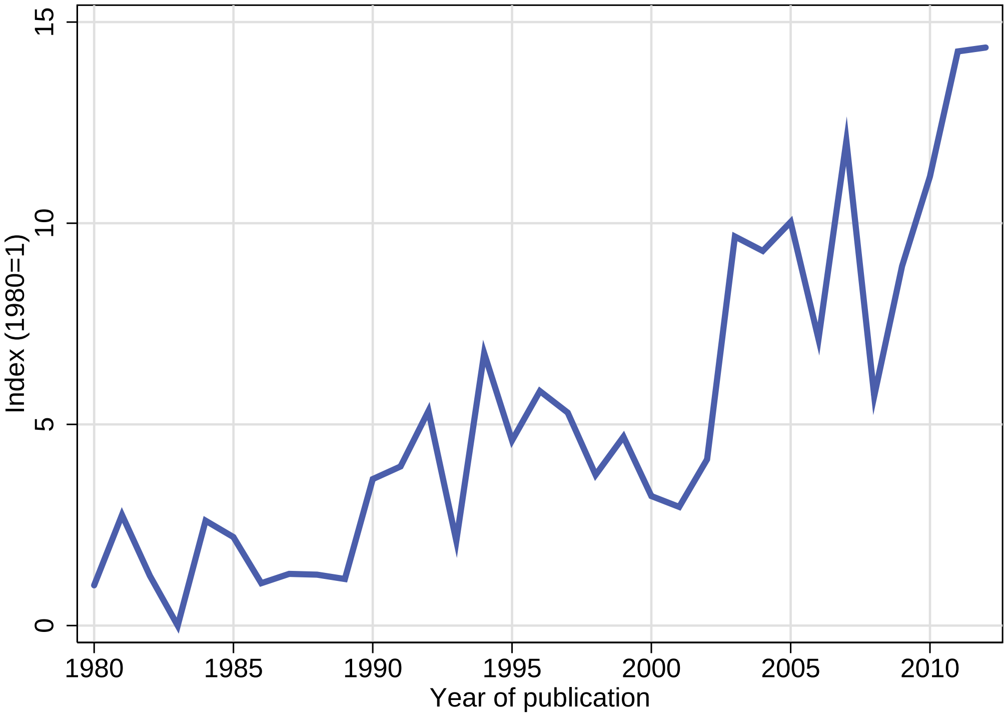
Smith, Anthony A. 1993. Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics* 8: S63-S84.

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.

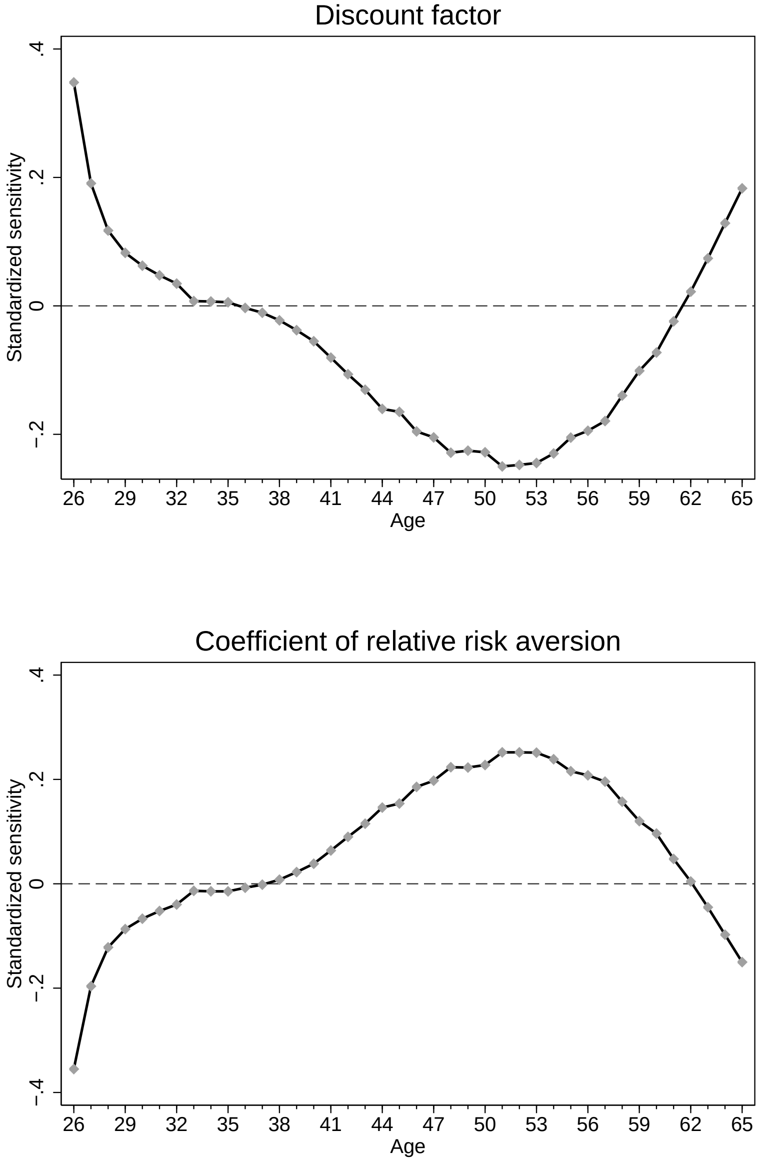
Vasnev, Andrey L. 2006. Local sensitivity in econometrics. In *CentER Dissertation Series*. Tilburg: CentER, Center for Economic Research.

Figure 1: Share of top journal articles containing the phrase “identified by”



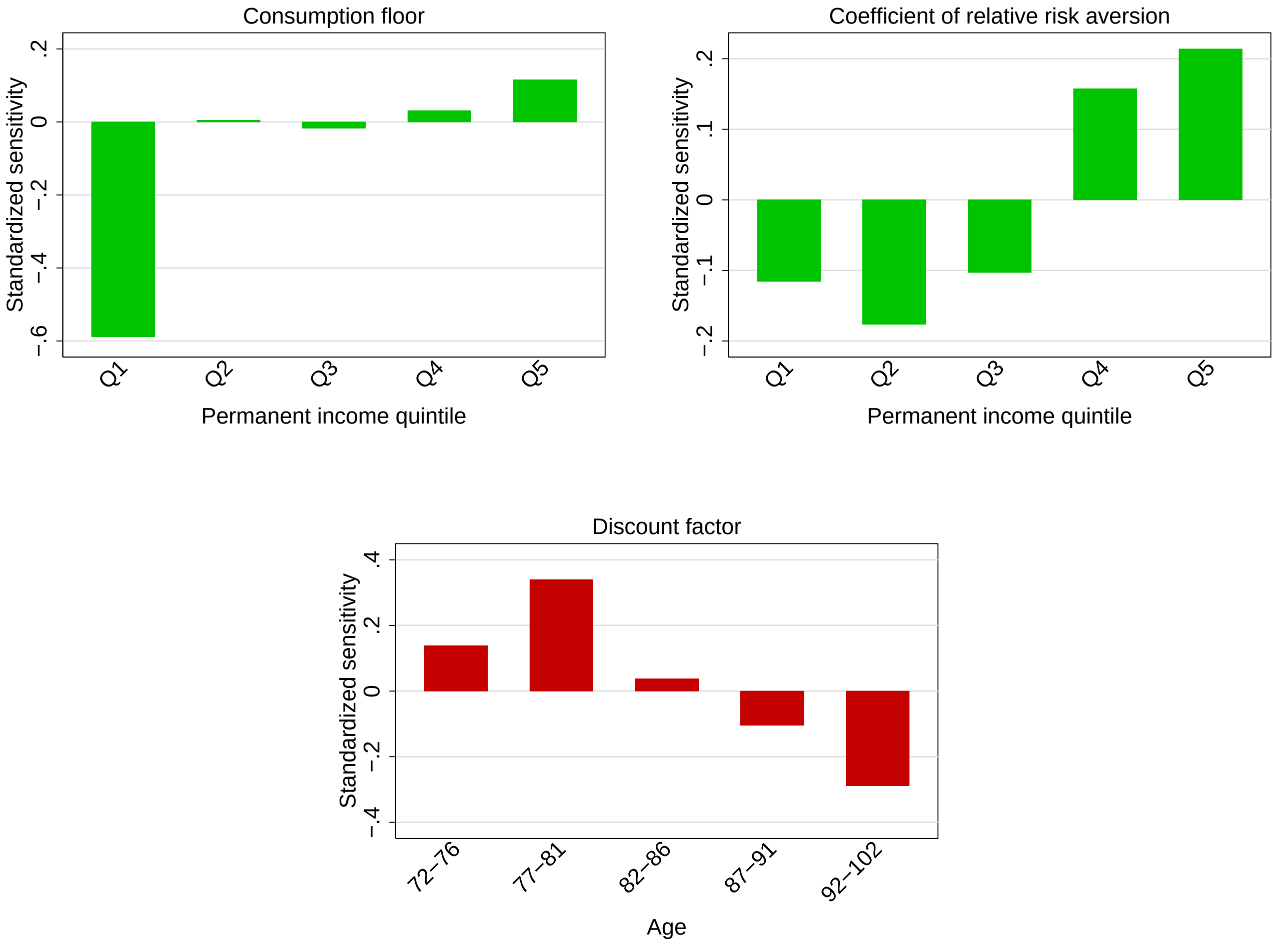
Notes: The plot shows an annual index of the share of articles published in the *American Economic Review,* the *Journal of Political Economy,* the *Quarterly Journal of Economics,* the *Review of Economic Studies,* and *Econometrica* containing the phrase “is identified by” or “are identified by” along with the word “data,” among all articles containing the word “data.” The index is constructed by dividing the share in each year by the share in 1980. Cases where the word “identified” is not used in the econometric sense are manually excluded.

Figure 2: Standardized sensitivity of select parameters in Gourinchas and Parker (2002)



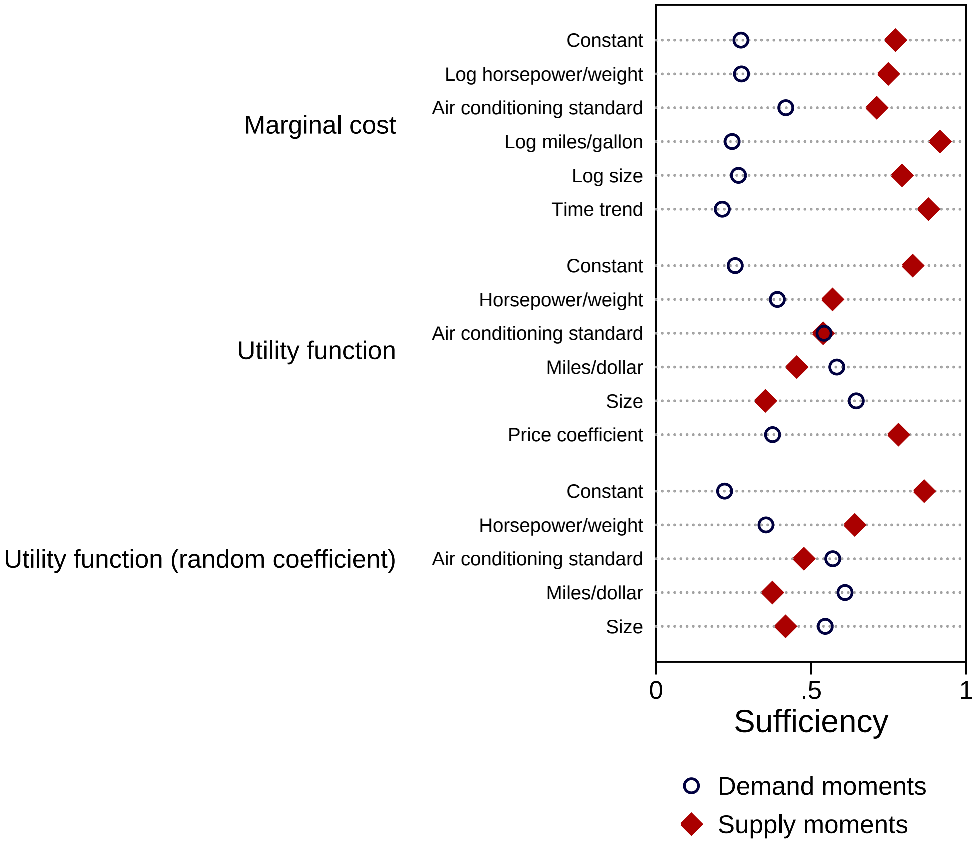
Notes: Each plot shows the standardized sensitivity of the parameter named in the plot title with respect to the full vector of estimation moments, which are the mean adjusted consumption levels at each age.

Figure 3: Standardized sensitivity of select parameters in De Nardi et al. (2010)



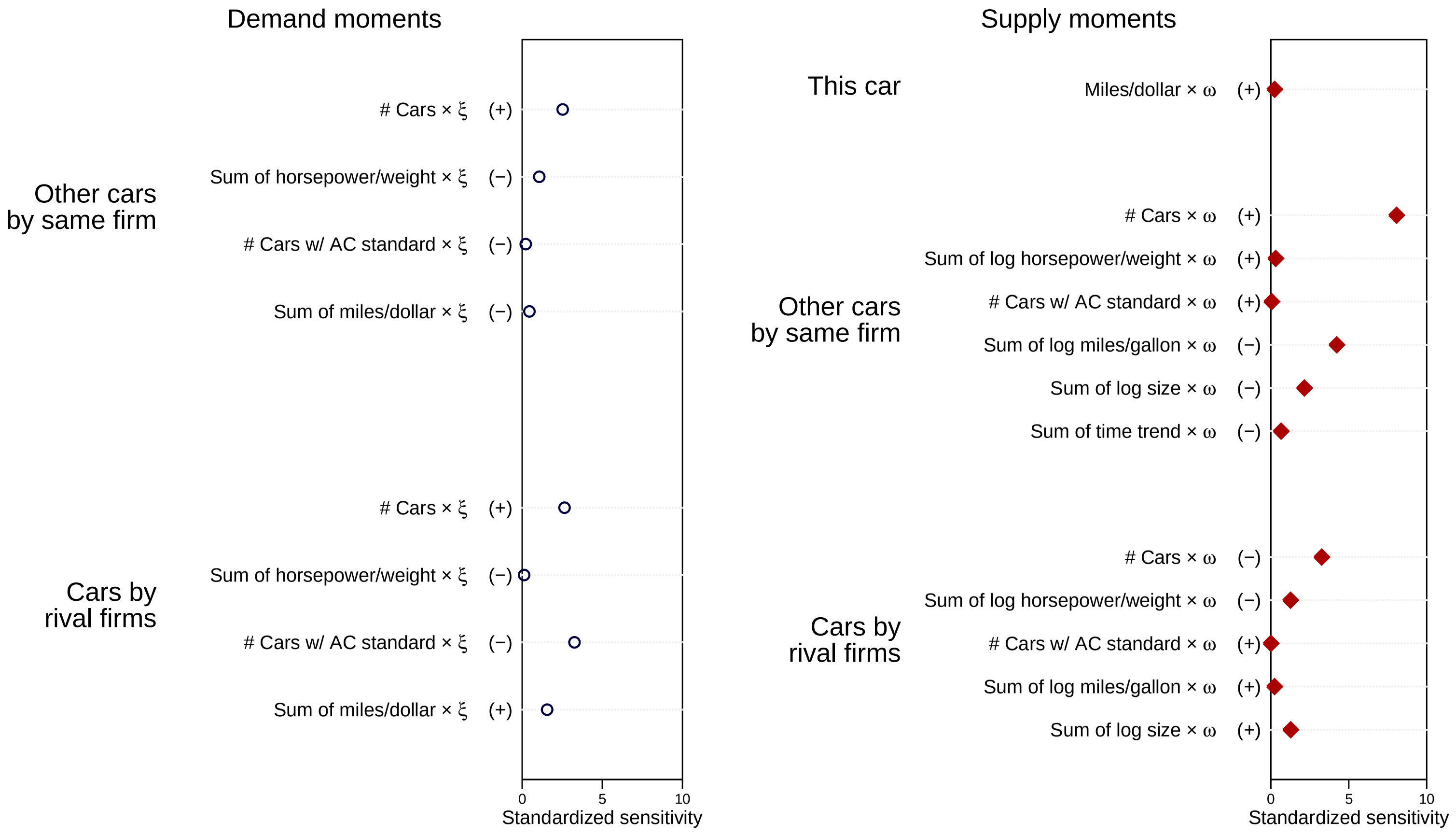
Notes: Each plot shows the standardized sensitivity of the parameter named in the plot title with respect to median asset holdings averaged by permanent income quintile (for the consumption floor and risk aversion) or averaged by age group (for the discount factor).

Figure 4: Sufficiencies for parameter estimates in Berry et al. (1995)



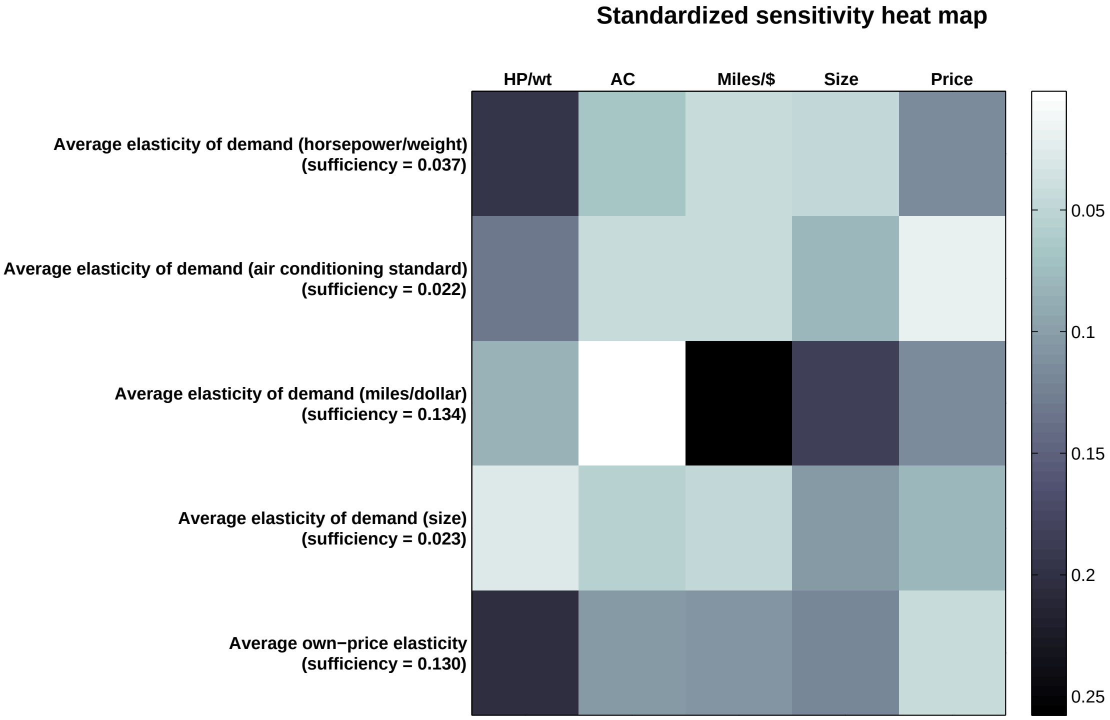
Notes: The plot shows the sufficiency of the demand-side and supply-side estimation moments for each parameter estimate.

Figure 5: Standardized sensitivity of average markup in Berry et al. (1995)



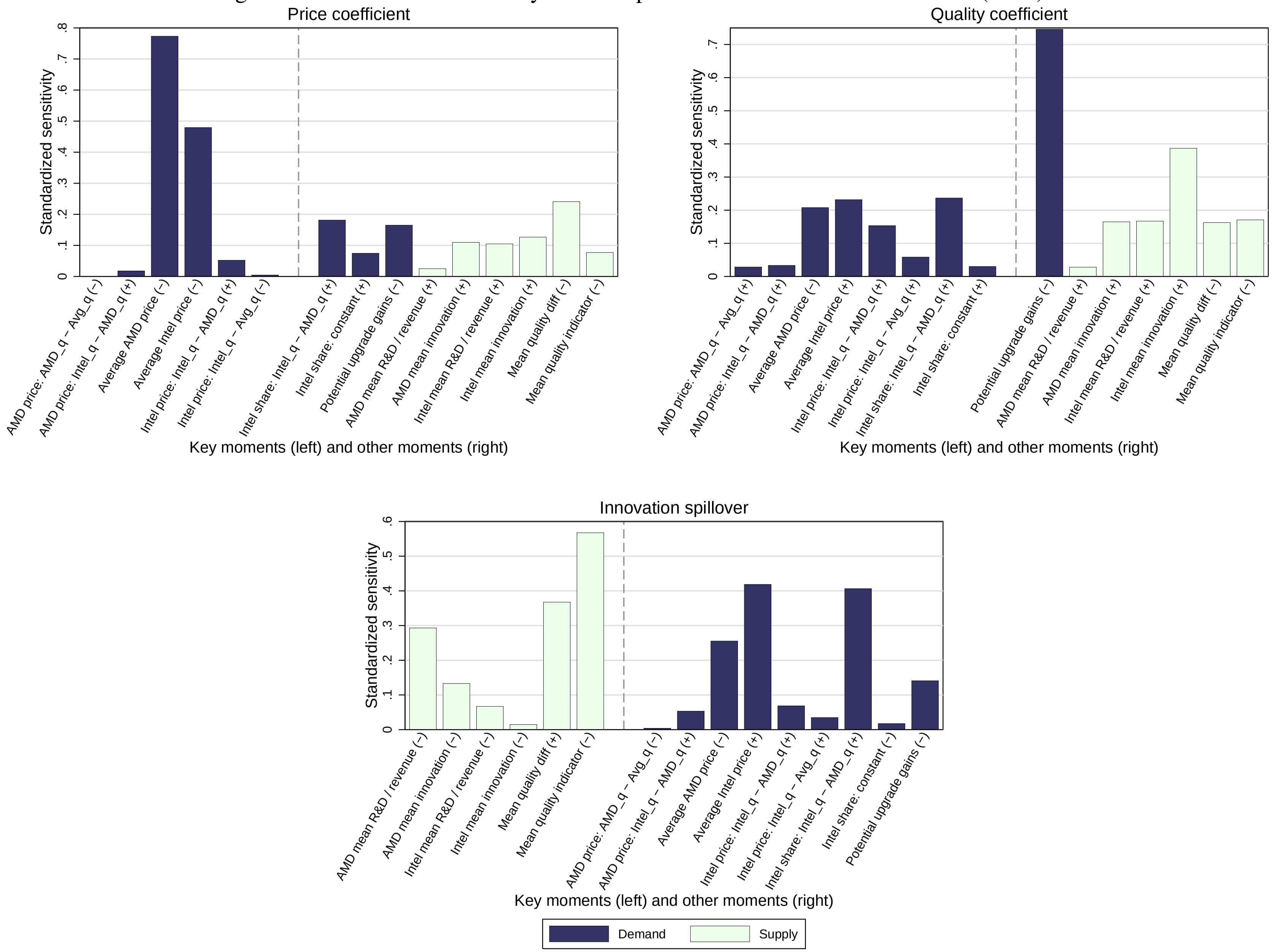
Notes: The plot shows the absolute value of the standardized sensitivity of the implied average markup with respect to the estimation moments, with the sign of sensitivity in parentheses. The demand-side moments are the product of demand-side instruments with expected unobserved quality ξ (computed as a residual inverted from market shares). The supply-side moments are the product of supply-side instruments with unobserved cost component ω (computed as a residual inverted from estimated marginal costs). All estimation moments (except for those corresponding to the "This car: Constant" instruments) use instruments from which the mean has been subtracted. While sensitivity is computed with respect to the complete set of estimation moments, the plot only shows those corresponding to the excluded instruments. To avoid collinearity, we drop three instruments: "Other cars by the same firm: Sum of size," "Cars by rival firms: Sum of size," and "Cars by rival firms: Sum of time trend."

Figure 6: Standardized sensitivity of elasticities of demand in Berry et al. (1995)



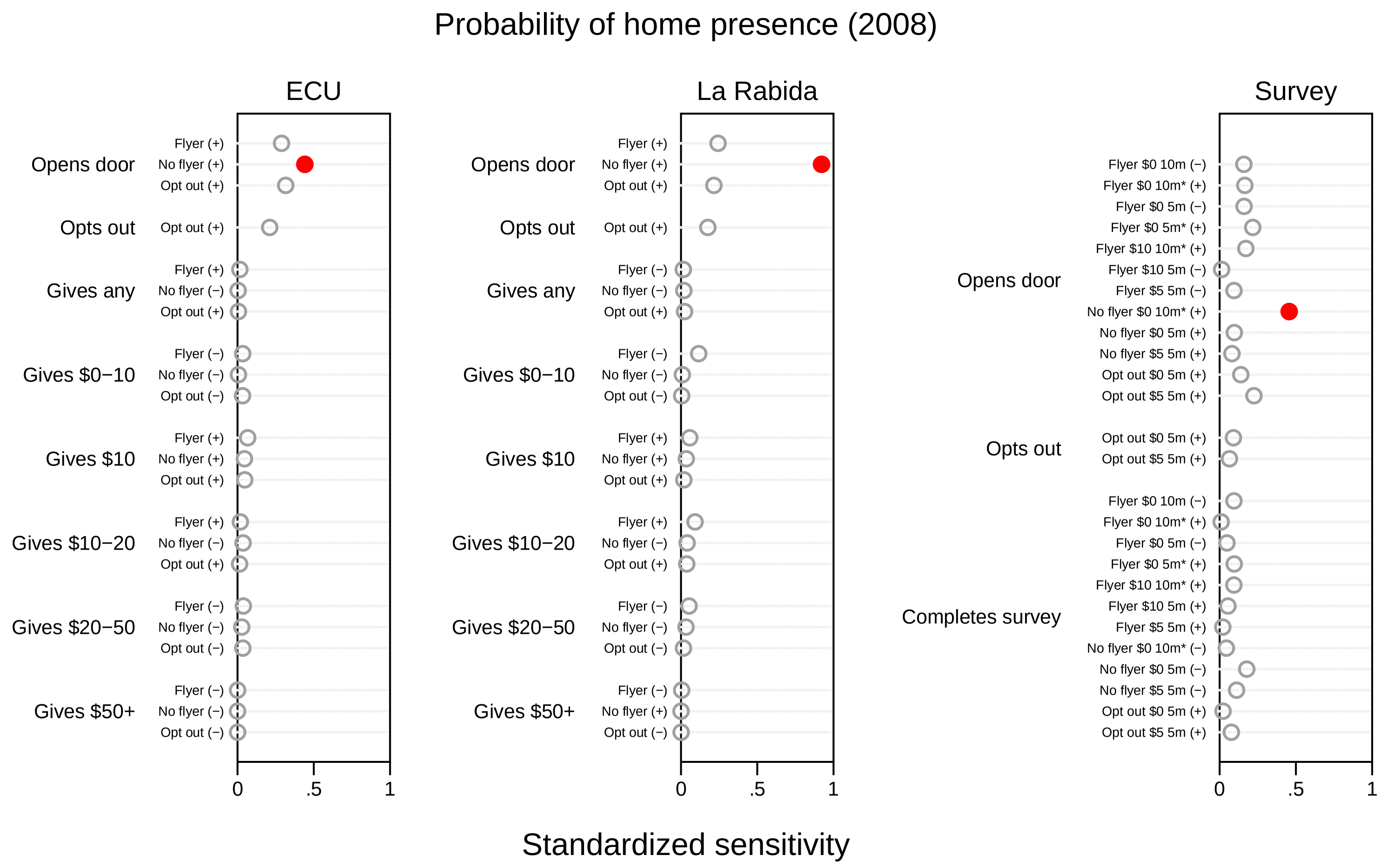
Notes: The plot shows a heat map of the absolute value of standardized sensitivity of the average own-price or own-characteristic elasticity of demand from the BLP model (in rows) with respect to the vector of analogous elasticities from a logit model with the same excluded instruments as the BLP model (in columns). The number in parentheses in each row is the sufficiency of the vector of logit model elasticities for the BLP model elasticity.

Figure 7: Standardized sensitivity of select parameters in Goettler and Gordon (2011)



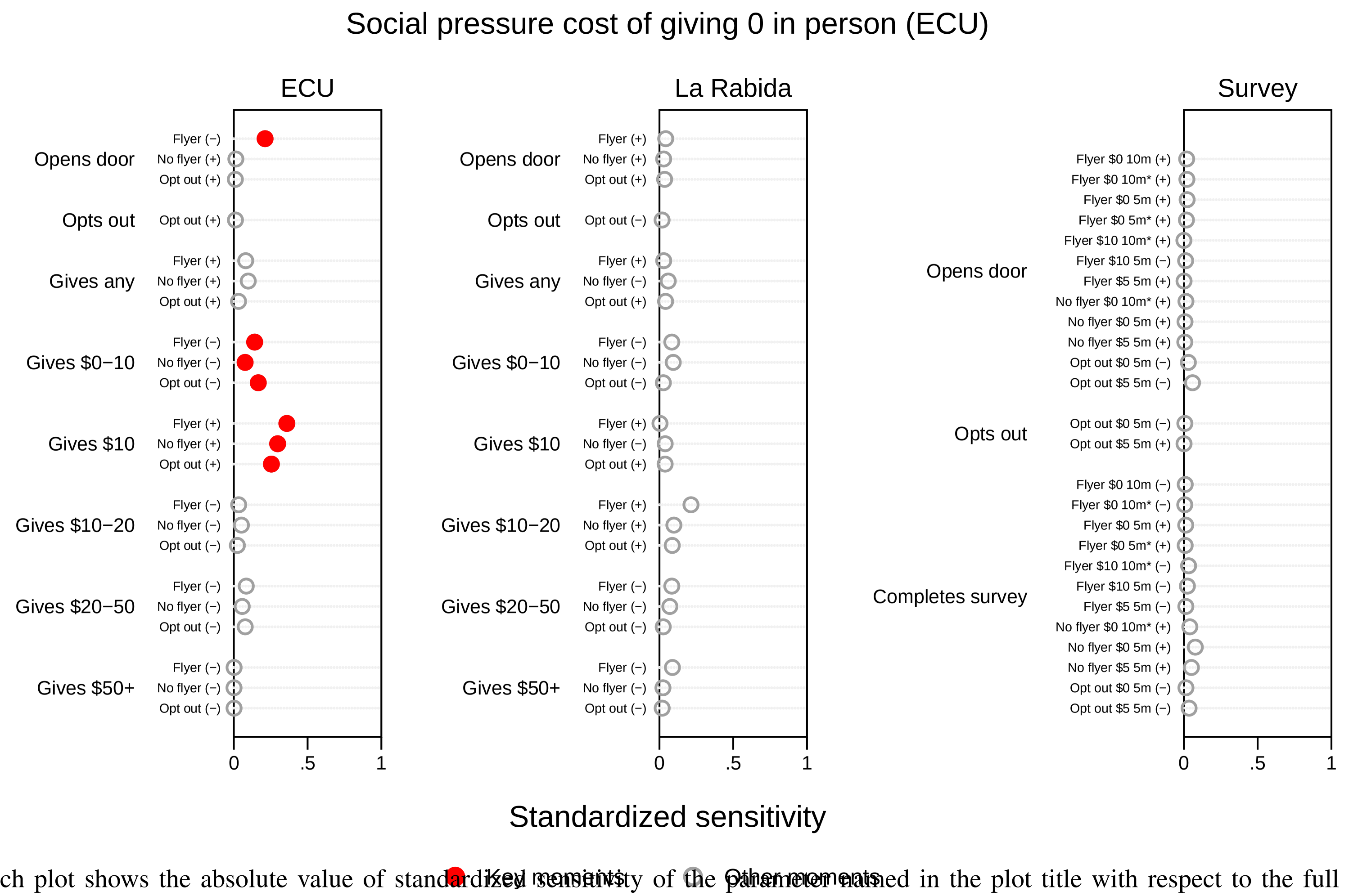
Notes: Each plot shows the absolute value of standardized sensitivity of the parameter named in the plot title with respect to the vector of estimation moments, with the sign of sensitivity in parentheses. A "key moment" is a moment that Goettler and Gordon (2011) highlight as especially sensitive to the given parameter. The plots use the following shorthand: "q" stands for quality of the product offered; "Avg_q" stands for the mean quality of products currently owned; "Intel share: X" stands for the coefficient on X from a regression of Intel share on the difference in quality between Intel and AMD; and "Y price: X" stands for the coefficient on X from a regression of firm Y's price on both the difference in quality between Intel and AMD, and the difference between Y's quality and the mean quality of owned products.

Figure 8: Standardized sensitivity of select parameters in DellaVigna et al. (2012)



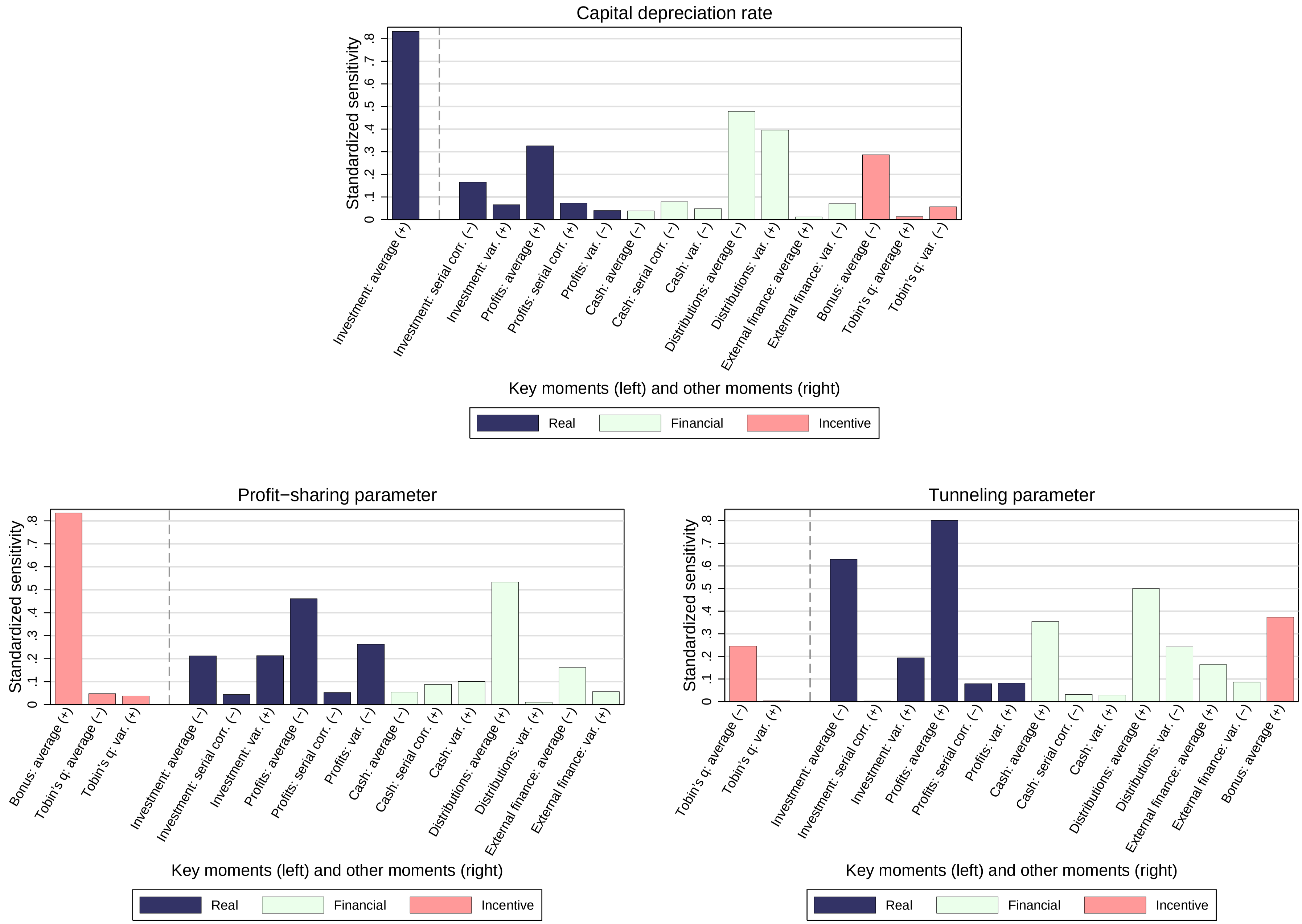
Notes: Each plot shows the absolute value of standardized sensitivity of the parameter named in the plot title with respect to the full vector of estimation moments, with the sign of sensitivity in parentheses. Each moment is the observed probability of a response for the given treatment group. Plot headings define the sub-experiment for each collection of moments; moments denoted with an asterisk are from 2008 data. The leftmost axis labels in larger font describe the response; the axis labels in smaller font describe the treatment group. Colored circles correspond to moments that DellaVigna et al. (2012) highlight as important for identifying the given parameter.

(Figure continues on next page.)



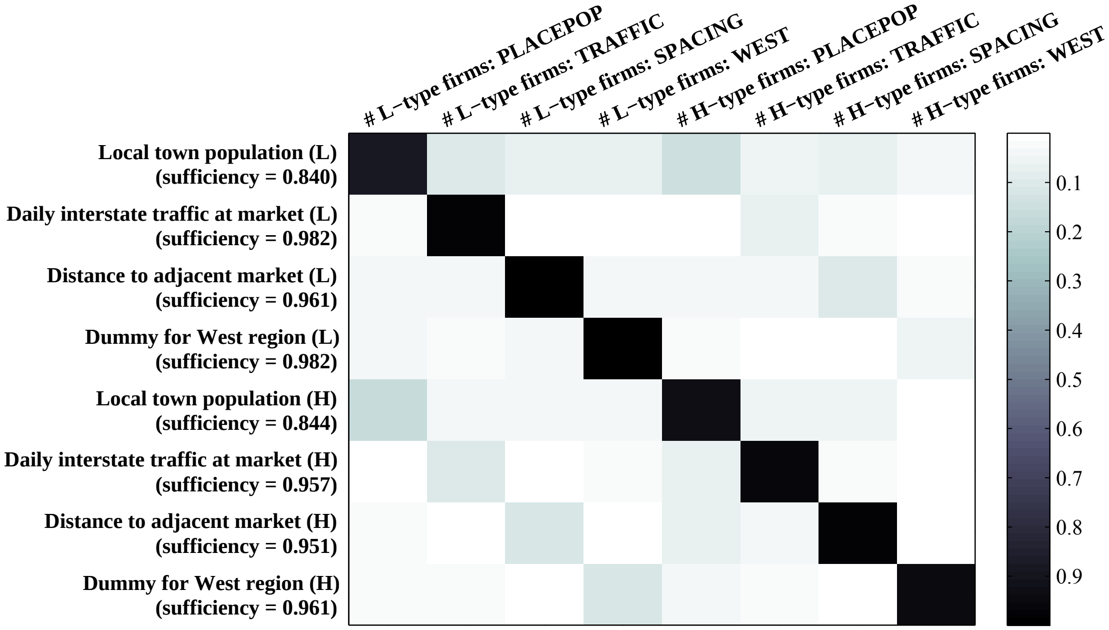
Notes: Each plot shows the absolute value of standardized sensitivity of the parameter named in the plot title with respect to the full vector of estimation moments, with the sign of sensitivity in parentheses. Each moment is the observed probability of a response for the given treatment group. The leftmost axis labels in larger font describe the response; the axis labels in smaller font describe the treatment group. Colored circles correspond to moments that DellaVigna et al. (2012) highlight as important for identifying the given parameter.

Figure 9: Standardized sensitivity of select parameters in Nikolov and Whited (2014)



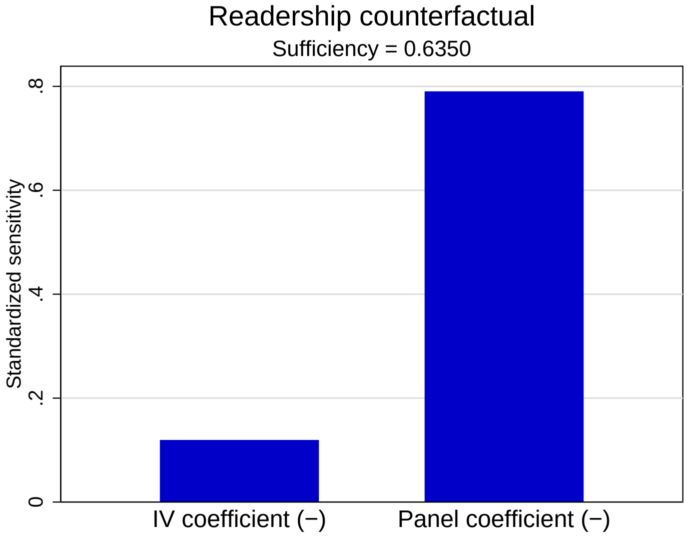
Notes: Each plot shows the absolute value of standardized sensitivity of the parameter named in the plot title with respect to the full vector of estimation moments, with the sign of sensitivity in parentheses. A "key moment" is a moment that Nikolov and Whited (2014) highlight as important for identifying the given parameter.

Figure 10: Standardized sensitivity of market characteristic parameters in Mazzeo (2002)



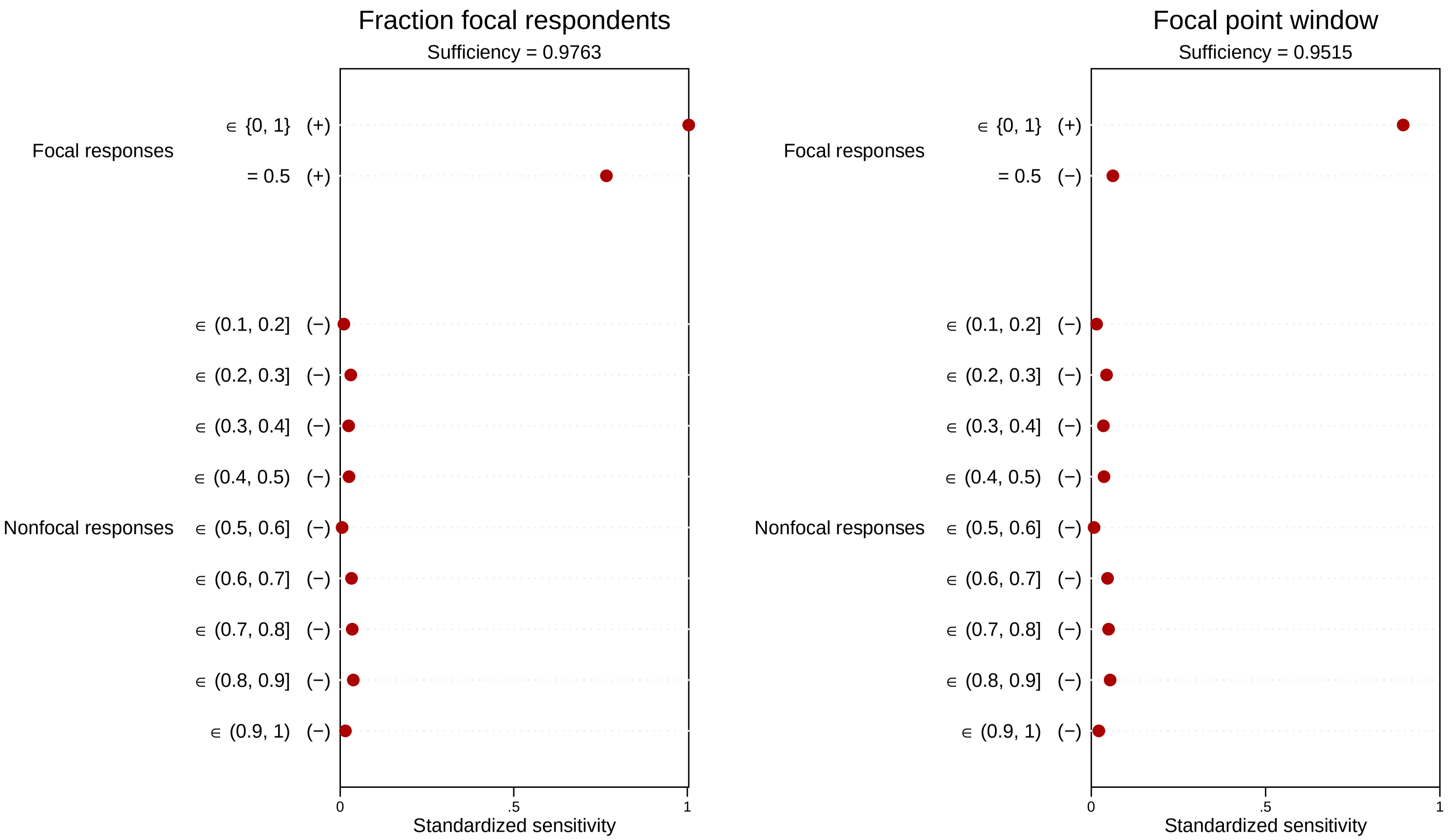
Notes: The plot shows a heat map of the absolute value of the standardized sensitivity of a model parameters (in rows) with respect to a vector of descriptive statistics (in columns). Each row also shows the sufficiency of the vector of statistics for the given parameter. Parameter names ending in “(L)” refer to effects on low-type payoffs, and parameter names ending in “(H)” refer to effects on high-type payoffs. The descriptive statistics are the coefficients from regressions of the number of low- and high-type firms on observable market characteristics. The model is the two-type Stackelberg model defined and estimated in Mazzeo (2002).

Figure 11: Standardized sensitivity of counterfactual estimate in Gentzkow (2007)



Notes: The plot shows the absolute value of standardized sensitivity of the readership counterfactual with respect to the two descriptive statistics listed on the x-axis, with the sign of sensitivity in parentheses and the sufficiency of the vector of descriptive statistics for the given parameter listed above the plot. The readership counterfactual is the change in readership of the print edition of the *Washington Post* when the post.com is removed from the choice set (Gentzkow 2007, table 10). The IV coefficient is the estimated coefficient from a two-stage least squares regression of last-five-weekday *Washington Post* print readership on last-five-weekday post.com readership, with a set of excluded instruments including Internet access at work (reported in Gentzkow 2007, table 4, IV specification (1)). The panel coefficient is the coefficient from an OLS regression of last-one-day print readership on last-one-day online readership controlling flexibly for readership of both editions in the last five weekdays. Each of these auxiliary regressions includes the standard set of demographic controls from Gentzkow (2007).

Figure 12: Standardized sensitivity of select parameters in Hendren (2013)



Notes: Each plot shows the absolute value of the standardized sensitivity of the parameter named in the plot title with respect to a vector of descriptive statistics. The descriptive statistics are the shares of responses that fall into each of the sets listed on the vertical axis. The sign of sensitivity is in parentheses and the sufficiency of the vector of descriptive statistics for the given parameter is listed above the plot. All parameter estimates except for the fraction focal respondents and the focal point window are held constant at their estimated values.

A Relationship to Alternatives

Here we relate our sensitivity measure to two alternative methods of developing intuition for the mapping from data to parameter estimates.

A.1 Inverse Sensitivity

Our sensitivity measure asks how the expected values of the parameters change as we vary the data features of interest. An alternative way to investigate what drives an estimator would be to ask how the expected values of the data features change when we vary the parameters. Intuitively, we might say that a particular $\hat{\theta}_p$ will depend heavily on a particular $\hat{\gamma}_j$ if varying θ_p in the model causes large changes in the expected value of $\hat{\gamma}_j$. This approach can easily be implemented by simulating data from the model at alternative parameter values. Goettler and Gordon (2011), Kaplan (2012), Morten (2013), and Berger and Vavra (2015) are examples of papers that refer to such simulations in their discussions of identification.[35]

This approach can be thought of as the "inverse" of our proposed sensitivity measure. To see why, suppose that $\hat{\theta}$ is a GMM estimator and Λ is sensitivity to moments. The alternative approach would infer that the j-th moment $\hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n} g_j(z_i, \theta_0)$ is an important driver of $\hat{\theta}_p$ if the absolute value of $\frac{\partial}{\partial \theta_p} \mathrm{E}\left[g_j(z_i, \theta)\right]\Big|_{\theta=\theta_0}$ is large. Notice that the matrix of these partial derivatives is simply the Jacobian G. Since $\Lambda = -\left(G^{'} W_g G\right)^{-1} G^{'} W_g$, we have $-\Lambda G = I$, and so when Λ is square $G = -\Lambda^{-1}$.

The intuitions delivered by G agree with those delivered by Λ when the model has a single parameter ($P = 1$) and $W_g = I$. In this case, $(G' W_g G)^{-1}$ is a constant, so $|\Lambda| \propto |G|$. If $\hat{\gamma}_j$ changes more than $\hat{\gamma}_k$ when we vary the single parameter θ, $\hat{\theta}$ will be more sensitive to $\hat{\gamma}_j$ than to $\hat{\gamma}_k$.

Outside of this special case, the intuitions from Λ and G can be very different. While examining G can be a useful way to build economic intuition about a model, we argue that it can be very misleading if interpreted as a guide to the sensitivity of an estimator to misspecification or to the similarity of an estimator to the one defined in an identification proof.

The reason that G is not a good guide to the sensitivity properties of an estimator is that it is not a property of an estimator; rather, it is a (local) property of a model. An easy way to see this is to note that G does not depend on the weight matrix W_g. For an overidentified model, this means

[35] Goettler and Gordon (2011) describe specific parameters as "primarily identified by" particular moments if those moments respond sharply to changes in those parameters (p. 1161). Kaplan (2012) writes: "I address the question of identification in three ways ... Third, below I provide an informal argument that each of the parameters has influence on a subset of the chosen moments and give some intuition for why this is the case" (p. 478). Morten (2013) writes: "As a check on how well the identification arguments for the simple model apply ... I simulate the dynamic model for a range of parameter values. I vary each parameter ... and then plot the responses of each of the... main moments as the parameter changes" (p. 33).

that G can't tell us which features of the data drive a particular $\hat{\theta}$. Consider our earlier example in which θ_0 is the population standard deviation of an exponential random variable. In this case, G tells us that θ is equally related to the mean and the standard deviation, because under the model both change by the same amount when we vary θ. By contrast, Λ reveals that the MLE puts weight only on the sample mean.

The reason that G is not a good guide to identification is that the relationship discussed in section 3.2 does not hold for G: it may be that $\hat{\theta}$ is in fact the hypothetical estimator $\Phi(\cdot)$, but that G assigns zero sensitivity to features that are needed for identification, and non-zero sensitivity to features that are not needed for identification. Recall our OLS example in which $\hat{\gamma} = \left[\begin{array}{cccc} \hat{\mu}_y & \hat{\sigma}_{xy} & \hat{\sigma}_x^2 & \hat{\mu}_x \end{array} \right]'$. The coefficient β is identified by σ_{xy} and σ_x^2 alone. Consistent with this, the row of Λ corresponding to $\hat{\beta}$ has non-zero entries for $\hat{\sigma}_{xy}$ and $\hat{\sigma}_x^2$ and zeros elsewhere. The corresponding column of G, however, has non-zero entries only for μ_y and σ_{xy} (assuming $\mu_x \neq 0$).[36] Changing β affects the mean of Y and its covariance with X, but leaves the mean and variance of X unchanged; however, β is not identified by the mean of Y and its covariance with X alone, and the mean of Y is not necessary for identification of β.

A.2 Dropping Moments

In the case of an overidentified MDE, an alternative way to check sensitivity to an empirical moment is to drop the moment and re-estimate the model. To fix ideas, assume that equation (3) has a solution when the j^{th} element of $\hat{g}(\theta)$ is excluded, and denote the resulting estimator by $\hat{\theta}^{\sim j}$. Comparing the parameters estimated with and without moment j amounts to calculating $\left(\hat{\theta} - \hat{\theta}^{\sim j}\right)$.

Suppose that the j^{th} moment (and only the j^{th} moment) is possibly misspecified. Then the following corollary of proposition 3 shows that the measure $\left(\hat{\theta} - \hat{\theta}^{\sim j}\right)$ combines information about sensitivity Λ with information about the degree of misspecification $\mu_{\eta j}$:

Corollary 1. *Suppose that the assumptions of proposition 3 are satisfied. Suppose that only the j-th element of $\hat{\gamma}$ is potentially misspecified, so $\hat{\theta}^{\sim j}$ is consistent and asymptotically normal satisfying equation (1), and the prior P places probability one on $\eta_k = 0 \forall k \neq j$. Then the asymptotic mean of $\hat{\theta} - \hat{\theta}^{\sim j}$ is $\Lambda_{\cdot j}\mu_{\eta j}$.*

[36] To restate this example as an MDE, let $\theta = \left[\begin{array}{cccc} \alpha & \beta & \sigma_x^2 & \mu_x \end{array} \right]'$ and $\hat{g}(\theta) = \hat{\gamma} - h(\theta)$ where $h(\theta) = \left[\begin{array}{cccc} \alpha + \beta\mu_x & \beta\sigma_x^2 & \sigma_x^2 & \mu_x \end{array} \right]'$. Then

$$G = \left[\begin{array}{cccc} -1 & -\mu_x & 0 & -\beta \\ 0 & -\sigma_x^2 & -\beta & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{array} \right].$$