cemmap

centre for microdata methods and practice

Estimation of a nonseparable heterogenous demand function with shape restrictions and Berkson errors

Richard Blundell
Joel Horowitz
Matthias Parey

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP67/18

Estimation of a Nonseparable Heterogeneous Demand Function with Shape Restrictions and Berkson Errors

Richard Blundell*, Joel Horowitz† and Matthias Parey‡

November 2018§

Abstract

Berkson errors are commonplace in empirical microeconomics and occur whenever we observe an average in a specified group rather than the true individual value. In consumer demand this form of measurement error is present because the price an individual pays is often measured by the average price paid by individuals in a specified group (e.g., a county). We show the importance of such measurement errors for the estimation of demand in a setting with nonseparable unobserved heterogeneity. We develop a consistent estimator using external information on the true distribution of prices. Examining the demand for gasoline in the U.S., accounting for Berkson errors is found to be quantitatively important for estimating price effects and for welfare calculations. Imposing the Slutsky shape constraint greatly reduces the sensitivity to Berkson errors.

JEL: C14, C21, D12

Keywords: consumer demand, nonseparable models, quantile regression, measurement error, gasoline demand, Berkson errors.

*Department of Economics, University College London (UCL), and Institute for Fiscal Studies (IFS). 7 Ridgmount Street, London WC1E 7AE, United Kingdom, email: r.blundell@ucl.ac.uk.

†Department of Economics, Northwestern University, and Cemmap. 2211 Campus Drive, Evanston, Illinois 60208, USA, email: joel-horowitz@northwestern.edu.

‡Department of Economics, University of Essex, and Institute for Fiscal Studies (IFS). Wivenhoe Park, Colchester CO4 3SQ, United Kingdom, email: m.parey@essex.ac.uk.

§We thank seminar participants at the 'Heterogeneity in Supply and Demand' conference at Boston College and the Bristol Econometric Study Group Annual Conference for helpful comments. We are grateful to Agnes Norris Keiller for invaluable help in assembling the GasBuddy data. We would also like to thank the ESRC Centre for the Microeconomic Analysis of Public Policy at IFS and the ERC advanced grant MICROCONLAB at UCL for financial support.

1 Introduction

We consider estimation of a demand model with nonseparable unobserved heterogeneity where the impact of price and income on household demand is the focus of interest. The analysis starts from the observation that datasets that are commonly used in household demand analysis often suffer from a particular type of measurement error in the covariates: Instead of observing the true price a household faces, the researcher observes a regional average price. Thus, only the average price in a specified group (e.g., a county) is observed. The resulting errors in the price variables are called *Berkson errors*.

Berkson measurement errors occur frequently in applied econometric analyses in which information on relevant covariates is not collected directly from households in a survey but is taken from an alternative data source and assigned to households based on their location. While covariates assigned in this way will often be highly correlated with the true covariates, they will not be identical as long as there is some variability in the covariate within the specified locality. Textbook analysis of this kind of model often focuses on the case when the model is linear in the covariate and the error is additive. In this case, Berkson errors do not lead to a bias. This is sometimes taken to mean that Berkson errors are unlikely to cause significant bias in applied analysis, compared to say classical measurement error.

In this paper, we argue that understanding the role of Berkson measurement errors in demand estimation is of growing relevance. The focus on understanding heterogeneity in responses motivates researchers to investigate behavior at different points in the distribution of unobserved heterogeneity, see e.g. Browning and Carro (2007). Moreover, researchers are increasingly interested in nonlinear models with non-separable unoberserved heterogeneity, see e.g. references in Cameron and Trivedi (2005); Blundell et al. (2012, 2017). Better data and increased computational power facilitate the study of models that do not impose linearity restrictions and, instead, allow flexible functional forms with a high degree of potential nonlinearity. Accordingly, nonlinear models are increasingly important in applications. In nonlinear models, Berkson errors are not innocuous and require careful treatment.

This paper develops a method for estimating a nonseparable demand model in the presence of Berkson errors, using a Maximum Likelihood Estimator (MLE). The standard quantile demand approach is inconsistent when prices are subject to Berkson errors. The maximum likelihood procedure we propose estimates all quantiles simultaneously, and a monotonicity constraint is used to ensure that the estimated quantiles do not cross. This estimator enables us to contrast the resulting estimates to results obtained assuming the absence of Berkson errors.

Delaigle et al. (2006) show the demand function is unidentified nonparametrically unless either the distribution of the Berkson error is known or can be estimated consistently from auxiliary data. Alternatively identification can be delivered if there is an instrument that is related to the true price in a suitable way (Schennach (2013)). We choose to follow the first of these approaches and use auxiliary data from external sources to inform us about the distribution of the Berkson error. We then assess the sensitivity to Berkson errors across different levels of the Berkson error variance. Finally, we note there is a potential for gasoline prices to be endogenous. To address this we develop a test for the exogeneity of covariates in the presence of Berkson errors.

We motivate and illustrate our analysis with an application to gasoline demand. Household travel surveys frequently assign gasoline prices from external sources based on the location of the household, leading to the presence of Berkson errors. A long-standing body of work has documented the importance of allowing for potential non-linearities in household gasoline demand (Hausman and Newey (1995); Yatchew and No (2001); Blundell et al. (2012)). The role of unobserved heterogeneity motivates a quantile modelling approach (Blundell et al. (2017); Hoderlein and Vanhems (2018)). These considerations suggest that nonlinearity plays an important role in this appliciation, highlighting the importance of Berkson errors in applied research and the need to treat them carefully.

We find that accounting for Berkson errors is quantitatively important. For example, Deadweight Loss measures derived from our estimates differ substantially when we allow for Berkson errors. In previous work we have investigated the role of shape restrictions in semiparametric or nonparametric estimation settings (Blundell et al. (2012, 2017)). In a setting with Berkson errors, we find that imposing shape restrictions, in the form of the Slutsky inequality, reduces the sensitivity of the estimates to the presence of Berkson errors.

The paper proceeds as follows. In the next section, we outline the demand model and introduce Berkson errors. Section 3 develops the MLE estimator. Section 4 presents the exogeneity test. We describe household gasoline data and the price data in Section 5. We also document how we use the gasoline price information from the GasBuddy website (`www.gasbuddy.com`) to provide external information on the distribution current local gasoline prices. The estimation results for the gasoline demand responses to prices and for deadweight loss welfare measures are presented in Section 6. Section 7 concludes.

2 Model

In this section we first outline the nonseparable demand model in the absence of Berkson errors. We then introduce Berkson errors into the model.

2.1 The Demand Model without Berkson Errors

To set the notation consider the demand function with nonseparable unobserved heterogeneity

$$Q = G(P, Y, U)$$

where Q is the quantity demanded, P the price, Y household income, and U unobserved heterogeneity. We assume that U is a scalar random variable that is statistically independent of (P, Y), and that $G(P, Y, U)$ is monotone increasing in its third argument.[1] We further assume without further loss of generality that $U \sim \mathrm{U}[0, 1]$.

Under these assumptions, the α quantile of Q conditional on (P, Y) is

$$Q_\alpha = G(P, Y, \alpha) \equiv G_\alpha(P, Y).$$

[1] The assumption of scalar unobserved heterogeneity (U) is restrictive but necessary to achieve point identification and to do welfare analysis. Hausman and Newey (2017) and Dette et al. (2016) discuss models with multi-dimensional unobserved heterogeneity. We address the possibility that P is endogenous in Section 4.

That is, the the conditional α quantile of Q recovers the demand function G, evaluated at $U = \alpha$.

2.2 The Demand Model with Berkson Errors

Suppose now that we do not observe the true price at which a transaction took place, which we refer to as $P^\star$. Instead, we observe a county average price P that is related to $P^\star$ by

$$P^\star = P + \epsilon,$$

where ϵ is an unobserved random variable, independent of P. The resulting errors in variables are called *Berkson errors* (Berkson (1950)).

With Berkson errors, the demand model becomes

$$Q = G(P + \epsilon, Y, U).$$

Berkson errors are common in economics data. For example, relevant covariates may not be surveyed or measured at the level of the household, but are instead approximated by a regional average from an external source. Importantly, Berkson errors in variables are different from classical errors in variables, where $P = P^\star + \epsilon$, with ϵ independent of $P^\star$.

The function G is unidentified nonparametrically unless either the distribution of ϵ is known or can be estimated consistently from auxiliary data (Delaigle et al. (2006)) or, alternatively, there is an instrument Z that is related to the true price $P^\star$ in a suitable way (Schennach (2013)). In this work we follow the first of these approaches, and use auxiliary data to inform us about the distribution of the Berkson error.

3 Estimation

3.1 A Maximum Likelihood Estimator

In this section we develop the Maximum Likelihood Estimation approach. The model is

$$Q = G(P + \epsilon, Y, U); \quad U \sim \mathrm{U}[0, 1].$$

Therefore,

$$\begin{aligned} P(Q \leq z|P, Y) &= P(G(P + \epsilon, Y, U) \leq z|P, Y) = P(U \leq G^{-1}(P + \epsilon, Y, z)|P, Y) \\ &= \int G^{-1}(P + \epsilon, Y, z) f_\epsilon(\epsilon) d\epsilon \\ &= E_\epsilon G^{-1}(P + \epsilon, Y, z), \end{aligned} \tag{1}$$

where $G^{-1}(\cdot, \cdot, z)$ is the inverse of G in the third argument.

The left-hand term of equation (1), $P(Q \leq z|P, Y)$, is identified by the sampling process. G^{-1} and G are identified nonparametrically if and only if G^{-1} is determined uniquely by

$$P(Q \leq z|P, Y) = E_\epsilon G^{-1}(P + \epsilon, Y, z).$$

This requires knowledge of $f_\epsilon(\epsilon)$; Delaigle et al. (2006) present a similar identification result for a conditional mean model.[2]

The truncated series

$$G^{-1}(P + \epsilon, Y, Q) \approx \sum_{j=1}^{J} \theta_j \Psi_j(P + \epsilon, Y, Q) \tag{2}$$

provides a flexible parametric approximation to G^{-1}. In the truncated series, J is the (fixed) truncation point, the Ψ_j 's are basis functions and the θ_j 's are Fourier coefficients. The data $\{Q_i, P_i, Y_i : i = 1, ..., n\}$ are a random sample of n households. The log-likelihood function for estimating parameter vector θ is the logarithm of the proba-

[2]Note that the identification condition can be formulated as a version of the completeness condition of Nonparametric Instrumental Variables (NPIV) models. See Newey and Powell (2003).

bility density of the data. This is:

$$\log L(\theta) = \sum_{i=1}^{n} log \sum_{j=1}^{J_n} \theta_j \int \left. \frac{\partial \Psi_j(P_i + \epsilon, Y_i, z)}{\partial z} \right|_{z=Q_i} f_\epsilon(\epsilon) d\epsilon$$

Maximum likelihood estimation of θ consists of maximizing $\log L(\theta)$ subject to the following constraints: first, that G^{-1} is non-decreasing in its third argument, and second, $0 \leq G^{-1} \leq 1$. The maximum likelihood procedure estimates all quantiles simultaneously, and by imposing the monotonicity constraint above ensures that the estimated quantiles do not cross.

3.2 Shape Restrictions

In some of the estimates we also impose the Slutsky shape restriction from consumer theory. Assuming quantity, income and prices for household i are measured in logs, and S_i reflects the budget share of household i, the Slutsky constraint, evaluated at (P_i, Y_i, U_i) can be written as

$$\frac{\partial Q}{\partial P}(P_i, Y_i, U_i) + \frac{\partial Q}{\partial Y}(P_i, Y_i, U_i)\ S_i \leq 0.$$

From $U = G^{-1}(P, Y, Q)$, we re-write the price and income effect in terms of G^{-1}, so that the Slutsky condition for household i is

$$\frac{\partial G^{-1}}{\partial P}(P_i, Y_i, Q_i) + \frac{\partial G^{-1}}{\partial Y}(P_i, Y_i, Q_i)\ S_i \geq 0. \tag{3}$$

The estimation then proceeds by maximizing the log-likehood as before, adding the constraint (3) for a set of households in the data. For the presentation of the results, we numerically invert the estimated function $\hat{G}^{-1}$ to obtain the corresponding demand function $\hat{G}$.

4 An Exogeneity Test

A common concern in demand estimation is the possible endogeneity of the price variable, where local prices are correlated with consumer preferences (see Blundell et al. (2012, 2017)). If a variable W is available as an instrument for the price, the researcher can test for the presence of endogeneity. In a nonparametric or flexible parametric model, such a test is likely to have better power properties than a comparison of the exogenous estimate with an instrumental variables (IV) estimate. We therefore develop an exogeneity test, which takes account of the presence of Berkson errors. In this section we state the test statistic and asymptotic approximation to its distribution. The corresponding derivations can be found in Appendix A.2.

Assume that the instrument, W, satisfies

$$P(U \leq \tau | W, Y) = \tau.$$

Let G_{EX}^{-1} denote the inverse demand function G^{-1}, described in Section 3, under the null hypothesis H_0 that P is exogenous. Under H_0

$$\begin{aligned} &\Pr\left[G_{EX}^{-1}(P+\varepsilon, Y, Q | W = w, Y = y) \leq \tau\right] \\ &= E \int I\left[G_{EX}^{-1}(P+\varepsilon, Y, Q | W = w, Y = y) \leq \tau\right] f_{\varepsilon}(\varepsilon) d\varepsilon = \tau \end{aligned} \tag{4}$$

for any (y, w) in the support of (Y, W). The exogeneity test statistic is based on a sample analog of this relation. Let f_{YW} denote the probability density function of (Y, W). Let K be a probability density function that is supported on $[-1, 1]$ and symmetrical around 0. Let $\{h_n : n = 1, 2, ...\}$ be a sequence of positive numbers that converges to 0 as $n \to \infty$. K is called a kernel function and $\{h_n\}$ is called a sequence of bandwidths. Denote the data by $\{Q_i, P_i, Y_i, W_i : i = 1, \ldots, n\}$. Let $\hat{f}_{YW}$ be a kernel nonparametric estimator of f_{YW}:

$$\hat{f}_{YW}(y, w) = \frac{1}{nh_n^2} \sum_{i=1}^{n} K\left(\frac{W_i - w}{h_n}\right) K\left(\frac{Y_i - y}{h_n}\right).$$

Let $\hat{G}_{EX}^{-1}$ denote the MLE of G_{EX}^{-1}. Define

$$S_n(y,w) = \frac{1}{nh^2}\sum_{i=1}^{n}\left\{\int I\left[\hat{G}_{EX}^{-1}(P_i+\varepsilon,Y_i,Q_i)\leq\tau\right]f_\varepsilon(\varepsilon)d\varepsilon K\left(\frac{W_i-w}{h_n}\right)K\left(\frac{Y_i-y}{h_n}\right)\right\}.$$

$S_n(y,w)/\hat{f}_{YW}(y,w)$ is a sample analog of the integral expression in (4). The test statistic is

$$T_n = nh_n^2\int\left[S_n(y,w)-\tau\hat{f}_{YW}(y,w)\right]^2 dwdy.$$

To obtain an asymptotic approximation to the distribution of T_n, assume without loss of generality that $(y,w)\in[0,1]^2$. Let $\{\hat{\lambda}_j : j=1,\ldots,n\}$ denote the eigenvalues of the operator

$$C(y_1,w_1;y_2,w_2) = \tau(1-\tau)\hat{f}_{YW}(y_1,w_1)\int K(\xi)K(\xi+\delta_W)K(\zeta)K(\zeta+\delta_Y)d\xi d\zeta.$$

Let $\{L_n : n=1,2,\ldots\}$ be an increasing sequence of positive constants such that $L_n\to\infty$ and $n^{-1/2}L_n^{3/2}\to 0$ as $n\to\infty$. Under regularity conditions that are stated in the appendix,

$$\left|T_n-\sum_{j=1}^{L_j}\hat{\lambda}_j\chi_j^2\right|\to^p 0$$

as $n\to\infty$, where the χ_j^2 s are independent random variables that are distributed as chi-square with one degree of freedom. The distribution of T_n can be approximated by that of

$$\omega = \sum_{j=1}^{L_n}\hat{\lambda}_j\chi_j^2.$$

The quantiles of the distribution of ω can be estimated with any desired accuracy by Monte Carlo simulation.

5 Data on Demand and Prices

5.1 The household gasoline demand

The data are from the 2001 National Household Travel Survey (NHTS), which surveys the civilian noninstitutionalized population in the United States. This is a household level survey conducted by telephone, and complemented by travel diaries and odometer readings.[3] These data provide information on the travel behavior of selected households. We focus on annual mileage by vehicles owned by the household.

In order to minimize heterogeneity in the sample, the following restrictions are imposed: We restrict attention to households with a white respondent, two or more adults, and at least one child under age 16. We drop households in the most rural areas, where farming activities are likely to be particularly important. We also omit households in Hawaii due to its different geographic situation compared to the continental states. Households without any drivers or where key variables are not observed are excluded, and we restrict attention to gasoline-based vehicles (excluding diesel, natural gas, or electricity based vehicles).[4] The sample we use is the same as in Blundell et al. (2017).

A key aspect of the data is that although odometer readings and fuel efficiencies are recorded, price information is not collected at the household level, reflecting the expense in collecting purchase diaries and the resulting burden for respondents (EIA (2003); Leckey and Schipper (2011)). Instead, in the NHTS gasoline prices are assigned the fuel cost in the local area, based on the location of the household (EIA (2003)). In Section 5.2 we document that households face price variability within local markets, and we use this information to assess the extent of Berkson errors.

The resulting sample contains 3,640 observations. Table 1 presents summary statistics. The reported means of our key variables correspond to about 1,250 gallons of gasoline per year, a gasoline price of \$1.33, and household income of about \$63,000. For reference, Table 2 presents baseline estimates of price and income elasticities from a log-log model

[3]See ORNL (2004) and Blundell et al. (2012) for further detail on the survey.

[4]We require gasoline demand of at least one gallon, and we drop one outlier observation where the reported gasoline share is larger than 1.

Table 1: Sample descriptives

	Mean	*St. dev.*
Log gasoline demand	7.127	0.646
Log price	0.286	0.057
Log income	11.054	0.580
Observations	3640	

Note: Table presents mean and standard deviations. See text for details.

of gasoline demand. In the mean regression model, we find a price elasticity of -0.83 and

Table 2: Log-log model estimates

	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$	OLS
	(1)	(2)	(3)	(4)
$\log(p)$	-1.00	-0.72	-0.60	-0.83
	[0.22]	[0.19]	[0.22]	[0.18]
$\log(y)$	0.41	0.33	0.23	0.34
	[0.02]	[0.02]	[0.02]	[0.02]
Constant	2.58	3.74	5.15	3.62
	[0.25]	[0.21]	[0.25]	[0.20]
N	3640	3640	3640	3640

Note: Dependent variable is log gasoline demand. See text for details.

an income elasticity of 0.34, similar to the elasticities reported in other studies of gasoline demand (see further Blundell et al. (2017)). Looking across quantiles, we find the lower quantile households to be more sensitive to changes in prices and income.

In the estimation below, the function G^{-1} is specified as a product of three Chebyshev polynomials, one each for P, Y, and Q. We use cubic polynomials in price and income, and a 7th-degree polynomial in quantity. The high-degree polynomial in quantity enables us to estimate differences in the demand function across quantiles of the distribution of unobserved heterogeneity.[5] When we impose the Slutsky constraint, using the observed data points in the sample, we restrict attention to those data points broadly in the areas

[5]We also trim the top and bottom 1 percent of the quantity distribution.

of the data which we are focusing our analysis on below.[6]

5.2 Dispersion in local gasoline prices

In this subsection, we present evidence on the within-market dispersion of gasoline prices. To gain insight into this, we draw on data from the gasoline price information website www.gasbuddy.com. Gasbuddy operates a website (and mobile app) where users report current local gasoline prices, and this information is then made available to other consumers. Atkinson (2008) compares gasoline price data from the same website for Canada with externally collected data and finds the crowdsourced data to be reliable.[7] To provide a description of the within-market price variability, we select seven counties in the U.S. as examples, and note the reported prices as shown on the website's map for each county on a given day. This results in a sample of 5,953 price observations.[8]

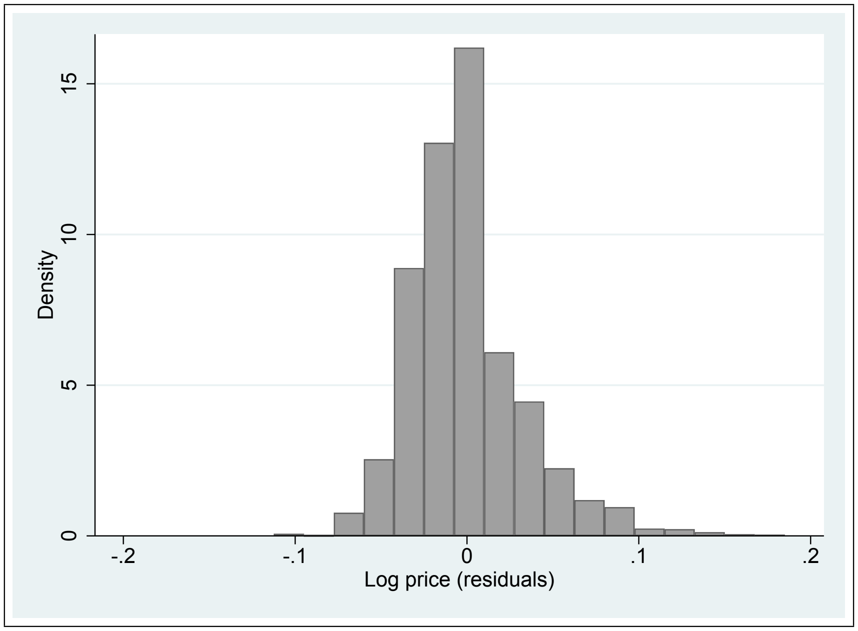
While it is possible that a limited amount of measurement error may result from the manual transcription of the gas prices shown on a map, this is unlikely to bias the resulting estimates systematically. Figure 1 shows a histogram of the gas prices collected, after removing county fixed effects. The price deviations are concentrated between -0.1 and $+0.1$, and the histogram suggests that a normal approximation of the within-market dispersion broadly captures the shape of the distribution. (Figure 2 shows individual histograms for each of these seven counties.) Table 3 shows the standard deviations of (log) prices, across the seven counties studied; these standard deviations vary between 0.024 and 0.043, with a weighted average of 0.033. In our analysis below, we therefore use a normal distribution with a standard deviation of 0.033 for the distribution of the Berkson error. Comparing this value to the reported standard deviation of 0.057 in the NHTS price variable (see Table 1) shows that a significant amount of price variability

[6] For this purpose, we add restrictions for data points between the 10th and the 90th percentile of the unconditional demand data, 0.2 to 0.36 in the log price dimension, and household income between 20,000 and 90,000 USD.

[7] In particular, 78.6% of reported prices were correct at some point during the relevant day, and, for price reports which differed, the mean difference was only 0.8 cents per litre (relative to a price level of around 100 cents per litre); furthermore, deviations were equally likely to be positive as negative, suggesting that there is no systematic tendency of over- or underreporting (Atkinson (2008)).

[8] The counties selected for this exercise were Cook County (IL), Dallas County (TX), Harris County (TX), Los Angeles County (CA), Maricopa County (AZ), Miami-Dade County (FL), and Queens County (NY). Data collection took place July 2014.

Figure 1: Histogram of GasBuddy price distribution (selected counties)



Note: Histogram shows distribution of gasoline prices collected for selected counties, after removing county effects. See text for details.

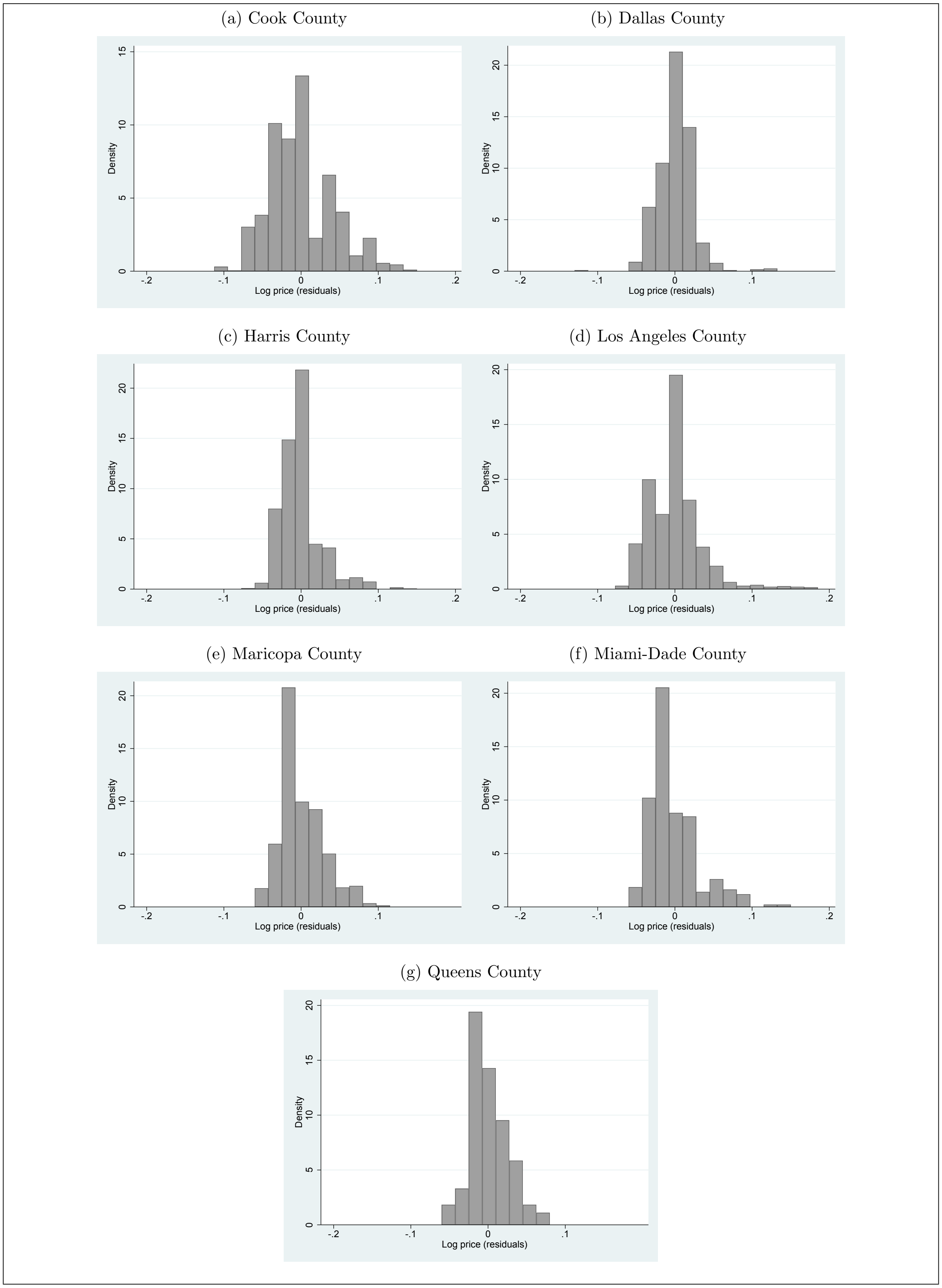
occurs within local markets, suggesting that the Berkson error is an important feature of the price variation in this sample.

5.3 Gasoline price cost shifter

To examine the exogeneity of prices we require a variable which is correlated with gasoline prices, but uncorrelated with the unobservable type of the household. Building on earlier work (Blundell et al. (2012)), we use transportation cost as a cost shifter. This reflects that the cost of transporting the fuel from the supply source is an important determinant of prices.

We measure transportation cost with the distance between one of the major oil platforms in the Gulf of Mexico and the state capital. The U.S. Gulf Coast region accounts for the majority of total U.S. refinery net production of finished motor gasoline and for almost two-thirds of U.S. crude oil imports. It is also the starting point for most major gasoline pipelines. We therefore expect that transportation cost increases with distance to the Gulf of Mexico (see Blundell et al., 2012, for further details and references). Appendix Figure A.1 shows the systematic and positive relationship between state-level

Figure 2: Histogram of GasBuddy price distribution (selected counties)



Source: Histogram shows distribution of gasoline prices collected on GasBuddy.com for selected counties, after removing county effects. Data collection took place July 2014. See text for details.

Table 3: Standard deviation of prices collected on GasBuddy.com for selected counties

County	Standard deviation	Observations
Cook County, IL	0.043	1131
Dallas County, TX	0.024	641
Harris County, TX	0.027	1371
Los Angeles County, CA	0.037	1323
Maricopa County, AZ	0.028	803
Miami-Dade County, FL	0.037	528
Queens County, NY	0.024	156
Weighted average	**0.033**	**5953**

Note: Table shows standard deviation and number of observations of prices reported on GasBuddy.com within the selected counties. Data collection took place July 2014. See text for details.

average prices and the distance to the Gulf of Mexico.

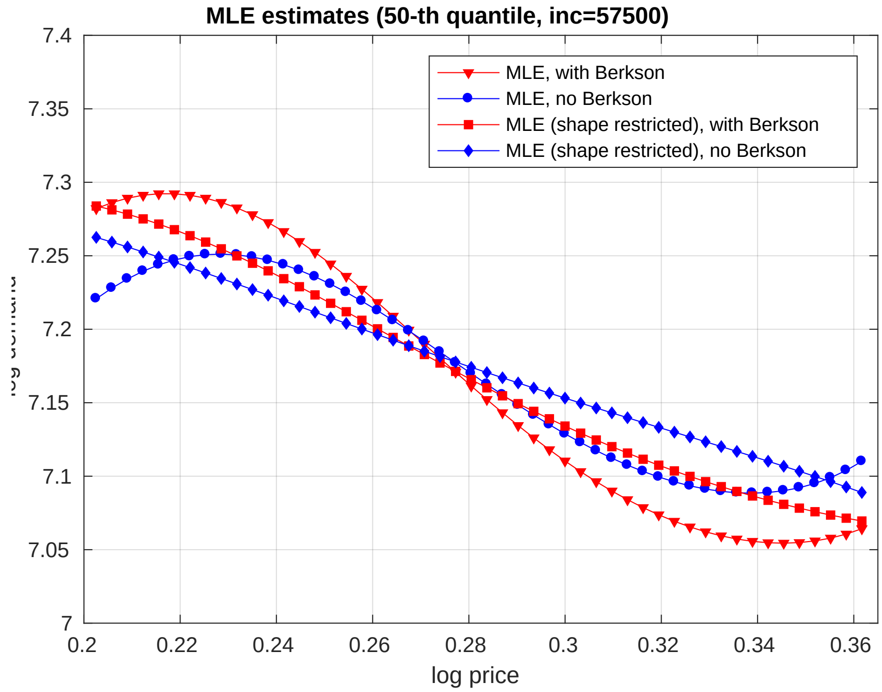
6 Empirical Results

6.1 Demand estimates

Figure 4 shows the ML estimates at the quartiles of the distribution of the unobserved heterogeneity, for the middle income group ($57,500). The round markers show the MLE estimates without taking account of Berkson errors; the upside down triangular markers show the MLE with Berkson error. As can be seen from the Figure, accounting for Berkson errors accentuates the variability in the demand estimates, and leads to relevant differences in the estimated price responsiveness. For the median, for example, shifting the price across the full range shown in the figure (from 0.20 to 0.36) leads to a fall in estimated (log) demand by 0.11 assuming the absence of Berkson errors, compared to 0.22 in the presence of Berkson errors. Note the non-monotonicity in the unconstrained demand curve estimates, which is an artifact of random sampling variation (see further Blundell et al. (2012, 2017)). This non-monotonicity appears to accentuate the sensitivity to the Berkson errors in this empirical example.

The square markers in Figure 3 and Figure 4 show the estimates when we impose Slutsky negativity. Although there is still a difference in the slope, the two sets of estimates

Figure 3: MLE estimates at the median (at middle income)

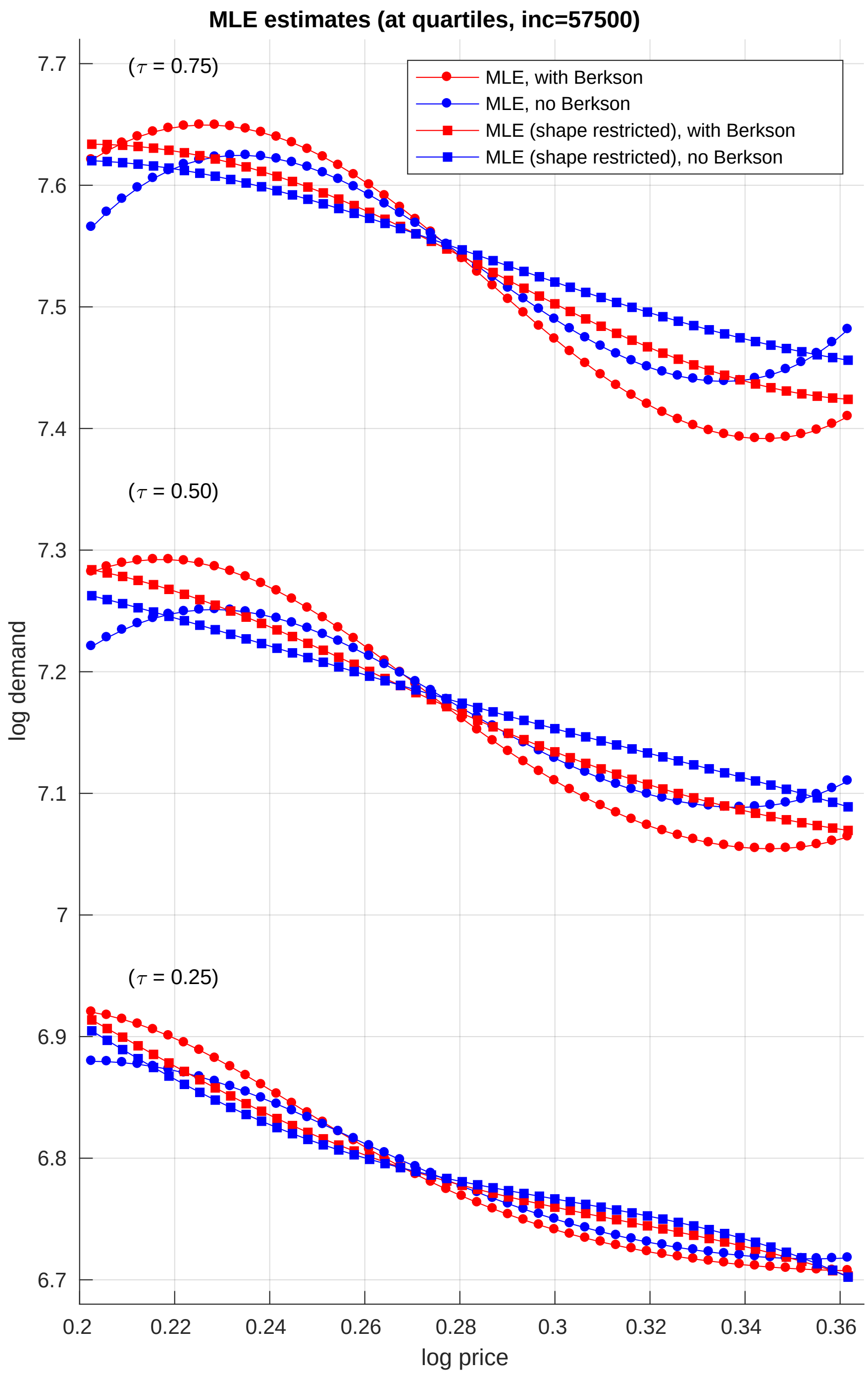


Note: The figure shows MLE estimates at the median ($\tau = 0.50$) for the middle income group. Lines shown in red are estimates accounting for Berkson error, lines shown in blue assume absence of Berkson error. The figure compares unconstrained estimates versus Slutsky-constrained estimates (see legend). See text for details.

are now much more similar. Looking across the different quantiles, we note a consistent finding that imposing the Slutsky inequality restriction removes non-monotonicity and delivers a smoother estimated demand curve much less sensitive to Berkson errors. This suggests that the estimates under the shape restriction are less sensitive to accounting for Berkson errors, reflecting the stabilizing effect of the shape restriction on the demand estimate.

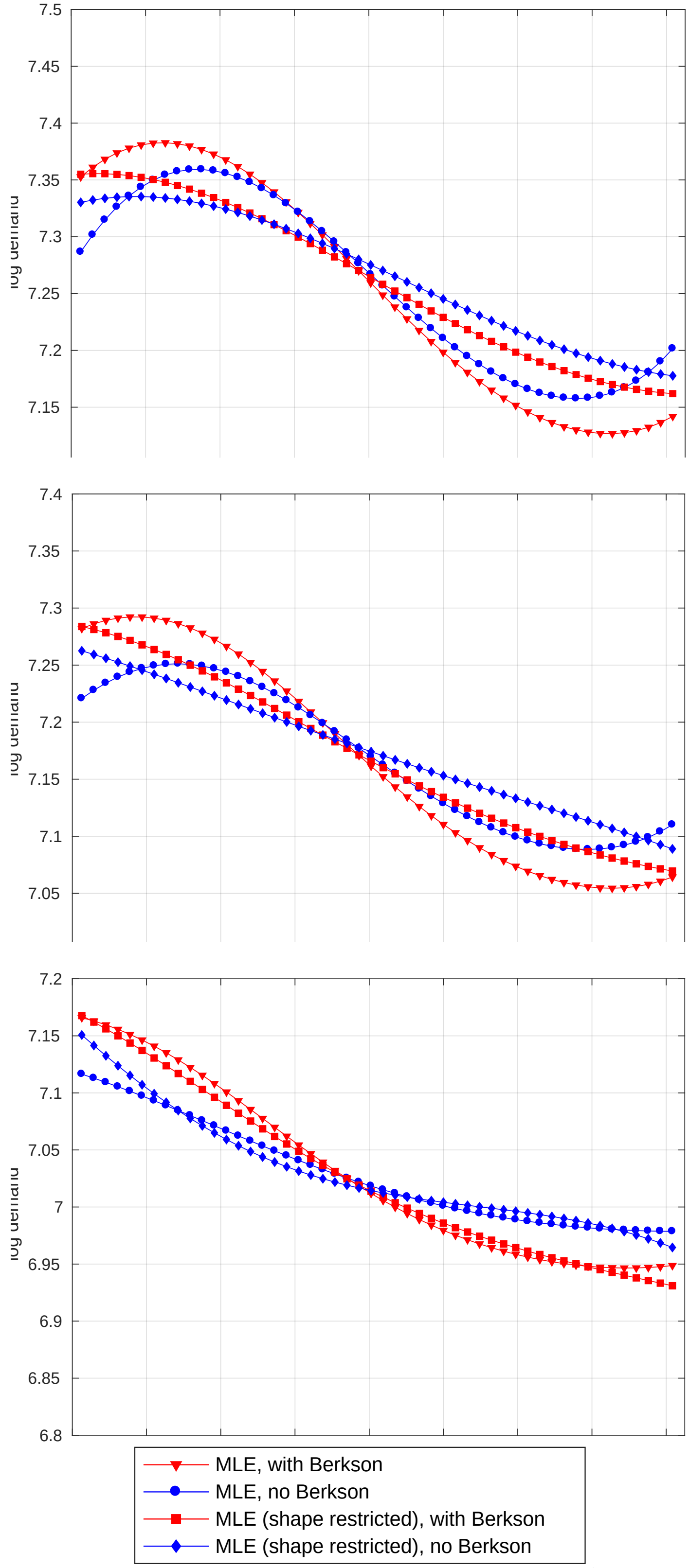
Figure 5 compares the estimated effect at the median across the income distribution, comparing \$ 72,500, \$ 57,500, and \$ 42,500, representing upper, middle and lower income households, respectively. These results highlight the importance of the Slutsky restriction in achieving monotonicity. In this way, these results not only provide demand function estimates that are consistent with consumer theory, but in addition attenuate sensitivity to Berkson errors. However, although the mitigation of sensitivity to Berkson errors through imposing the Slutsky restriction is a clear empirical finding of our analysis, we do not claim that it is a theoretical necessity.

Figure 4: MLE estimates across quartiles (at middle income)



Note: The figure shows MLE estimates at the three quartiles (upper quartile, $\tau = 0.75$, median, $\tau = 0.50$, and lower quartile, $\tau = 0.25$) for the middle income group. Lines shown in red are estimates accounting for Berkson error, lines shown in blue assume absence of Berkson error. The figure compares unconstrained estimates versus Slutsky-constrained estimates (see legend). See text for details.

Figure 5: MLE estimates across the income distribution (at $\tau = 0.50$)



Note: The figure shows MLE estimates for the three income groups (top panel: 'high' income, corresponding to \$72,500, middle panel: 'medium' income, corresponding to \$57,500, and bottom panel: 'low' income, corresponding to \$42,500) at the median ($\tau = 0.50$). Lines shown in red are estimates accounting for Berkson error, lines shown in blue assume absence of Berkson error. The figure compares unconstrained estimates versus Slutsky-constrained estimates (see legend). See text for details.

Figure 6 compares the estimates for different magnitudes of the Berkson error, varying the standard deviation with factor 2 and factor 0.5, respectively. For small standard deviations (panel (b)), the presence of Berkson error makes very little difference to the demand estimates. However for larger standard deviation of the Berkson errors (panel (c)), the differences become quantitatively very important. This is especially pronounced for the unconstrained estimates.

6.2 Estimating the welfare loss of gasoline taxation

The estimates of the demand function can be used to estimate welfare measures such as deadweight loss (DWL). We consider a hypothetical tax change which moves the price from p^0 to p^1 in a discrete fashion (see Blundell et al. (2017)). Let $e(p)$ denote the expenditure function at price p and a reference utility level. The DWL of this price change is then given by

$$L(p^0, p^1) = e(p^1) - e(p^0) - (p^1 - p^0)\, H_\alpha\left[p^1, e(p^1)\right],$$

where $H_\alpha(p, y)$ is the Marshallian demand function. $L(p^0, p^1)$ is computed by replacing e and H with consistent estimates. The estimator of e, $\hat{e}$, is constructed by numerical solution of the differential equation

$$\frac{d\hat{e}(t)}{dt} = \hat{H}_\alpha\left[p(t), \hat{e}(t)\right] \frac{dp(t)}{dt},$$

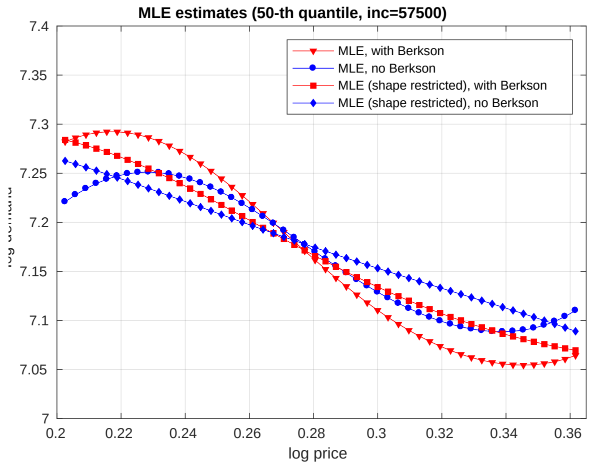
where $[p(t), \hat{e}(t)]$ $(0 \le t \le 1)$ is a price-(estimated) expenditure path.

Deadweight Loss (DWL) estimates are reported in Table 4. Looking at the unconstrained estimates, the table shows the strong quantitative difference in the DWL figures between the estimates with Berkson error (columns (1)-(2)) versus those without (columns (3)-(4)). In many cases, the estimates with Berkson errors but not the Slutsky restriction are more than twice as large as those assuming absence of Berkson errors.

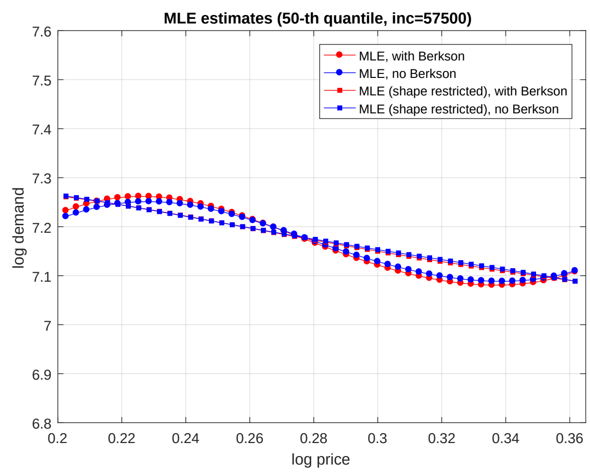
Regarding the constrained estimates, however, the DWL figures are now much closer together and often of similar order of magnitude. This underlines a key point from

Figure 6: Comparison of different magnitudes of the Berkson error

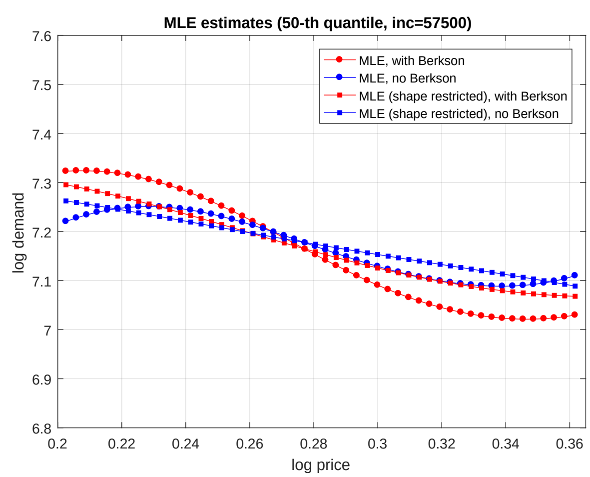
(a) factor 1



(b) factor 0.5



(c) factor 2



Note: The figure compares the baseline estimates in panel (a) to estimates with different standard deviation of the Berkson error. Panel (b) reduces the Berkson error standard deviation by factor 0.5, and panel (c) increases it by factor 2. Estimates shown for the median, at the middle income group. Lines shown in red are estimates accounting for Berkson error, lines shown in blue assume absence of Berkson error. Round markers indicate unconstrained estimates, square markers indicate Slutsky-constrained the demand curve estimates in the previous subsection, the Slutsky constrained demand estimates reduce sensitivity to the presence of Berkson errors.

6.3 Exogeneity test

In this section we report the empirical results for the endogeneity test. To simplify the computation, we implement the univariate version of the test. For this purpose, we stratify the sample along the income dimension in three groups: a low-income group of households (household income between \$35,000 and \$50,000), a middle-income group of households (between \$50,000 and \$65,000), and an upper-income group of households (between \$65,000 and \$80,000). The test is then performed for each income group. The results are shown in Table 5.

We find we do not reject exogeneity for any of the three income groups. This conclusion remains unchanged when we consider moderate variation in the extent of the Berkson error, multiplying the standard error of the Berkson error by a factor of 0.8 and 1.2, respectively, as shown in the table. The critical values shown in the table do not take account of the fact that we perform the test three times (for each of the three income groups). One possibility for adjusting the size for a joint 0.05 level test would be a Bonferroni adjustment. The adjusted p-value for a joint 0.05 level test of exogeneity is $1-(0.95)^{(1/3)} = 0.01695$, at each of the three income groups. Using this more conservative cutoff would strengthen our conclusion. Based on these results, endogeneity is unlikely to be a first-order issue for our estimates.

7 Conclusions

It has long been understood that in a mean regression model with a linear effect of a covariate with Berkson errors and an additive error term, the coefficients in an OLS regression are unbiased. Recent advances in methods, data, as well as computational capacity, together with a desire for understanding the effect of heterogeneity in the studied population, have led to a growing interest in nonlinear models. In nonlinear models, the role of Berkson errors is much less well understood, and ignoring these errors in general leads to a bias in the estimates. This motivates our interest in investigating the effect of Berkson errors, and methods for addressing their presence in the data. We conduct this analysis in the context of a quantile regression model, where the covariates enter through a flexible parametric specification, allowing for potential nonlinearity in the effects. Our application of interest is a gasoline demand model with unobserved heterogeneity, where the price is measured with Berkson error.

The presence of Berkson errors is a frequent feature of economic data. It occurs, for example, when the covariate is measured as a regionally aggregated average, masking within-region variability. The data generating process features the covariate which includes the Berkson error but its error-free value is unobserved by the researcher. This naturally raises the question how much difference recognizing the presence of Berkson error may make.

We derive a maximum likelihood estimator, which enables us to carry out consistent estimation in the presence of Berkson errors with a known density. The paper also develops a test for exogeneity of the Berkson covariate in the presence of an instrument.

We apply the method to the demand for gasoline in the U.S. We examine demand curves in which we impose the Slutsky inequality constraint and those that do not. The unconstrained estimated demand function display non-monotonicity in the price of gasoline. This estimated demand function is substantially affected by Berkson errors. The estimates which do not take account of the Berkson errors understate the variability in the price effect. These results show that accounting for Berkson error can have a substantial effect on the estimated demand function in a standard demand application. In turn, these estimates result in differences in DWL estimates for given price changes. In a number of cases, the DWL estimates recognizing the presence of Berkson errors are more than twice as large as estimates assuming the absence of Berkson errors. Thus, Berkson errors can have quantitatively large effects.

In our application, the estimated demand function is weakly non-monotonic in the price. As Blundell et al. (2012, 2017) explain, this can be due to the effects of random sampling errors on the estimate. We overcome this problem by imposing the Slutsky constraint on the structural demand function estimates, as a way of adding structure to the estimation problem. When the Slutsky restriction is imposed, the estimated demand function is well-behaved and the effects of Berkson errors are greatly attenuated. These results illustrate that in a setting where measurement error increases the uncertainty of the estimates, shape restrictions such as the Slutsky constraint can be particularly useful for providing additional structure to improve the estimation.

Table 4: DWL estimates

	income	with Berkson errors		without Berkson errors	
		DWL per tax	DWL per income	DWL per tax	DWL per income
		(1)	(2)	(3)	(4)
A. Upper quartile (τ=0.75)					
	high	0.132	6.82	0.053	2.98
unconstrained	middle	0.130	7.92	0.055	3.59
	low	0.105	7.79	0.043	3.35
	high	0.117	6.18	0.094	5.14
constrained	middle	0.121	7.44	0.093	5.92
	low	0.114	8.37	0.066	5.04
B. Median (τ=0.50)					
	high	0.126	4.66	0.061	2.39
unconstrained	middle	0.114	4.90	0.062	2.79
	low	0.089	4.60	0.052	2.80
	high	0.108	4.08	0.096	3.66
constrained	middle	0.109	4.70	0.092	4.04
	low	0.105	5.33	0.070	3.66
C. Lower quartile (τ=0.25)					
	high	0.114	2.92	0.077	2.02
unconstrained	middle	0.088	2.65	0.074	2.24
	low	0.067	2.44	0.069	2.51
	high	0.098	2.57	0.102	2.64
constrained	middle	0.093	2.78	0.094	2.80
	low	0.093	3.27	0.083	2.96

Note: DWL shown corresponds to a price change from the 5th to the 95th percentile in the data. Income level 'high' corresponds to \$72,500, 'medium' to \$57,500, and 'low' to \$42,500. 'DWL per income' is re-scaled by $\times 10^4$ for readibility.

Table 5: Exogeneity test

	test statistic	crit value (5%)	p-value	reject?
(a) HIGH INCOME (N=578)				
baseline case	0.1575	0.4000	0.4490	no
reduced Berkson error, factor 0.8	0.1629	0.4000	0.4291	no
increased Berkson error, factor 1.2	0.1443	0.4000	0.5009	no
(b) MEDIUM INCOME (N=555)				
baseline case	0.2257	0.4033	0.2459	no
reduced Berkson error, factor 0.8	0.1879	0.4033	0.3444	no
increased Berkson error, factor 1.2	0.2617	0.4033	0.1781	no
(c) LOW INCOME (N=580)				
baseline case	0.1338	0.4042	0.5427	no
reduced Berkson error, factor 0.8	0.1490	0.4042	0.4799	no
increased Berkson error, factor 1.2	0.1777	0.4042	0.3768	no

Note: Income range 'high' refers to \$65,000-\$80,000, 'medium' to \$50,000-\$65,000, 'low' to \$35,000-\$50,000. Exogeneity test is conducted separately for each income range. Bonferroni-adjusted p-value for a joint 0.05 level test of exogeneity is 0.01695. See text for details.

References

Atkinson, Benjamin, "On Retail Gasoline Pricing Websites: Potential Sample Selection Biases and their Implications for Empirical Research," *Review of Industrial Organization* 33 (2008), 161–175.

Berkson, Joseph, "Are There Two Regressions?" *Journal of the American Statistical Association* 45 (1950), 164–180.

Blundell, Richard, Joel Horowitz, and Matthias Parey, "Nonparametric Estimation of a Nonseparable Demand Function under the Slutsky Inequality Restriction," *Review of Economics and Statistics* 99 (2017), 291–304.

Blundell, Richard, Joel L. Horowitz, and Matthias Parey, "Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation," *Quantitative Economics* 3 (March 2012), 29–51.

Browning, Martin and Jesus Carro, "Heterogeneity and Microeconometrics Modeling," in Blundell, Richard, Whitney Newey, and Torsten Persson (eds.), "Advances in Economics and Econometrics," Cambridge University Press (2007), pp. 47–74.

Cameron, A. Colin and Pravin K. Trivedi, *Microeconometrics. Methods and Applications*, Cambridge University Press (2005).

Delaigle, Aurore, Peter Hall, and Peihua Qiu, "Nonparametric methods for solving the Berkson errors-in-variables problem," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (2006), 201–220.

Dette, Holger, Stefan Hoderlein, and Natalie Neumeyer, "Testing multivariate economic restrictions using quantiles: the example of Slutsky negative semidefiniteness," *Journal of Econometrics* 191 (2016), 129–144.

EIA, "Supplemental Energy-related Data for the 2001 National Household Travel Survey," (2003). Appendix K - Documentation on estimation methodologies for fuel economy and fuel cost. `https://www.eia.gov/consumption/residential/pdf/appendix_k_energy_data.pdf`.

Hausman, Jerry A. and Whitney K. Newey, "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss," *Econometrica* 63 (November 1995), 1445–1476.

Hausman, Jerry A and Whitney K Newey, "Nonparametric welfare analysis," *Annual Review of Economics* 9 (2017), 521–546.

Hoderlein, Stefan and Anne Vanhems, "Estimating the distribution of welfare effects using quantiles," *Journal of Applied Econometrics* 33 (2018), 52–72.

Leckey, Tom and Mark Schipper, "Extending NHTS to Produce Energy-Related Transportation Statistics," (2011). Presentation at National Household Travel Survey Data: A Workshop, June 2011. Available at `http://onlinepubs.trb.org/onlinepubs/conferences/2011/NHTS1/Leckey.pdf`.

ORNL, "2001 National Household Travel Survey. User Guide," (2004). Oak Ridge National Laboratory. Available at `http://nhts.ornl.gov/2001/`.

Pollard, D., *Convergence of Stochastic Processes (Springer Series in Statistics)*, Springer (1984).

Schennach, Susanne M, "Regressions with Berkson errors in covariates – A nonparametric approach," *The Annals of Statistics* 41 (2013), 1642–1668.

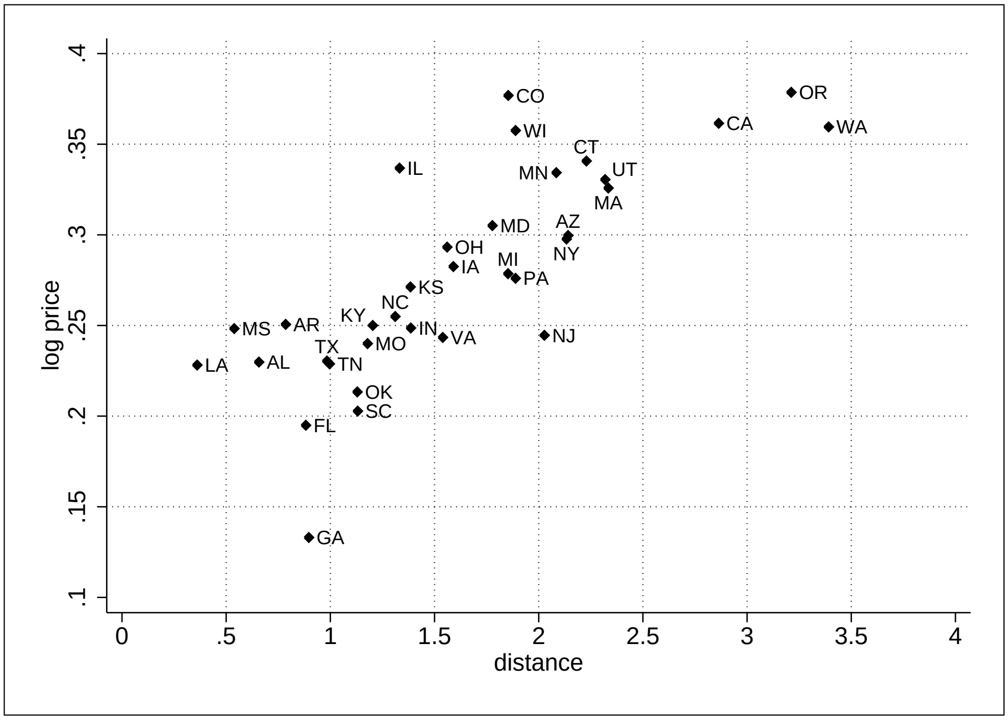
Spokoiny, Vladimir and Mayya Zhilova, "Supplement to: Bootstrap confidence sets under model misspecification," (2015). DOI:10.1214/15-AOS1355SUPP.

Yatchew, Adonis and Joungyeo Angela No, "Household Gasoline Demand in Canada," *Econometrica* 69 (November 2001), 1697–1709.

A Appendix

A.1 Additional Tables and Figures

Figure A.1: Instrumental Variable for Price: Distance to the Gulf of Mexico



Note: Price of gasoline and distance to the Gulf of Mexico. Distance to the respective state capital is measured in 1000 km. Source: BHP (2012, Figure 5).

A.2 Exogeneity Test

The argument that follows uses linear functional notation. In this notation,

$$Pg = \int gdP; \quad P_n g = \int gdP_n$$

for any function $g(\cdot)$, where P and P_n, respectively, are the distribution and empirical distribution functions of the random argument of g.

To obtain an asymptotic approximation to the distribution of T_n, make:

Assumption 1. *(i) G_{EX}^{-1} is a known bounded function $g(\cdot,\cdot,\cdot,\theta)$, where $\theta \in \mathbb{R}^d$ for some $d < \infty$ is a constant parameter whose maximum likelihood estimate is denoted by $\hat{\theta}$ and whose true but unknown population value is denoted by θ_0.*

(ii) $n^{1/2}\left(\hat{\theta} - \theta_0\right) \to^d N(0, V)$ for some non-singular covariance matrix V.

(iii) The first and second derivatives of g with respect to its third argument are bounded and continuous uniformly over θ in a neighborhood of θ_0 and the other arguments of g.

Assumption 2. *(i) K is a probability density function that is symmetrical about 0 and supported on $[-1,1]$.*

(ii) $n^{1/2}h/(\log n)^{\gamma} \to \infty$ as $n \to \infty$ for some $\gamma > 1/2$.

Define

$$G_{EX}^{-1}(\cdot,\cdot,\cdot) = g(\cdot,\cdot,\cdot,\theta).$$

Define

$$\begin{aligned} R_n(y,w,\varepsilon) =& \frac{1}{nh^2}\sum_{i=1}^{n} I\left[\hat{G}_{EX}^{-1}\left(P_i+\varepsilon, Y_i, Q_i\right) \leq \tau\right] K\left(\frac{W_i - w}{h}\right) K\left(\frac{Y_i - y}{h}\right) \\ =& \frac{1}{h^2} P_n \left\{ I\left[\hat{G}_{EX}^{-1}\left(P+\varepsilon, Y, Q\right) \leq \tau\right] K\left(\frac{W - w}{h}\right) K\left(\frac{Y - y}{h}\right)\right\}. \end{aligned}$$

Define

$$\begin{aligned} R_{n1}(y,w,\varepsilon) = h^{-2}\left(P_n - P\right) \Big\{ \Big(& I\left[\hat{G}_{EX}^{-1}\left(P+\varepsilon, Y, Q\right) \leq \tau\right] - I\left[G_{EX}^{-1}\left(P+\varepsilon, Y, Q\right) \leq \tau\right]\Big) \\ & K\left(\frac{W - w}{h}\right) K\left(\frac{Y - y}{h}\right)\Big\} \end{aligned}$$

and

$$\begin{aligned} R_{n2}(y,w,\varepsilon) =& h^{-2} P \left\{ I\left[\hat{G}_{EX}^{-1}\left(P+\varepsilon, Y, Q\right) \leq \tau\right] K\left(\frac{W - w}{h}\right) K\left(\frac{Y - y}{h}\right)\right\} \\ & + h^{-2}\left(P_n - P\right) \left\{ I\left[G_{EX}^{-1}\left(P+\varepsilon, Y, Q\right) \leq \tau\right] K\left(\frac{W - w}{h}\right) K\left(\frac{Y - y}{h}\right)\right\}. \end{aligned}$$

Then $R_n = R_{n1} + R_{n2}$. In linear functional notation, $\hat{G}_{EX}^{-1}$ is treated as a fixed (non-random) function in the integrals.

Under Assumption 1, $\hat{G}_{EX}^{-1} - G_{EX}^{-1} = O_p\left(n^{-1/2}\right)$. Therefore, it follows from Lemma 2.37 of Pollard (1984) that

$$R_{n1}(y,w,\varepsilon) = O_p\left[\frac{(\log n)^{\gamma}}{nh}\right]$$

uniformly over (y, w, ε). It further follows that

$$\begin{aligned}
R_n(y,w,\varepsilon) &= h^{-2} P\left\{ I\left[\hat{G}_{EX}^{-1}(P+\varepsilon,Y,Q) \le \tau\right] K\left(\frac{W-w}{h}\right) K\left(\frac{Y-y}{h}\right)\right\} \\
&\quad + h^{-2}(P_n - P)\left\{ I\left[G_{EX}^{-1}(P+\varepsilon,Y,Q) \le \tau\right] K\left(\frac{W-w}{h}\right) K\left(\frac{Y-y}{h}\right)\right\} + O_p\left[\frac{(\log n)^\gamma}{nh}\right] \\
&= h^{-2} P\left\{ I\left[\hat{G}_{EX}^{-1}(P+\varepsilon,Y,Q) \le \tau\right] - I\left[G_{EX}^{-1}(P+\varepsilon,Y,Q) \le \tau\right]\right\} K\left(\frac{W-w}{h}\right) K\left(\frac{Y-y}{h}\right) \\
&\quad + h^{-2} P_n\left\{ I\left[G_{EX}^{-1}(P+\varepsilon,Y,Q) \le \tau\right] K\left(\frac{W-w}{h}\right) K\left(\frac{Y-y}{h}\right)\right\} + O_p\left[\frac{(\log n)^\gamma}{nh}\right] \\
&\equiv \; R_{n3}(y,w,\varepsilon) + R_{n4}(y,w,\varepsilon) + O_p\left[\frac{(\log n)^\gamma}{nh}\right].
\end{aligned}$$

Under Assumption 1, $(\hat{\theta} - \theta_0) = O_p(n^{-1/2})$. It follows from standard arguments for kernel estimators that $R_{n3}(y, w, \varepsilon) = O_p(n^{-1/2})$ uniformly over (y, w, ε). Therefore, by Assumption 2,

$$R_n(y,w,\varepsilon) = R_{n4}(y,w,\varepsilon) + O_p(n^{-1/2}) \tag{5}$$

uniformly over (y, w, ε).

Now consider $R_{n4}(y, w, \varepsilon)$. Because $U = G_{EX}^{-1}(P + \varepsilon, Y, Q)$,

$$\begin{aligned}
R_{n4}(y,w,\varepsilon) &= h^{-2} P_n\left[I(U \le \tau) K\left(\frac{W-w}{h}\right) K\left(\frac{Y-y}{h}\right)\right], \\
R_{n4}(y,w,\varepsilon) - \tau \hat{f}_{YW}(y,w) &= h^{-2} P_n\left\{ \left[I(U \le \tau) - \tau\right] K\left(\frac{W-w}{h}\right) K\left(\frac{Y-y}{h}\right)\right\},
\end{aligned}$$

and

$$P\left[R_{n4}(y,w,\varepsilon) - \tau \hat{f}_{YW}(y,w)\right] = 0. \tag{6}$$

Therefore, $R_{n4}(y, w, \varepsilon) - \tau \hat{f}_{YW}(y, w)$ is a mean-zero stochastic process. The covariance function of this process is $[C(y_1, w_1; y_2, w_2) + o(1)]/(nh^2)$, where

$$C(y_1,w_1;y_2,w_2) = \tau(1-\tau) f_{YW}(y_1,w_1) \int K(\xi)K(\xi+\delta_W)K(\zeta)K(\zeta+\delta_Y)\,d\xi d\zeta,$$

where $\delta_W = (w_1 - w_2)/h$ and $\delta_Y = (y_1 - y_2)/h$. It follows from (5) and (6) that

$$S_n(y,w) - \tau \hat{f}_{YW}(y,w) = \frac{1}{h^2} P_n\left\{ \left[I(U \le \tau) - \tau\right] K\left(\frac{W-w}{h}\right) K\left(\frac{Y-y}{h}\right)\right\} + O_p\left(n^{-1/2}\right).$$

Define the stochastic process

$$
\begin{aligned}
Z_n(y,w) &= n^{1/2}h^{-1}P_n\left\{[I(U\le\tau)-\tau]\,K\left(\frac{W-w}{h}\right)K\left(\frac{Y-y}{h}\right)\right\}\\
&= \frac{1}{n^{1/2}}h^{-1}\sum_{i=1}^{n}[I(U_i\le\tau)-\tau]\,K\left(\frac{W_i-w}{h}\right)K\left(\frac{Y_i-y}{h}\right)\\
&= n^{1/2}h[S_n(y,w)-\tau\hat{f}_{YW}(y,w)]+O_p(h).
\end{aligned}
$$

Let $\{\psi_j : j=1,2,\dots\}$ be the eigenfunctions of $C(y_1,w_1;y_2,w_2)$ and $\{\lambda_{nj} : j=1,2,...\}$ be the eigenvalues. The ψ_j's form a complete, orthonormal basis for $L_2[-1,1]^2$. $Z_n(y,w)$ has the representation

$$
Z_n(y,w)=\sum_{k=1}^{\infty}\hat{b}_{nk}\psi_k(y,w)
$$

where

$$
\hat{b}_{nk}=\int Z_n(y,w)\psi_k(y,w)dydw.
$$

Moreover,

$$
E\hat{b}_{nk}=0
$$

and

$$
E(\hat{b}_{nk}\hat{b}_{nl})=\lambda_{nk}\delta_{kl}+o(1)
$$

for all k and l, where δ_{kl} is the Kronecker delta. In addition,

$$
T_n=\sum_{k=1}^{\infty}\hat{b}_{nk}^2.
$$

Let $\{L_n : n=1,2,\dots\}$ be an increasing sequence of positive constants such that $L_n\to\infty$ as $n\to\infty$. Define

$$
\tilde{T}_n=\sum_{k=1}^{L_n}\hat{b}_{nk}^2.
$$

Then

$$|\tilde{T}_n - T_n| \to^p 0.$$

Let V_{L_n} denote the $L_n \times L_n$ diagonal matrix whose (l, l) element is λ_{nl}. Let ω be a $L_n \times 1$ random vector with the $N(0, V_{L_n})$ distribution, and let $\|\cdot\|$ denote the Euclidean norm. It follows from Theorem A.1 of Spokoiny and Zhilova (2015) that for any $z > max(4, L_n)$ and some constant $C_4 < \infty$,

$$\left| \; P\left(\tilde{T}_n \leq z\right) - P\left(\|\omega\|^2 \leq z\right) \right| \leq C_4 n^{-1/2} L_n^{3/2}.$$

Assume that $n^{-1/2} L_n^{3/2} \to 0$ as $n \to \infty$. Then

$$P\left(T_n \leq z\right) - P\left(\|\omega\|^2 \leq z\right) \to 0$$

as $n \to \infty$, and the distribution of T_n can be approximated by that of $\|\omega\|^2$. This is

$$\|\omega\|^2 = \sum_{j=1}^{L_n} \lambda_{nj} \chi_j^2,$$

where the χ_j^2s are independent random variables that are distributed as chi-square with one degree of freedom. Estimate the λ_{nj}'s by the eigenvalues of the empirical covariance operator of Z_n.

Table A.1: DWL estimates with confidence intervals

	with Berkson errors		without Berkson errors	
income	DWL per tax	DWL per income	DWL per tax	DWL per income
	(1)	(2)	(3)	(4)
A. Upper quartile (τ=0.75)				
high	0.132	6.82	0.053	2.98
	[0.062; 0.209]	[3.787; 10.645]	[-0.012; 0.100]	[-0.313; 5.583]
middle	0.130	7.92	0.055	3.59
	[0.066; 0.199]	[4.624; 11.925]	[-0.002; 0.103]	[0.310; 6.709]
low	0.105	7.79	0.043	3.35
	[0.038; 0.210]	[3.442; 15.530]	[-0.022; 0.116]	[-1.301; 9.202]
B. Median (τ=0.50)				
high	0.126	4.66	0.061	2.39
	[0.043; 0.193]	[2.022; 7.074]	[-0.003; 0.118]	[0.142; 4.616]
middle	0.114	4.90	0.062	2.79
	[0.052; 0.177]	[2.500; 7.555]	[0.011; 0.117]	[0.730; 5.258]
low	0.089	4.60	0.052	2.80
	[0.025; 0.189]	[1.554; 9.847]	[-0.007; 0.125]	[-0.149; 6.745]
C. Lower quartile (τ=0.25)				
high	0.114	2.92	0.077	2.02
	[0.017; 0.183]	[0.711; 4.668]	[0.011; 0.155]	[0.471; 4.090]
middle	0.088	2.65	0.074	2.24
	[0.016; 0.154]	[0.640; 4.590]	[0.021; 0.145]	[0.757; 4.407]
low	0.067	2.44	0.069	2.51
	[-0.028; 0.172]	[-0.609; 6.351]	[-0.024; 0.151]	[-0.462; 5.496]

Note: Table shows unconstrained DWL estimates with 90% confidence intervals, based on 499 bootstrap replications. DWL shown corresponds to a price change from the 5th to the 95th percentile in the data. Income level ‘high’ corresponds to \$72,500, ‘medium’ to \$57,500, and ‘low’ to \$42,500. ‘DWL per income’ is re-scaled by $\times 10^4$ for readibility. See text for details.