A Semiparametric Analysis of Gasoline Demand in the US: Reexamining The Impact of Price

Manzan, sebastiano and Zerom, Dawit
California State University Fullerton

December 2008

A Semiparametric Analysis of Gasoline Demand in the US: Reexamining The Impact of Price

Sebastiano Manzan[†] and Dawit Zerom[‡]

[†] *Department of Economics and Finance, Baruch College (CUNY), USA*

[‡] *Mihaylo College of Business and Economics*

California State University at Fullerton, USA

Abstract

The evaluation of the impact of an increase in gasoline tax on demand relies crucially on the estimate of the price elasticity. This paper presents an extended application of the Partially Linear Additive Model (PLAM) to the analysis of gasoline demand using a panel of US households, focusing mainly on the estimation of the price elasticity. Unlike previous semi-parametric studies that use household-level data, we work with vehicle-level data within households that can potentially add richer details to the price variable. Both households and vehicles data are obtained from the Residential Transportation Energy Consumption Survey (RTECS) of 1991 and 1994, conducted by the US Energy Information Administration (EIA). As expected, the derived vehicle-based gasoline price has significant dispersion across the country and across grades of gasoline. By using a PLAM specification for gasoline demand, we obtain a measure of gasoline price elasticity that circumvents the implausible price effects reported in earlier studies. In particular, our results show the price elasticity ranges between -0.2, at low prices, and -0.5, at high prices, suggesting that households might respond differently to price changes depending on the level of price. In addition, we estimate separately the model to households that buy only regular gasoline and those that buy also midgrade/premium gasoline. The results show that the price elasticities for these groups are increasing in price and that regular households are more price sensitive compared to non-regular.

Keywords: semiparametric methods, partially linear additive model, gasoline demand.
JEL codes: C14, D12

Forthcoming at **ECONOMETRIC REVIEWS.**

1 Introduction

A recent report by the US Department of Energy (2004) estimates that fuel consumption in 2003 contributed to 32% of US and 7.5% of world emissions of carbon dioxide. Thus, policies aimed at decreasing gasoline demand are likely to have a noticeable impact in addressing the environmental consequences of emissions of carbon dioxide and local air pollutants. Two recent studies by the US Congressional Budget Office (2002, 2003) examine different policy instruments; namely, increasing the standards for the average fuel economy of vehicles, gasoline taxes, and programs of cap-and-trade[1]. Comparing the costs and benefits of the three instruments, the studies conclude that increasing gasoline taxes might be the most effective way to influence demand. A higher gasoline tax would affect fuel demand in the short term and also encourage households to replace the stock of vehicles with more efficient ones in the longer run. In addition, it would spread the cost of the tax increase between producers and consumers (of gasoline) and encourage different gas-reduction activities. Price elasticity plays an important role in evaluating the impact of gasoline tax. Consequently, there has been a considerable amount of research interest in the estimation of gasoline demand models that focus mainly on the estimation of price elasticity. Dahl and Sterner (1991) and Graham and Glaister (2002) provide extensive surveys of the literature on the estimation of gasoline price elasticity. Empirical evidence from both cross-sectional and time series studies generally suggest that the price elasticity demand for gasoline is estimated in the range between -0.5 and -1.1. However, studies considering more recent data typically find lower estimates. Based on household data from the late 1980s and 1990s, **?** and Nicol (2003) estimated the price elasticity of gasoline demand in the range between -0.2 to -0.4. A study by the US Department of Energy (1996) provides a price elasticity value of -0.38, and this value is adopted by the Congressional Budget Office (2002, 2003) in evaluating the impact of an increase in gasoline tax. **?** estimate a structural model on a panel of U.S. states (for the period 1966 to 2001) and estimate a (long-run) price elasticity of gasoline demand between -0.33 and -0.42[2]. In a recent paper that assesses the optimal level of taxation in US, Parry and Small (2005) uses a price elasticity of -0.55 as a compromise between recent low and past high estimates.

[1] In this case the government fixes a limit to the emission of carbon dioxide and producers or importers of gasoline are allowed to trade allowances for the emissions deriving from the consumption of their gasoline sales.

[2] The elasticity in their model is a function of income and fuel price (among others). If these variables are set at their average values they obtain an elasticity of -0.42 while it is lower when income and fuel price are set at the 1997-2001 average value.

By carefully addressing some data issues, we provide new empirical results on the analysis of US gasoline demand, focusing mainly on the price elasticity. We analyze household data (including the vehicle-level information) from the Residential Transportation Energy Consumption Surveys (RTECS) of 1991 and 1994. RTECS has been administered by the Energy Information Administration (EIA) from 1979 until 1994, when it was terminated for budgetary reasons. Using the 1988 and 1991 RTECS data, Schmalensee and Stoker (1999) find some relevant nonlinearities when modeling the gasoline demand by using partially linear models. They allowed the income and age variables to have a general nonparametric shape while the other control variables being linear (demographic and location variables). Within the partially linear framework, Schmalensee and Stoker (1999) also consider gasoline price to have a nonparametric effect on demand. However, they obtain a price function that is upward sloping for a range of fuel prices in the middle of the distribution and is negatively sloped in the rest of the interval of variation. Using similar semiparametric techniques, Hausman and Newey (1995) also found a similar effect for the pooled RTECS from 1979 until 1981. Puzzled by this "implausible" price effect and further scrutinizing the price data in RTECS, Schmalensee and Stoker (1999) argue that the price variable provided in RTECS is unreliable. As a proxy for the price variable (per household), RTECS assigns each household an average fuel cost per gallon purchased, where the total expenditure is determined using average regional gasoline prices. This procedure assumes that all the households living in a broadly defined area such as a region (e.g., the Mid-West) face the same gasoline price.

While the immediate goal of our paper is to address the empirical problem raised by Schmalensee and Stoker (1999), the paper has a much wider scope. The main contributions of the paper are outlined as follows. First, we tackle the problem of estimating price elasticity from RTECS household data. In a follow up study to Schmalensee and Stoker (1999), Yatchew and No (2001) use Canadian household data from the National Private Vehicle Use Survey, conducted by Statistics Canada between October 1994 and September 1996. Using the "complete" price data and applying a similar semiparametric specification as in Schmalensee and Stoker (1999), Yatchew and No (2001) obtain plausible nonparametric price elasticity. In this paper, we exploit instead the detailed information on the "vehicles" owned by households as reported in the RTECS. Such details include the type of vehicle(s), type and grade (regular, midgrade, or premium) of gasoline purchased, and the price of the last fuel purchase. By carefully studying these detailed information, we are able to assign to households an average (over the vehicles) gasoline price that maintains the geographical variability in gasoline prices (compared to the RTECS

procedure that destroys this variability). Unlike the price variable in RTECS (used in Schmalensee and Stoker, 1999), the derived vehicle-based gasoline price has significant dispersion across regions and across grades of gasoline.

Second, we use the partially linear additive model (hereinafter PLAM) as a reduced form model for the gasoline demand. PLAM is a semiparametric specification in the sense that it involves both a nonparametric and a parametric (linear) part. Compared to the partially linear model proposed by Robinson (1988) and applied to gasoline demand by Schmalensee and Stoker (1999), the model assumes additivity of the nonparametric component. Introducing this assumption delivers more efficient estimates of the parametric effects and easier interpretation of the relationship among the variables that enter nonparametrically. In addition, The PLAM set-up also allows interactions among the variables of the nonparametric part by incorporating them within the linear part. We estimate the model following the kernel-based approach proposed by Manzan and Zerom (2005). The resulting estimator of the linear parameters are root-n consistent and asymptotically normal distributed. A convenient feature of this estimator is that it is semiparametric efficient in the sense of Chamberlain (1992) (when the error is homoscedastic). This is an attractive feature compared to other kernel-based estimators (e.g., Fan *et al.*, 1998; Fan and Li, 2003; and Moral and Rodriguez-Poo 2004). In our gasoline demand analysis, the linear part includes up to 20 demographic and location variables (these are mainly dummy and discrete variables) while the nonparametric part contains $\log price$, $\log age$ and $\log income$. The nonparametric treatment of the price effect is able to show that our vehicle-based gasoline price solves the implausible price effect that arises when the price provided in RTECS is used.

Focusing on the price effect, the main empirical findings of the paper can be summarized as follows. The partial nonparametric price effect is appropriately downward sloping, and the corresponding elasticity (the derivative of the price effect curve) ranges between -0.2, at low prices, and -0.5, at high price values. This result suggests that households might respond differently to price changes depending on the level of the fuel price. The availability of the vehicle information allows us to further investigate this issue by considering separately the households that consume only "regular" gasoline and those that purchase "non-regular" grades of gasoline[3]. The estimation results for the two groups show that regular users are more sensitive to price changes (estimated elasticity of -0.54) compared to non-regular users (that have an elasticity of -0.33). The price elasticity of regular

[3]A household with more than one car might use midgrade or premium for one vehicle and regular for the others or they might use midgrade or premium fuel for all vehicles.

gasoline has a tendency to increase from -0.3 toward -0.7 at high prices. Instead, the demand for "non-regular" fuel is quite inelastic at low prices and becomes increasingly reactive at high prices. This is an interesting result since it provides evidence on the different characteristics and behavior of households buying regular and non-regular gasoline. Separate analysis of the two groups also shows some relevant differences in the effects of income, age and number of drivers in the household.

The remainder of the paper is organized as follows. In Section (2), we describe the semi-parametric method for the estimation of the PLAM. In Section (3), we apply the PLAM to investigate the US gasoline demand based on household-level vehicles data from the RTECS. Several empirical results are also discussed. Finally, Section (4) concludes the paper.

2 Description of the methodology

Semi-parametric methods have become increasingly popular in empirical work. The widespread acceptance of these methods derives from their flexible specification, which allows for some variables to be linearly related to the dependent variable without imposing stringent restrictions on other variables whose relationship may be difficult to parameterize. These models allow for a more general specification compared to the linear regression model, while retaining ease of interpretability. Various demand studies have successfully employed semi-parametric methods to tackle the problem of finding appropriate ways of modeling the effects of expenditure on consumer demand (e.g., Blundell and Duncan (1998), and Blundell *et al.* (1998)). There has also been growing interest in the application of semi-parametric methods to analyze the demand for gasoline in U.S. and Canada based on household survey data (e.g., Hausman and Newey (1995), Schmalensee and Stoker (1999), Coppejans (2003) and Yatchew and No (2001)).

In this paper, we consider the partially linear additive model (hereinafter PLAM) which has the following form

$$Y_i = \beta_0 + X_i'\beta + m_1(Z_{1i}) + \ldots + m_q(Z_{qi}) + u_i \qquad (i = 1, \ldots, n), \tag{1}$$

where Y_i is a scalar dependent variable, β_0 is a scalar parameter, X_i is a $p \times 1$ vector of explanatory variables, $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of unknown parameters, $Z_i = (Z_{1i}, \ldots, Z_{qi})'$ is a $q \times 1$ vector of explanatory variables, $m_1(\cdot), \cdots, m_q(\cdot)$ are unknown real-valued smooth functions, and u_i is an unobservable random variable that

satisfies $E[u_i|X_i, Z_i] = 0$. The semiparametric structure of the model derives from the linearity assumption for the effect of the X_i variables, while the Z_i's are not restricted to any particular functional form. The model is partially additive in the sense that the nonparametric part is characterized by the sum of the $m_j(Z_{j,i})$ rather than being a fully nonparametric function of all the Z_i variables. The additive structure helps reduce the curse of dimensionality problem because the additive components can be estimated at the one-dimensional nonparametric rate. Moreover, unlike the purely additive model, the PLAM allows interaction terms among the elements of Z_i enter the linear part of the model. This is possible as PLAM permits X_i to be a deterministic, but non additive, function of Z_i.

Various methods have been proposed to estimate the parametric part of the PLAM. Recent approaches include those of Fan *et al.* (1998), Fan and Li (2003), Moral and Rodriguez-Poo (2004) and Hengartner and Sperlich (2005) using kernel-based methods, while Li (2000) introduced a series-based estimator. In this paper, we follow the kernel-based approach of Manzan and Zerom (2005). This estimator has two advantages compared to these alternative estimators. First, it achieves the semiparametric efficiency bound (Chamberlain, 1992) of the partially linear additive model under the assumption of homoscedastic errors. In addition, it is computationally more efficient because it requires $\mathcal{O}(n^2)$ operations, while marginal integration estimators involve an increase of computations by the order of the sample size n. In the rest of the Section we briefly describe the estimation method and refer to Manzan and Zerom (2005) for a more detailed discussion.

Assume the additive components in (1) satisfy the identification assumption $E[m_j(Z_{ji})] = 0$ for all $j = 1, \ldots, q$. Denote by Z_{ji} the j-th element of Z_i and W_{ji} the set of all Z_i variables excluding $Z_{j,i}$, i.e. $W_{ji} = (Z_{1,i}, \ldots, Z_{j-1,i}, Z_{j+1,i}, \ldots, Z_{q,i})'$. Define a generic instrument function $\phi(z_j, w_j)$ as follows,

$$\phi(z_j, w_j) = \frac{p_z(z_j)p_w(w_j)}{p(z_j, w_j)}$$

where $p_z(\cdot)$ and $p_w(\cdot)$ represent the density functions of Z_{ji} and W_{ji}, respectively, and $p(\cdot)$ is the joint probability function of $Z = (Z_j, W_j)$. The function $\phi(z_j, w_j)$ has the following properties: (1) $E[\phi(Z_{ji}, W_{ji})|Z_{ji} = z_j] = 1$, and (2) $E[\phi(Z_{ji}, W_{ji})m_k(Z_{ki})|Z_{ji} = z_j] = 0$ for $k \neq j$. Then, multiplying each side of Equation (1) by the above instrument and

taking conditional expectations on $Z_{ji} = z_j$, we obtain

$$Y^*_{i,j} = m_j(Z_{ji}) + (X^*_{i,j})'\beta \quad (j = 1, \ldots, q). \tag{2}$$

where Y^*_{ji} and X^*_{ji} denote the $E[\phi(Z_{ji}, W_{ji})Y_i|Z_{ji} = z_j]$ and $E[\phi(Z_{ji}, W_{ji})X_i|Z_{ji} = z_j]$, respectively. Adding the above q-equations in (2) and subtracting the result from (1) gives

$$Y_i - Y^*_i = (X_i - X^*_i)'\beta + u_i, \tag{3}$$

where $Y^*_i = \sum_j^q Y^*_{i,j}$ and $X^*_i = \sum_j^q X^*_{i,j}$. This Equation shows that the role of the function $\phi(z_j, w_j)$ is to reduce the PLAM in Equation (1) to a linear-like model. Then, an estimator of β can simply be derived by OLS regression of the deviation $Y_i - Y^*_i$ on $X_i - X^*_i$[4].

The estimation of β depends on Y^*_i and X^*_i that are unknown quantities. Manzan and Zerom (2005) propose replacing these quantities by their kernel estimators. Let $\hat{A}^*_i = \sum_{j=1}^q \hat{A}^*_{i,j}$, denotes an estimator of A^*_i (where A^*_i is either Y^*_i or X^*_i). The kernel-based estimator of $\hat{A}^*_{i,j}$ is

$$\hat{A}^*_{i,j} = \frac{1}{(n-1)b} \sum_{\ell \neq i}^{n} K\Big(\frac{Z_{j\ell} - Z_{ji}}{b}\Big) \frac{\hat{p}_w(W_{j\ell})}{\hat{p}(Z_{j\ell}, W_{j\ell})} A_\ell \quad (i = 1, \ldots, n; \quad j = 1, \ldots, q), \tag{4}$$

where $K(\cdot)$ is a kernel function, b is a bandwidth (or smoothing parameter), and $\hat{p}_w(\cdot)$ and $\hat{p}(\cdot)$ are kernel-smoothers of the corresponding densities. Note that the $\hat{A}^*_{i,j}$ is a leave-out estimator in the sense that the i-th observation (A_i, Z_i) is not used in the estimation. The estimator of β is obtained by OLS regression of $Y_i - \hat{Y}^*_i$ on $X_i - \hat{X}^*_i$. Under some regularity conditions, Manzan and Zerom (2005) show that $\hat{\beta}$ is $n^{1/2}$-consistent and asymptotically normally distributed.

The implementation of the kernel smoothers $\hat{Y^*}_i$ and $\hat{X^*}_i$ requires choices to be made on both the bandwidth b and the type of kernel function $K(\cdot)$. We use bandwidths b that decrease to 0 at the rate $n^{-2/7}$ and a standard Gaussian kernel function. The above rate for b and the choice of the Gaussian kernel are consistent with Assumption A2 for $q < 4$ (see Manzan and Zerom, 2005). In the application to be discussed in section (3), $q < 4$ and hence the above choices are optimal. In addition, we allow b to adapt to the variability of the variable Z_{ji}. Hence, the bandwidth is given by $b_j = a\ \sigma_j\ n^{-2/7}$, where

[4]In empirical work, one may also be interested in estimating the intercept β_0. It is easy to see that when $\beta_0 \neq 0$, equation (3) would become $Y_i - Y^*_i = (1-q)\beta_0 + (X_i - X^*_i)'\beta + u_i$. Hence, we would instead regress $(Y_i - Y^*_i)$ on $(1, (X_i - X^*_i)')'$ so as to incorporate the estimation of the intercept.

σ_j denotes the standard deviation of Z_{ji}. Using this argument, the problem of bandwidth choice reduces to the choice of a. To select this, we use a cross-validation (CV) procedure over different values of a[5].

Now, we discuss how one can estimate the additive non-parametric components of the PLAM. Based on (2) and using the estimator $\hat{\beta}$, we can compute $\hat{m}_j(\cdot)$ as

$$\hat{m}_j(Z_{ji}) = \hat{Y}^*_{i,j} - (\hat{X}^*_{i,j})'\hat{\beta} \quad (j = 1, \ldots, q), \tag{5}$$

where $\hat{A}^*_{i,j}$ (A can be Y or X) is defined in (4). Because $\hat{\beta} = \beta + O_p(n^{-1/2})$ and this rate is surely faster than the possible rates of convergence of the kernel smoothers $\hat{Y}^*_{i,j}$ and $\hat{X}^*_{i,j}$, the asymptotic distribution of the additive components $\hat{m}_j(\cdot)$ will remain unaffected by the estimation of β. In this way, the estimation of β and that of the additive nonparametric components can be done in a single step without a need for extra computations to recover the additive components.

However, the estimation of the nonparametric components as in (5) does not lead to efficient estimates. Using the terminology in Linton (1996) and Kim *et al.* (1999), the additive estimates are oracle inefficient. They are inefficient in the sense that if

$$m_1(z_1), m_2(z_2), \cdots, m_{j-1}(z_{j-1}), m_{j+1}(z_{j+1}), \cdots, m_q(z_q)$$

were known, $m_j(z_j)$ could be estimated with a smaller variance. Because the empirical results of this paper are highly dependent on the precise estimation of the nonparametric components, ensuring their efficiency is vital. For example, the price effect (the main focus of the paper) will be modeled as being nonparametric in section (3). Following the approach of Kim *et al.* (1999), we implement a one-step backfitting procedure in order to attain efficiency. First, use $\hat{\beta}$ to compute $\hat{Y}_i = Y_i - X_i'\hat{\beta}$. Second, for each $j \in (1, 2, \cdot, q)$, compute partial residuals $\hat{\varepsilon}^j_i = \hat{Y}_i - \sum_{k \neq j} \hat{m}_k(Z_{ki})$ where the $\hat{m}_k(\cdot)$ estimates are obtained from (5). Finally, apply a local linear smoothing of $\hat{\varepsilon}^j_i$ on Z_{ji}. Let's denote the resulting nonparametric component estimators by $\hat{m}^e_j(\cdot)$. It should be noted that in the implementation of the one-step backfitting, one needs to choose a different bandwidth (other than the ones used in the computation of $\hat{m}_j(\cdot)$) for $\hat{m}^e_j(\cdot)$. The asymptotic theory of

[5]The CV procedure selects a to minimize the following quantity,

$$\hat{a} = \min_a \sum_{i=1}^{n} \{(Y_i - \hat{Y}^*_i) - (X_i - \hat{X}^*_i)'\hat{\beta}\}^2$$

where $\hat{Y}^*_i$ and $\hat{X}^*_i$ are leave-out estimators in Equation (4) where the the i-th observation is not used in the estimation. The motivation for the above minimization step comes from the formulation in (3).

local linear smoothing suggests that the bandwidth be chosen as $\sim cn^{-1/5}$. Following this, and allowing different smoothing for different j, we choose the corresponding bandwidth of $\hat{m}_j^e(\cdot)$ by $c\,\sigma_j\,n^{-1/5}$ where σ_j is the standard deviation of the variable Z_{ji}. In section (3), we have experimented with several values of c before settling for a final value.

Finally, we outline a procedure for calculating point-wise confidence intervals of the non-parametric estimates $\hat{m}_j^e(\cdot)$. Because the asymptotic variance of $\hat{m}_j^e(\cdot)$ is a very complicated function of unknown quantities (see Kim *et al.*, 1999), we use the alternative route of bootstrap methods. Given $\hat{\beta}$ and $\hat{m}_j^e(\cdot)$, the residuals of the PLAM in Equation (1) are given by

$$\widehat{u}_i = Y_i - X_i'\hat{\beta} - \sum_{j=1}^{q} \hat{m}_j^e(Z_{ji}). \tag{6}$$

We resample the residuals according to the wild bootstrap method of Liu (1988). This consists of drawing from the centered residuals, $\tilde{u}_i = \hat{u}_i - \frac{1}{n}\sum_i \widehat{u}_i$, according to the following scheme

$$\tilde{u}_{i,s} = \begin{cases} \alpha\tilde{u}_i & \text{with probability } p = (\sqrt{5}+1)/(2\sqrt{5}) \\ \gamma\tilde{u}_i & \text{with probability } 1-p \end{cases}$$

where $\alpha = (\sqrt{5}-1)/2$, $\gamma = (\sqrt{5}+1)/2$, and s indicates the number of bootstrap replications ($s = 1, \cdots, S$). A bootstrap replicate is then obtained as follows

$$Y_{i,s} = X_i'\hat{\beta} + \sum_{j=1}^{q} \hat{m}_j^e(Z_{j,i}) + \tilde{u}_{i,s}.$$

For each replicate $(X_i, Z_i, Y_{i,s})$, we compute the nonparametric component (denoted by $m_j^{e,s}(z_j)$) at fixed values $Z_{ji} = z_j$. Then, bootstrap confidence interval for $m_j(z_j)$ is simply calculated using the appropriate percentiles of $\{m_j^{e,s}(z_j)\}_{s=1}^{S}$.

3 Empirical Results

In this section we investigate the US demand for gasoline using household-level data from the RTECS of 1991 and 1994. A study by Schmalensee and Stoker (1999) applies a partially linear model for the pooled 1988 and 1991 samples and is able to uncover some interesting empirical regularities. We complement their analysis in at least two important aspects. First, we use the PLAM set-up as a reduced form model for gasoline demand. To the extent that PLAM is a plausible specification for modeling gasoline demand, our theoretical result suggests that ignoring additivity will lead to a less efficient estimator of the linear parameters. Furthermore, additivity facilitates easy interpretation

of non-parametric estimates. Second, Schmalensee and Stoker (1999) concluded, using their semiparametric approach, that the price data given in RTECS could not be used to estimate the price effect (or price elasticity). We address this data problem by deriving an alternative price variable.

Table (1) provides a summary of the descriptive statistics of the variables of interest. In the Appendix, we provide details of how the data were constructed. The 1991 and the 1994 survey data comprise a total of 3045 and 3002 households, respectively. In our analysis, we remove those households that have zero miles driven, gallons consumed, number of drivers and vehicles owned. The resulting dataset has 2697 observations in 1991 and 2563 in 1994. The means and standard deviations of the continuous variables do not vary significantly between the two surveys. However, the discrete variables show some differences between the surveys. The fraction of households living in urban areas increases from 28.4% to 42.4% while those of both suburban and rural areas become lower. This is due to the change of the area classification from 3 to 4 groups. For the 1994 survey we refer to urban as the "city" area and to suburban as the sum of "town" and "suburbs". In the 1991 survey we used "inside central city" for the urban area and "outside central city" for the suburban area dummy variable. The regional dummy variables also show some changes between the surveys. In 1994 there is an increase of more than 3% of households living in the East-North Central, South and West-South Atlantic regions. A corresponding decrease is observed in the New England and West-North Central regions. The lifecycle dummy variables (defined in RTECS by 9 categories that combine age, number of children and household size) are similar in both survey years with approximately 40% of households with the oldest child aged below 17, a similar fraction of households composed of 2 adults, and the remaining 20% of singles.

`Table (1) here`

3.1 Empirical Specification

We consider a basic reduced form model for household gasoline demand which is given as follows,

$$\log gals_i = m(\log price_i,\ \log age_i,\ \log income_i,\ X_i) \ +\ u_i \tag{7}$$

where $gals_i$ is gasoline consumption of household i measured in gallons, $price_i$ is the average cost per gallon, age_i is the age of the household i head, $income_i$ is the annual income of a household and X_i is a vector of household characteristics: number of drivers

in the household ($\log drivers$), household size ($\log hldsize$), and dummy variables for residence (urban, suburban and rural) and for the lifecycle categories. The error u_i satisfies $E[u_i|\log price_i, \log age_i, \log income_i, X_i]$=0. In the above model, there are more than 20 predictors in which *price*, *income* and *age* are continuous and the remainder are discrete. Because of such a large number of predictors, we decide to use a semiparametric specification for $m(\cdot)$ where the three continuous variables are modeled nonparametrically and the discrete variables are entered linearly. This will greatly reduce the dimensionality of the problem while allowing flexible modeling of the price-income-age structure of demand. To this end, we consider three semiparametric models.

The first model is partially linear model,

$$\log gals_i = m_{P,A,I}(\log price_i,\ \log age_i,\ \log income_i) + X_i'\beta + u_i \tag{8}$$

where $m_{P,A,I}(\cdot)$ is an unknown smooth function. The second model is a refinement of (8) where the nonparametric function $m_{P,A,I}(\cdot)$ is additive while allowing linear bivariate interactions among $\log price$, $\log age$ and $\log income$, i.e.,

$$\log gals_i = m_P(\log price_i) + m_A(\log age_i) + m_I(\log income_i) + X_i'\beta + I_i'\delta + u_i \tag{9}$$

where $m_P(\cdot)$, $m_A(\cdot)$ and $m_I(\cdot)$ are unknown univariate smooth functions, and I_i is a vector containing ($\log price \times \log age$), ($\log price \times \log income$), and ($\log age \times \log income$). This model circumvents the curse of dimensionality problem in model (8). To see if the PLAM specification in (9) is supported by the data, we conduct a test of the null model (9) against (8). Using the specification test of Aït Sahalia *et al.* (2001), we can not reject model (9) at the 5% level.

Based on model (9), we also conducted both individual and joint-tests of the interaction coefficient δ. Both tests strongly indicate that none of the interactions are significant at the 10% level. Thus, we further reduce model (9) where interactions are eliminated from the specification,

$$\log gals_i = m_P(\log price_i) + m_A(\log age_i) + m_I(\log income_i) + X_i'\beta + u_i. \tag{10}$$

Unlike in model (9), the three nonparametric estimates in (10) represent partial effects of $\log price_i$, $\log age_i$ and $\log income_i$, respectively. Thus, we can interpret these estimates as nonparametric elasticities.

Based on model (10), we also consider the possible bias in the estimate of price elasticity due to the possible endogeneity of *price*. Yatchew and No (2001) suggest that fuel price and gasoline consumption might be negatively correlated. Households that drive more are likely to come across a wider range of prices and have lower average cost per gallon. In this case the nonparametric estimator is not consistent (i.e. overestimates the true responsiveness of demand to price) due to the correlation between the error term in Equation (10) and the $\log price$ variable.

We follow the approach of Blundell *et al.* (1998) to account for the possible endogeneity of the price variable. Assume there is a set of instrumental variables S_i such that

$$\log price_i = S_i'\pi + v_i \tag{11}$$

with $E(v_i|S_i) = 0$. We can then include the residuals v_i in Equation (10), that is,

$$\log gals_i = m_P(\log price_i) + m_A(\log age_i) + m_I(\log income_i) + \rho v_i + X_i'\beta + u_i \tag{12}$$

where we assume that $E(u_i|\log price_i,\ \log age,\ \log income,\ X_i,\ v_i) = 0$. Under these assumption, the resulting estimator of $m_P(\cdot)$ is consistent. The null hypothesis of exogeneity of the price variable can be tested using the least squares estimator of ρ. Equation (12) is estimated by including in the PLAM specification the fitted residuals $\hat{v}_i$ from the first-stage regression in Equation (11). Doing so will also not affect the asymptotic distribution of $\hat{\beta}$; see for example Newey *et al.* (1999).

To capture the above form of endogeneity, we may use the average intra-city price (i.e. an average over a neighborhood where the household resides) as the instrument for the household level price. But, these data is not available. Constrained by this data problem, we consider regional dummy variables instead. Using regional dummy variables as instruments, we can not reject the null hypothesis that price is exogenous; see Table (4). However, as one referee correctly points out regional dummy variables might not be valid instruments because they may correlate with gasoline consumption due to differences in land use patterns, in the density of development and in state size. So, our test may not fully address the endogeneity problem. On the other hand, looking at our nonparametric density weighted price elasticity estimates (see the following sections), they do not appear to be much larger than the values found in the literature. Note that the effect of endogeneity is to overestimate the elasticities. In view of this we think that the possible endogeneity in prices may not have caused serious bias in our estimates.

3.2 Results and Discussion

We begin by discussing the method RTECS uses to calculate the price variable and the undesirable consequence of this procedure on price-elasticity estimates when PLAM is implemented. This problem emerged from the analysis of the RTECS data in Schmalensee and Stoker (1999). To tackle this problem, we use the vehicle information in the RTECS to assign a more appropriate price measure to each household. We also obtain some interesting empirical results by estimating separate PLAMs for different categories (categorized by gasoline type use) of households.

3.2.1 Implausible price effect

The use of semi-parametric methods in Hausman and Newey (1995) and Schmalensee and Stoker (1999) suggested a puzzling property of the price effect on gasoline consumption. The non-parametric estimated price function (that relates price with gasoline demand) is upward sloping for a range of fuel prices in the middle of the distribution and is negatively sloped in the rest of the interval of variation. Schmalensee and Stoker (1999) investigated this implausible effect and attributed this finding to the price measure constructed by RTECS. They computed the price effect from the nonparametric estimate of the function $m_{P,I}(.,.)$ by slicing the curve along the income dimension. The $m_{P,I}(\cdot,\cdot)$ was estimated in the framework of the partial linear model using the approach of Robinson (1988).

RTECS does not collect fuel purchase diaries[6]. Instead, the total fuel expenditure is calculated based on the miles traveled (reported by the household for each vehicle owned) and a price is assigned based on the region of residence and grade of gasoline purchased. The price data are provided by the Bureau of Labor Statistics (BLS) at an aggregate level for each of 4 census regions (North-East, Mid-West, South, and West[7]) and for different grades (regular, midgrade, and premium). The problem with this procedure is that all households in a broad area as a Census region are assumed to face the same

[6]The EIA stopped collecting purchase diaries starting from the 1988 RTECS while earlier surveys contained also this information. Hausman and Newey (1995) considered the 1979, 1980 and 1981 surveys and they found the upward sloping demand although the price measure is based on diary of fuel purchases. Schmalensee and Stoker (1999) considered the 1988 and 1991 surveys where in both years the price measure was constructed by RTECS.

[7]The Census regions can be further partitioned in Census Divisions:

- North-East: New England and Middle Atlantic
- Mid-West: East-North Central and West-North Central
- South: South Atlantic, East-South Atlantic, and West-South Atlantic
- West: Mountain and Pacific.

gasoline price. However, this assumption is not realistic due to differences in state gasoline tax and intra-regional differences in prices. Schmalensee and Stoker (1999) considered the RTECS average cost per gallon as a measure of price (defined as total household expenditure divided by total gallons purchased). Figure (1) shows the scatter plot of the log average cost versus fuel consumed, and the smoothed distribution of the log fuel price. We consider all the households surveyed in 1991 and 1994 (a total of 5260 households). Further, we also report plots for the groups of households consuming only one grade (regular, midgrade, or premium) of gasoline for all the vehicles owned[8].

Figure (1) here

Consistent with the observation of Schmalensee and Stoker (1999), the scatter plots show that the gasoline price clusters around few values corresponding to the regional prices assigned by RTECS. The procedure creates an artificial discreteness in the price variable because it destroys the intra-regional variation in prices. This effect largely explains the bi-modal shape of the (smoothed) price densities for both the aggregate households and when they are segmented by grade of fuel purchased.

We estimate the PLAM specification in Equation (12) using the average cost (the price variable) calculated by RTECS. Figure (2) shows $\hat{m}_P(\log price)$ with bootstrap confidence intervals. It is clear from the non-parametric price curve that the same problem pointed out by Hausman and Newey (1995) and Schmalensee and Stoker (1999) also arises in the pooled sample of 1991 and 1994[9]. The demand for gasoline is upward sloping in the price range between \$1.1 and \$1.2. This price region is associated with a transition from households consuming mostly "regular" gasoline toward mostly "non-regular" (those households purchasing only midgrade or premium, or different fuel grades for the vehicles in the household). The discreteness of the price measure implies that for fuel prices between \$1.1 and \$1.2 there is an abrupt increase of the fraction of households purchasing non-regular fuel. These households are characterized by consuming (on average) more gasoline compared to regular ones. The upward sloping price curve can thus be interpreted as the result of the sudden concentration (artificially created by the price discreteness) of high consuming non-regular households that have a determinant role (at least locally) in determining the shape of the nonparametric estimator.

Figure (2) here

[8]The sample includes also 1398 households that have more than one vehicle and purchase different gasoline grades.

[9]The 1994 data has not been investigated by Schmalensee and Stoker (1999).

3.2.2 The vehicle based price measure

As the above result suggests, the lack of diaries of fuel purchases complicates the analysis of the relation between fuel price and quantity consumed. However, as we mentioned previously, RTECS also collects information on the last fuel purchase of households. Such information includes fuel price, fuel type, and grade for each vehicle in the household. These details are useful sources of information about the gasoline price faced by households that is neglected in the procedure described above[10].

A possible drawback of the vehicle information data is the presence of missing values. Some households did not provide information for any of their vehicles while others reported information for some or all the cars owned. Table (2) shows the number of households for which we have partial or complete vehicle information (in the Table indicated as *valid*) and those who did not provide any information[11]. Pooling the surveys of 1991 and 1994 we have a total of 5260 households. For 3020 of these households we have (partial or full) vehicles information. The Table reports some summary statistics of the main variables for the subset of households that reported prices and the full sample. The subsample represents closely the characteristics of the complete sample. The averages of the variables of interest (gallons consumed, household income, number of drivers) are very similar. Also, the distribution of the type of gasoline consumed in the subsample reflects quite well the complete sample. The only difference consists of the share of households having only one car. Their fraction decreases from 28% to 21% in the subsample. This effect is due to our choice of considering valid the households that have price information for at least one vehicle. It implies that our sub-sample slightly over-represents the households having more than one car and under-represents those that have only one vehicle. Overall, the descriptive statistics indicate that the selection of the sub-sample of households in the rest of our analysis should not significantly bias our results.

`Table (2) here`

Table (3) shows the average real prices[12] of the different gasoline grades for each of the 9

[10]RTECS collects this information during a phone interview with the household between January and March of the year following the survey. A concern with using the last fuel price is that it might not be representative of the average price faced by households during the survey year. For 1994 the EIA Petroleum Marketing Annual reports an average price (for all grades) of around 73.6, while it ranged between 70.5 and 71.3 during January and March 1995 (when the interview takes place). The difference is not very large. Hence, we believe the last fuel price represents a good proxy for the average price paid by households during the survey year.

[11]We decided to consider as *missing* the households that did not report information for any of the vehicles owned. Instead, we consider as *valid* those units that reported information for at least one vehicle.

[12]We deflated prices in 1994 to 1991 levels using the CPI Index.

Census divisions based on the vehicle-based price data from 1991 and 1994. In this case the unit of analysis is the vehicle: we pooled all the vehicles in the surveys and segmented them by division and by gasoline grade. We also report the standard deviation of the price and the number of vehicles in the category. The first aspect that emerge is the significant *inter*-divisional (and of course inter-regional) variation in fuel prices. In 1991, a group of divisions had an average price for regular gasoline around \$1 and the other group (New England, Mid-Atlantic, and Pacific) above \$1.1. The difference is probably due to higher gasoline taxes in some states. Another fact that emerge from the Table is the significant *intra*-divisional variation. The standard deviations vary between 0.077\$ (regular in New England) and 0.177\$ (premium in the Pacific division). It is thus clear that the vehicle information delivers a price measure that accounts for the *intra-regional* dispersion in prices that is neglected when assigning a common regional price to all households as in the RTECS methodology.

Table (3) here

We assign an average cost to each household which was defined as total expenditure (calculated using the last fuel price) divided by the total gallons consumed. For the households that reported prices for only part of their cars, we impute a value given by the average of the prices reported for vehicles in the same division and using the same grade. In this way, we use the last fuel price to assign the missing observations an average price that is more detailed compared to the RTECS procedure (at the division level instead of regional). The average cost, $price_i$, for household i is given by

$$price_i = \frac{\text{Total Expenditure of hld } i}{\text{Total Gallons hld } i} = \frac{\sum_{k=1}^{K} price_{i,k} gals_{i,k}}{\sum_{k=1}^{K} gals_{i,k}}$$

where $price_{i,k}$ denotes the last fuel price reported by household i for vehicle k, $gals_{i,k}$ the gallons consumed by the same vehicle and K is the total number of cars owned by household i. Figure (3) is similar to Figure (1) with the difference that the vehicle information is used to calculate the average fuel price. The scatter plots of the log gallons consumed and the log price does not show the clusters of observations that characterizes Figure (1). In addition, the range of price variation is much wider compared to the RTECS measure. This is due to the effect of accounting for the *intra-divisional* dispersion of prices[13]. The bi-modality that was apparent for the RTECS price measure has now

[13]Figure (3) shows that there are some extreme prices in the right tail of the price distribution. We checked the price data for these households; they are mainly consuming midgrade and premium gasoline and living in the Pacific division. They reported a price for the last fuel purchase between 1.70\$ and 2\$.

disappeared. In this sense, the vehicle based price measure is a realistic indicator of the fuel cost faced by households and should not be affected by the problems discussed in the previous Section.

Figure (3) here

3.2.3 Corrected price effect

We now consider the model in Equation (12)[14] where the price variable is represented by the average cost based on the vehicle information. For comparison purposes, we also report the estimation results of Equation (10) for the 1991, 1994 and the pooled households data (where we exclude the price effect as in Schmalensee and Stoker (1999)). For the latter case, we adopt the specification with $\log age$ and $\log income$ treated additively (but not price) and, as a proxy for the price effect, we also include regional dummy variables in the linear part of the PLAM specification. Figure (4) shows the estimated components (with bootstrap-based confidence intervals) for $\log price$, $\log age$ and $\log income$ along with the estimated price elasticity[15]. Table (4) reports the density-weighted average derivatives for the additive components and the estimated coefficients for the PLAM model. The comparison of the PLAM estimation based on the 3020 households (using the new price variable) and the pooled 1991 and 1994 surveys (5260 observations) with regional dummy variables does not show significant differences in the results. Thus, the selection of the subsample of households that reported fuel prices for their vehicle does not bias significantly the estimates of the other components. The estimation on the full sample available for 1991 and 1994 shows that there is some variation in the magnitude of the coefficients for some variables but the results are quite close to the estimates for the pooled case.

Table (4) here

These results confirm that the use of the vehicle-based data does not substantially alter the conclusion from the household-level data while permitting the estimation of the price elasticity. We summarize the results of the PLAM estimation for vehicle-based data

[14]We selected the bandwidth based on the CV search described in Section (2) for different values of the constant a in $b_j = a\ \sigma_j\ n^{-2/7}$ (for j=1,2, and 3). The optimal values used in the application are 0.11 for $\log price$, 0.34 for $\log age$ and 0.73 for $\log income$. In estimation we trim the 5% of observations in the low density region of the explanatory variables.

[15]The elasticity curve is derived from the one-step back-fitting procedure (that implements a local linear smoothing) discussed in Section (2) of the paper. The standard error for the estimated price elasticity is obtained by bootstrap.

as follows. The first interesting result of the analysis is that the estimated $\log price$ component is negatively sloped in the complete range of variation of the variable. Panel (c) of Figure (4) shows the nonparametric estimate of the price elasticity. For low prices it is close to -0.2 and increases toward -0.5 for high prices suggesting that gasoline demand becomes more responsive to price changes when the fuel price is high. The density-weighted average derivative is equal to -0.35. A possible interpretation of this finding is the heterogeneity in the grade purchasing decision of households. At low prices, most households consume regular gasoline while high prices are typical of those households that purchase midgrade or premium gasoline. In the next section we segment the sample in groups based on the gasoline grade purchased. We distinguish between households that bought regular gasoline for all their vehicles (the "regular" households) and those that bought (for at least one of their vehicles) midgrade and/or premium (the "non-regular" households).

The estimated $\log age$ component shows a similar pattern to that previously found by Schmalensee and Stoker (1999). It is flat for households aged below 50 and slopes down significantly for higher ages. The $\log income$ variable has a density-weighted average derivative of 0.16 and the component does not appear to deviate significantly from linearity.

Figure (4) here

Table (4) also reports the estimated coefficients for the variables that enter the PLAM specification in a linear fashion. The $\log drivers$ variable is highly significant with an estimated elasticity of 0.69. Households living in urban area consume (on average) less compared to those living in suburbs, while the opposite is true for those residing in rural areas. The lifecycle variable reveals that households with the oldest child aged between 7 and 15 and singles aged below 35 consume (on average) significantly more. However, households composed of 1 or more adults aged above 60 tend to consume significantly less. Accounting for endogeneity of the price variable shows that the null hypothesis of $\rho = 0$ cannot be rejected at standard significance levels.

3.2.4 Heterogeneity of households

As we discussed above, the estimated price component reveals an interesting feature of a larger elasticity (in absolute value) for higher prices compared to low prices. To investigate further this issue we segment the 3020 households in two groups[16]: those consuming (for

[16]Yatchew and No (2001) conduct a similar analysis where they segment households based on the decision to purchase regular, medium or premium gasoline. We decided to divide our sample in "regular"

all their vehicles) regular gasoline (1682 households) and those that consume non-regular (1338 households). The second group includes households that purchase only midgrade or premium gasoline and those that buy different grades (regular/midgrade/premium) for their vehicles.

We estimate the PLAM specification in Equation (12) separately for "regular" and "non-regular" households. Table (5) reports the estimation results for the two groups. Some interesting results emerge from the comparison. First, the estimated (density-weighted) average price derivative for regular users is equal to -0.54 and for non-regular to -0.33. Although the price elasticities have large standard errors, households that buy exclusively regular gasoline seem to be more sensitive to price changes compared to households that purchase non-regular grades. This result seems to be at odd with our earlier finding (based on all households) that the estimated price elasticity increases at higher prices. An intuitive interpretation of the aggregate result is that we should expect low elasticity for regular households (since regular gasoline buyers are likely to concentrate at the lower end of the price distribution) and high elasticity for non-regular households (that characterize the upper end of the price range of variation). However, our results from the separate regression for regular and non-regular suggest the opposite interpretation. The key to understand this is the fact that non-regular households consume (on average) more gallons of gasoline compared to the other group. They are thus characterized for being less price sensitive and for consuming more gallons of gasoline. Panel (a) of Figure (5) shows the estimated price component for the two groups together with the aggregate one. The estimated price component for all households lies between the component for non-regular households (top) and regular (bottom) since it can be interpreted as a weighted average of the two curves. At low gas prices[17] a large fraction of households consumes regular gasoline and the aggregate component is close to the regular one. Increasing the price, the aggregate curve shifts toward the non-regular component due to the higher weight

and "non-regular" in order to have a large number of observations in each group. The households that reported prices for their vehicles is composed of 3020 observations of which 1682 consumed regular for all their vehicles, 319 purchased exclusively midgrade, 190 only premium, and the remaining 829 bought different grades.

[17]The price distribution for the regular and non-regular groups are shown in Panel (c) of Figure (5). It is interesting to notice that the two (smoothed) densities overlap in a range of gasoline prices between \$0.90 and \$1.22. This is due to two reasons. The first is because of geographical dispersion in prices. From Table (3), we can notice that there are division (e.g., W/N Central) where premium gasoline is cheaper than regular in other divisions (e.g., New England and Pacific). The second reason for this wide overlap of the two price distributions has to do with the way we constructed the regular and non-regular groups. While the former is composed of households only buying regular gasoline for all of their vehicles, the latter is characterized by those households that buy, for at least one of their vehicles, non-regular gasoline.

of the non-regular households. Panel (b) shows the estimated elasticities for the regular and non-regular group together with the aggregate one discussed in the previous Section. The price elasticity of regular households is close to -0.20 at low prices and increases toward -0.70 at high prices. However, the gasoline demand of non-regular households is quite inelastic at low prices and increases to -0.50 for high prices. For gas prices above $1.05, the aggregate price elasticity has a magnitude similar to the non-regular estimated elasticity suggesting that the respective components are parallel (in that price range).

Table (5)

Figure (5)

The regressions results for regular and non-regular households also reveal some other interesting differences between the groups. The role of the $\log age$ is remarkably different for regular and non-regular users. For regular households it has a negative elasticity (equal to -0.34). However, for non-regular users there hardly exist an age effect. Panel (d) of Figure (5) gives a graphical intuition for this result. The additive $\log age$ component for regular users has a very similar pattern to the pooled case. It starts flat and then rapidly slopes downwards when the householder age increases. However, for non-regular users the estimated component is approximately flat in the range of variation of the $\log age$ variable. This result suggests that the demand for non-regular gasoline is not influenced by age.

The groups are also heterogeneous in their elasticities to income and the number of drivers in the household. Non-regular households have a significantly larger income elasticity compared to regular (0.20 and 0.13, respectively) while the opposite effect holds for the drivers effect (0.54 and 0.78, respectively). Households that consume non-regular gasoline are more responsive to changes in income compared to regular gasoline, and less sensitive to changes in the number of drivers.

4 Conclusion

In this paper we apply the Partially Linear Additive Model (PLAM) to model gasoline demand in the United States. The flexibility of the semiparametric specification derives from the possibility of including variables both in a parametric and nonparametric fashion. In addition, for each variable treated non-parametrically we estimate a component that allows an easy graphical interpretation of the relationship with the dependent variable.

We estimate the model following the approach of Manzan and Zerom (2005). Compared to alternative estimators, the adopted estimator is semi-parametrically efficient, has better finite sample properties and it is computationally more convenient.

On the empirical side, we reexamine the issue of the price elasticity of gasoline demand in the United States discussed by Schmalensee and Stoker (1999). Using the RTECS data, we construct an average fuel cost for each household based on "vehicles" information contained in the survey. This allows us to overcome the difficulties encountered by Schmalensee and Stoker (1999), who use the average cost provided by RTECS. In particular, we show that there is significant dispersion in gasoline prices across the US and across grades of fuel. By estimating the PLAM specification with $\log price$, $\log income$ and $\log age$ treated non-parametrically (but additively), we find a density weighted price elasticity of around -0.35. The non-parametric estimate of the price elasticity also shows the tendency to increase (in absolute value) at higher prices. This suggests that households might respond differently to price changes depending on the level of price.

We further investigate the above empirical result by splitting the households in the sample in two groups depending on the grade of gasoline purchased. The estimation results for the two groups show that regular users are more sensitive to price changes (estimated elasticity of -0.54) compared to non-regular users (that have an elasticity of -0.33). The price elasticities for regular and "non-regular" households have a similar pattern: they are quite inelastic at low prices and become increasingly responsive for high prices. Separate analysis of the two groups also shows significant differences in the effects of income, age and number of driver.

Finally, it is worth noting that while our estimated density-weighted average price elasticity of -0.35 is well within the range found in the literature, the dependence of the price elasticity on the level of price (and fuel grade) is a new empirical finding. In light of this result, further empirical investigation with more recent data is warranted[18].

[18]As we mentioned earlier, the last RTECS was run by the EIA in 1994. For 2001, EIA provides an equivalent of the RTECS based on information collected by the National Household Travel Survey (NHTS) of the U.S. Department of Transportation. However, there are relevant differences between the original RTECS and the 2001 RTECS that significantly limit its use. First, the NHTS does not collect fuel purchase diaries and household expenditure is constructed in RTECS based on retail gasoline prices in the state of residence of the household. This procedure is affected by the same problems discussed in Section (3.2.1). In addition, no information is provided on the gasoline grade purchased for the vehicles in the households. This prevents a detailed analysis of heterogeneity in gasoline demand between households buying regular and non-regular fuel. For these reasons we could not include more recent data in our analysis.

Appendix: Data Description

The data consists of the 1991 and 1994 RTECS that are publicly available at

`http://www.eia.doe.gov/emeu/rtecs/`

The EIA stopped the RTECS in 1994 and hence prevented us from studying more recent periods. The survey reports files that include information on characteristics of the households and of the owned vehicles. The data used in the paper are extracted from the following survey files:

- `househld`: contains information about *households characteristics*, such as: total gallons purchased, income, number of drivers, members of the household, age of the householder, location variables (area, census division and region), lifecycle variable (composition and age of the household members), total miles driven, and fuel expenditure.

- `veconexp`: contains information about each (up to a maximum of 8) *vehicle* owned by the household. The vehicle characteristics reported are: total gallons consumed, total fuel cost, and average cost (per vehicle). The average cost is determined by the EIA procedure to assign average prices in the census region where the household lives and based on the type of gasoline purchased. This file is related to information that the EIA obtained by the household or assigned by the agency.

- `vehchar5` (`veh5` in 1994 survey): contains information about each vehicles last fuel purchase; the information concerns: price, type, and grade of the last fuel purchase and MPG (Miles per Gallon) estimate. Additional information contained in the file is the age of the usual driver, if the vehicle is used to commute to work, and the number of miles to commute. The information contained in this file is based on responses given by the household during a phone conversation as part of the survey.

- `fueltype`: information about each vehicle type and grade of fuel purchased. Fuel type is classified in 4 categories: gasoline, diesel, gasahol and propane. Vehicles are also classified by fuel grade that can be regular, premium, midgrade, and both regular and premium.

The `veconexp` data is based on the VMT (Vehicles-Miles Traveled) based on the households reports of odometer readings. From this information the EIA adopts a vehicle-specific MPG (Miles per Gallons) estimate[19] to calculate the amount of gallons consumed

[19]The estimate is provided by the Environmental Protection Agency (EPA) and is specific to the type of vehicles considered and the fuel type purchased.

by each vehicle in the household. The sum of the gallons consumed per vehicle provides the total gallons of fuel consumed by the household. All other information about the household is based on a phone interview conducted as part of the survey. Households characteristics are included in the `househld` file, while `vehchar5` contains information about the last fuel purchase (price and type). In this file some data are missing. Some households failed to report the price and/or type of fuel purchased for all their vehicles, whereas other households reported information for only part of the vehicles owned. It is interesting to notice that there are two sources of information on the gasoline type purchased: the file `veconexp` contains the type used by EIA to calculate the MPG, while `vehchar5` reports the information provided by the respondents. As mentioned above, for some vehicles this information is missing. However, when the gasoline type is reported in `vehchar5` it is also equal to the information reported in `veconexp`. This suggests that EIA used the vehicle information provided by the respondents to attribute a gasoline type to each vehicle. However, it is not clear from the documentation how they attributed the type of gasoline when this information was not provided by the respondents (the missing data mentioned earlier).

References

Aït Sahalia, Y., Bickel, P.J. and Stoker, T.M. (2001). Goodness-of-fit tests for regression using kernel methods. *Journal of Econometrics*, **105,** 363–412.

Blundell, R. and Duncan, A. (1998). Kernel regression in empirical microeconomics. *Journal of Human Resources*, **33,** 62–87.

Blundell, R., Duncan, A. and Pendakur, K. (1998). Semiparametric estimation and consumer demand. *Journal of Applied Econometrics*, **13,** 435–461.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, **60,** 567–596.

Congressional Budget Office (2002). Reducing gasoline consumption: three policy options. *CBO study.*

Congressional Budget Office (2003). The economic costs of fuel economy standards versus a gasoline tax. *CBO study.*

Coppejans, M. (2003). Flexible but parsimonious demand designs: The case of gasoline. *Review of Economics and Statistics*, **85,** 680–692.

Dahl, C. and Sterner, T. (1991). Analysing gasoline demand elasticities: a survey. *Energy Economics*, **13,** 203–210.

Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low dimensional components in additive models. *Annals of Statistics*, **26,** 943–971.

Fan, Y. and Li, Q. (2003). A kernel-based method for estimating additive partially linear models. *Statistica Sinica*, **13,** 739–762.

Graham, D.J. and Glaister, S. (2002). The demand for automobile fuel: A survey of elasticities. *Journal of Transportation Economics and Policy*, **36,** 1–26.

Hausman, J.A. and Newey, W.K. (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica*, **63,** 1445–1476.

Hengartner, N.W. and Sperlich, S. (2005). Rate optimal estimation with the integration method in the presence of many covariates. *Journal of Multivariate Analysis*, **95,** 246–272.

Kim, W., Linton, O.B. and Hengartner, N.W. (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, **8,** 278–297.

Li, Q. (2000). Efficient estimation of additive partial linear models. *International Economic Review*, **41,** 1073–1091.

Linton, O.B. (1996). Efficient estimation of additive nonparametric regression models. *Biometrika*, **84,** 469–474.

Liu, R.Y. (1988). Bootstrap procedure under some non-*i.i.d.* models. *Annals of Statistics*, **16,** 1696–1708.

Manzan, S. and Zerom, D. (2005). Kernel estimation of a partially linear additive model. *Statistics & Probability Letters*, **72,** 313–322.

Moral, I. and Rodriguez-Poo, J.M. (2004). An efficient marginal integration estimator of a semiparametric additive modeling. *Statistics and Probability Letters*, **69,** 451–463.

Newey, W., Powell, J. and Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, **67,** number 3, 565–603.

Nicol, C.J. (2003). Elasticities of demand for gasoline in Canada and the United States. *Energy Economics*, **25,** 201–214.

Parry, I.W.H. and Small, K.A. (2005). Does Britain or the United States have the right gasoline tax? *American Economic Review*, **95,** 1276–1289.

Robinson, P. (1988). Root-n consistent semiparametric regression. *Econometrica*, **56,** 931–954.

Schmalensee, R. and Stoker, T.M. (1999). Household gasoline demand in the United States. *Econometrica*, **67,** 645–662.

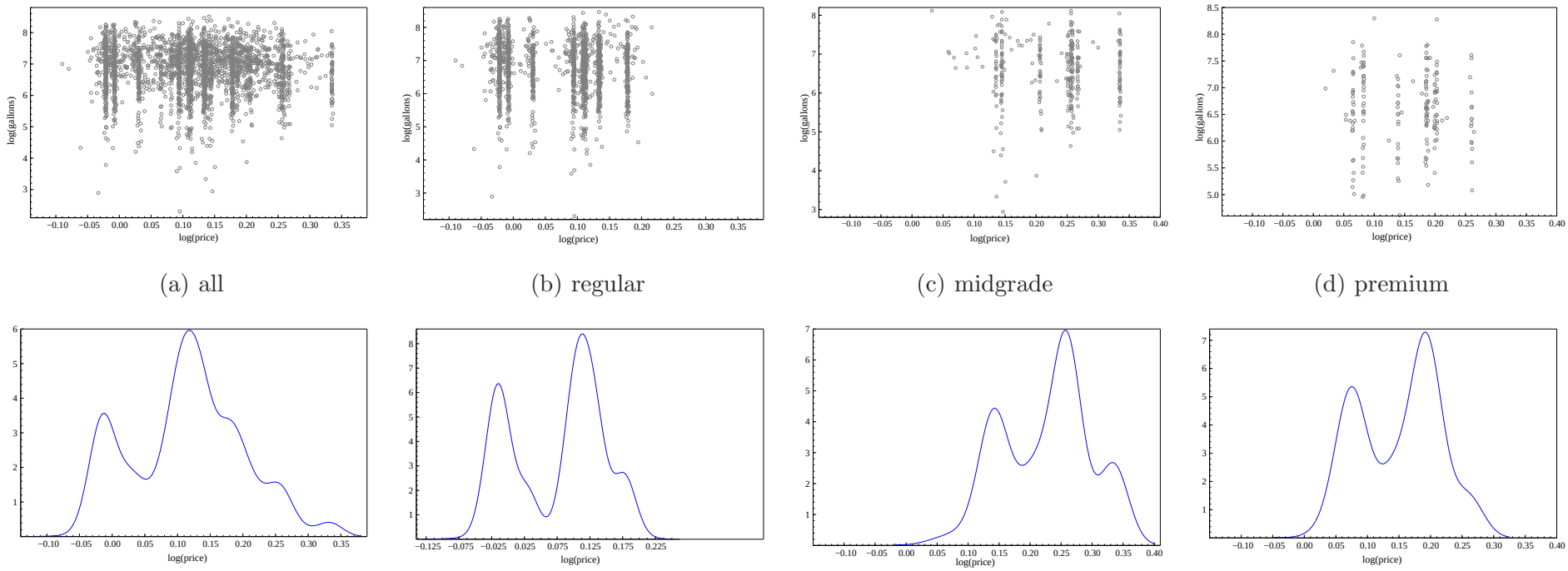
US Department of Energy (1996). Policies and measures for reducing energy related greenhouse gas emissions: lessons from recent literature. *Report No. DOE/PO-0047.*

US Department of Energy (2004). Emissions of greenhouse gases in the United States 2003. *Report No. DOE/EIA-0573.*

Yatchew, A. and No, J.A. (2001). Household gasoline demand in Canada. *Econometrica*, **69,** 1697–1709.

Variables	1991		1994	
	Mean	St. Dev.	Mean	St. Dev.
log(*gallons*)	6.75	0.718	6.76	0.743
log(*income*)	3.31	0.794	3.15	0.736
log(*drivers*)	0.548	0.4	0.54	0.402
log(*hld size*)	0.882	0.536	0.863	0.528
log(*age*)	3.76	0.363	3.8	0.363
Residence Dummy Variables (in % of total):				
Urban	0.284		0.424	
Suburban	0.445		0.384	
Rural	0.271		0.192	
Region Dummy Variables (in % of total):				
New England	0.075		0.049	
Middle Atlantic	0.128		0.127	
East North Central	0.141		0.172	
West North Central	0.143		0.088	
South Atlantic	0.117		0.183	
East South Atlantic	0.082		0.066	
West South Atlantic	0.08		0.114	
Mountain	0.084		0.062	
Pacific	0.148		0.136	
Lifecycle Dummy Variables (in % of total):				
Oldest Child < 7 years	0.127		0.112	
Oldest Child 7-15 years	0.214		0.198	
Oldest Child 16-17 years	0.072		0.076	
Two Adults, Head < 35 years	0.084		0.084	
Two Adults, Head 35-59 years	0.16		0.182	
Two Adults, Head ≥ 60 years	0.16		0.165	
One Adult, Head < 35 years	0.045		0.036	
One Adult, Head 35-59 years	0.065		0.068	
One Adult, Head ≥ 60 years	0.071		0.078	

Table 1: Descriptive statistics for the RTECS data of 1991 and 1994.



(a) all (b) regular (c) midgrade (d) premium

Figure 1: RTECS price measure defined as log average cost for the households in the 1991 and 1994 surveys and for those using only one grade of gasoline (the remaining 1398 households purchased different grades for their vehicles). (`top`) Scatter plot of gallons of gasoline consumed by an household and the average price, (`bottom`) smoothed density of the log(*price*) attributed by RTECS to household i. The gasoline price for 1994 is deflated to 1991 levels by the CPI index.

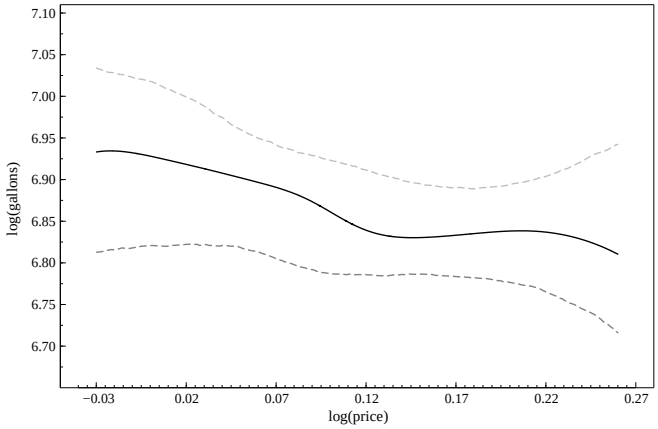


Figure 2: The estimated price component $m_p[\log(price_i)]$ for the PLAM specification in Equation (12) when the RTECS price measure is considered. The estimate is based on the pooled 1991 and 1994 surveys (5260 households). 95% confidence intervals obtained by bootstrap.

	1991		1994		Pooled	
	Valid	All	Valid	All	Valid	All
log(gallons)	6.86	6.75	6.86	6.76	6.85	6.75
	(0.68)	(0.72)	(0.76)	(0.74)	(0.72)	(0.73)
log(age)	3.78	3.76	3.82	3.80	3.80	3.78
	(0.34)	(0.36)	(0.34)	(0.36)	(0.34)	(0.36)
log(income)	3.45	3.31	3.17	3.06	3.31	3.19
	(0.73)	(0.79)	(0.68)	(0.73)	(0.72)	(0.77)
log(drivers)	0.61	0.55	0.58	0.54	0.60	0.54
	(0.39)	(0.40)	(0.39)	(0.40)	(0.39)	(0.40)
log(hld size)	0.91	0.88	0.88	0.86	0.90	0.87
	(0.52)	(0.53)	(0.52)	(0.53)	(0.52)	(0.53)
log*(gallons) by gasoline grade:*						
Regular	6.78	6.69	6.81	6.72	6.80	6.71
	(0.72)	(0.74)	(0.77)	(0.74)	(0.75)	(0.74)
Midgrade	6.62	6.48	6.54	6.50	6.58	6.49
	(0.71)	(0.69)	(0.89)	(0.80)	(0.80)	(0.75)
Premium	6.67	6.50	6.55	6.51	6.61	6.51
	(0.63)	(0.69)	(0.73)	(0.75)	(0.69)	(0.72)
More grades	7.12	7.07	7.15	7.16	7.13	7.11
	(0.49)	(0.50)	(0.55)	(0.52)	(0.52)	(0.51)
Gasoline Grade (in % of total):						
Regular	0.55	0.53	0.56	0.58	0.56	0.55
Midgrade	0.11	0.13	0.10	0.12	0.11	0.13
Premium	0.05	0.06	0.07	0.09	0.06	0.07
More Grades	0.28	0.28	0.26	0.21	0.27	0.24
Number of Vehicles (in % of total):						
One	0.21	0.28	0.22	0.31	0.22	0.30
Two	0.40	0.39	0.39	0.39	0.40	0.39
Three	0.24	0.20	0.23	0.18	0.23	0.20
More	0.15	0.12	0.14	0.10	0.15	0.11
Total	1571	2697	1449	2563	3020	5260

Table 2: Summary statistics for the full sample and the subsample of households that reported the price of the last fuel purchase. In parenthesis the standard deviations of the households characteristic variables. For gasoline grade and number of vehicles we reported percentages of households belonging to each category.

	1991			1994		
	Regular	Midgrade	Premium	Regular	Midgrade	Premium
New England	115.92	126.53	135.05	109.05	114.21	125.01
	(7.71),[118]	(7.86),[32]	(9.82),[40]	(9.58),[89]	(9.44),[18]	(10.11),[28]
Mid Atlantic	110.31	119.19	131.85	105.31	113.14	121.32
	(9.30),[254]	(13.02),[31]	(12.89),[88]	(8.95),[217]	(7.33),[48]	(10.23),[75]
E/N Central	101.97	110.55	117.38	96.72	102.16	109.43
	(9.32),[320]	(11.77),[33]	(17.62),[52]	(7.05),[346]	(9.75),[67]	(12.02),[76]
W/N Central	100.4	99.08	108.72	95.12	98.81	104.25
	(10.6),[344]	(9.61),[39]	(12.86),[61]	(9.28),[190]	(8.16),[26]	(7.14),[23]
South Atlantic	103.36	112	121.69	96.72	103.88	113.99
	(9.19),[182]	(8.11),[48]	(8.68),[65]	(8.95),[270]	(13.99),[82]	(8.93),[84]
E/S Atlantic	102.34	109.25	115.33	96.48	104.49	113.45
	(7.85),[151]	(8.69),[24]	(9.68),[46]	(6.75),[117]	(5.68),[23]	(113.45),[53]
W/S Atlantic	102.77	112.55	117.07	96.91	106.01	109.45
	(8.93),[137]	(12.75),[29]	(11.56),[59]	(7.06),[181]	(5.38),[39]	(8.95),[63]
Mountain	102.71	103.5	111.65	107.69	112.1	117.24
	(8.56),[198]	(10.95),[12]	(10.79),[26]	(8.40),[131]	(5.69),[13]	(9.65),[22]
Pacific	111.40	113.62	129.63	111.82	120.04	127.61
	(11.94),[258]	(14.23),[29]	(17.88),[95]	(7.48),[198]	(8.84),[36]	(9.55),[60]

Table 3: Average Real Prices in $ cents per gallon based on vehicles data. The number in $(\cdot)$ is the standard deviation of the price per division and per grade of gasoline and $[\cdot]$ the number of vehicles for each entry.

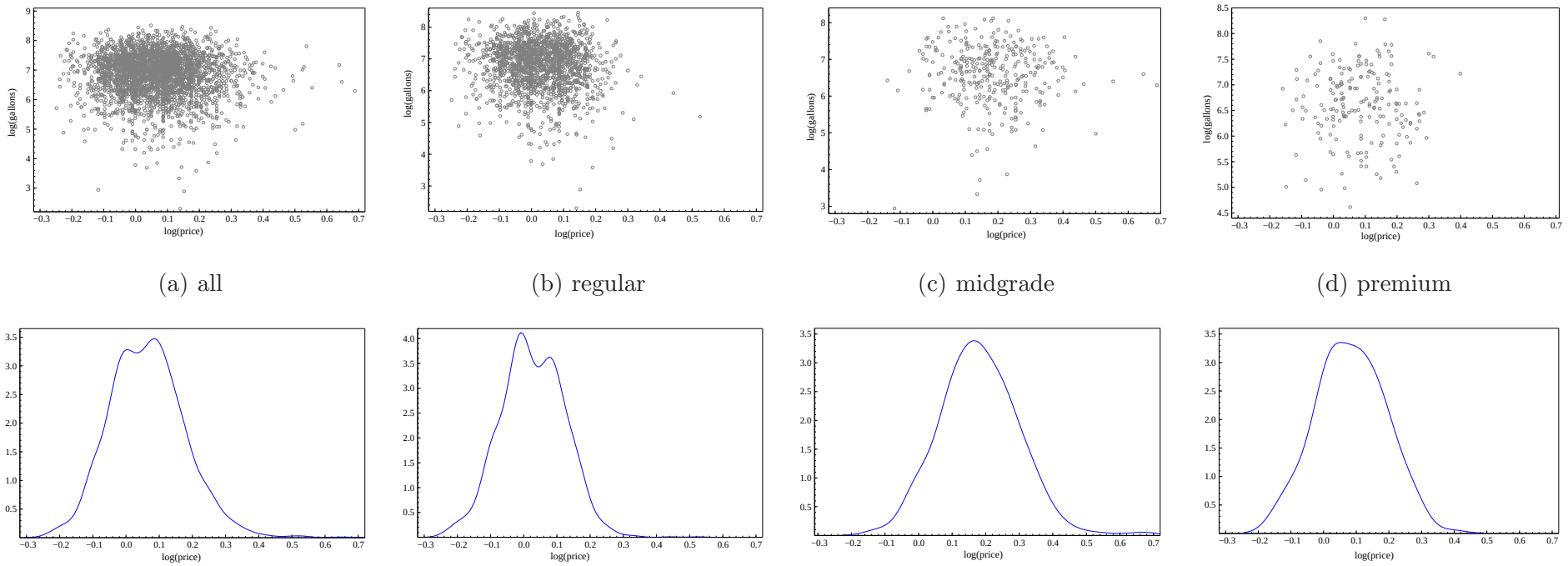
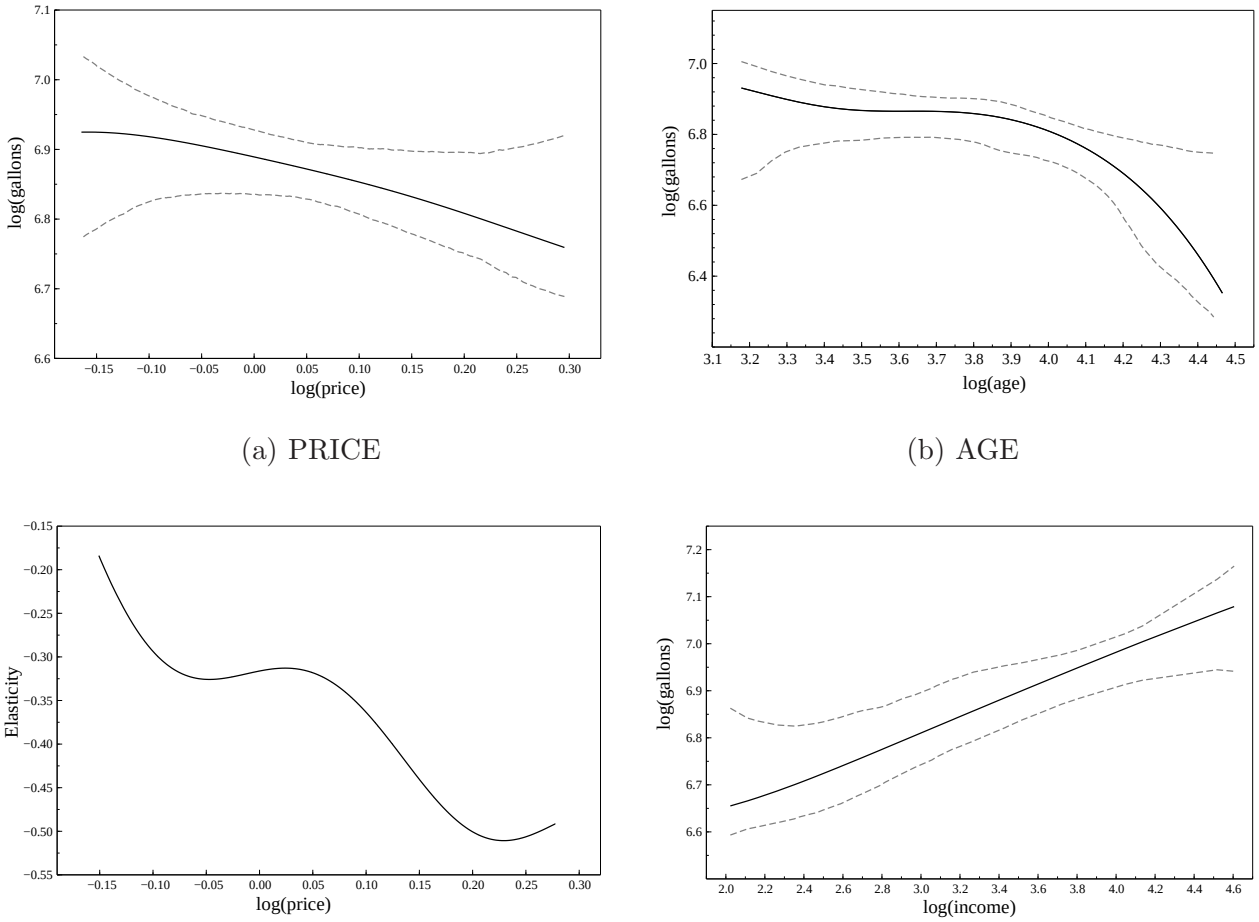


Figure 3: Price measure based on the vehicles price information for the households in the 1991 and 1994 surveys and for those using only one grade of gasoline (the remaining 829 households are those that purchase more than one grade for their vehicles). (`top`) Scatter plot of gallons of gasoline consumed by an household and the average price, (`bottom`) smoothed density of the $\log(price)$ attributed to household i. The gasoline price for 1994 is deflated to 1991 levels by the CPI index.

	1991		1994		Pooled 91&94		Valid	
	Av. Der.	Std. Err.	Av. Der.	Std. Err.	Av. Der.	Std. Err.	Av. Der.	Std. Err.
log(price)							-0.355	0.117
log(age)	-0.165	0.041	-0.139	0.053	-0.22	0.051	-0.11	0.043
log(income)	0.20	0.017	0.147	0.022	0.132	0.016	0.162	0.017
	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
log(drivers)	0.649**	0.044	0.667**	0.0458	0.651**	0.0319	0.692**	0.0451
log(hld size)	0.116*	0.056	0.0529	0.0586	0.0823*	0.0404	0.078	0.0552
First-Stage Residuals							0.099	0.143
Residence Dummy Variables:								
area - urban	-0.165**	0.027	-0.139**	0.0258	-0.135**	0.0185	-0.113**	0.024
area - rural	0.086**	0.0283	0.175**	0.0321	0.124**	0.0211	0.165**	0.0266
Lifecycle Dummy Variables:								
lifecycle - 7<child<15	0.0935*	0.0417	0.0293	0.0443	0.055	0.0305	0.0964*	0.0408
lifecycle - 16<child<17	0.0362	0.0568	0.0429	0.0576	0.0234	0.0408	0.0386	0.0526
lifecycle - 2+adlts<35	0.0368	0.0561	-0.0244	0.0605	0.0289	0.0412	0.052	0.0578
lifecycle - 35<2+adlts<59	0.0822	0.0537	0.0107	0.0574	0.0214	0.0396	0.095	0.052
lifecycle - 2+adlts>60	-0.015	0.0646	-0.126	0.0728	-0.176**	0.0484	-0.088	0.0624
lifecycle - 1adlts<35	0.301**	0.0864	0.0353	0.0954	0.169**	0.0641	0.194*	0.089
lifecycle - 35<1adlts<59	0.0829	0.0849	-0.13	0.0887	-0.0443	0.0616	0.020	0.0816
lifecycle - 1adlts<60	-0.208*	0.0915	-0.442**	0.099	-0.426**	0.0667	-0.364**	0.0895
Division Dummy Variables:								
div. - mid atl.	-0.039	0.0501	-0.0511	0.06	-0.0373	0.0384		
div. - E/N central	0.0748	0.0491	0.0898	0.0579	0.0891*	0.0373		
div. - W/N central	0.128**	0.0496	0.155*	0.0637	0.127**	0.0392		
div. - S central	0.113*	0.0508	0.082	0.0574	0.111**	0.0375		
div. - E/S central	0.098	0.0554	0.151*	0.0675	0.124**	0.043		
div. - W/S central	0.137*	0.0556	0.0957	0.0612	0.118**	0.0406		
div. - mountain	0.116*	0.0548	0.143*	0.0687	0.13**	0.0431		
div. - pacific	0.0246	0.0487	0.0643	0.0599	0.0448	0.0379		
R^2	0.385		0.409		0.392		0.395	
N	2697		2563		5260		3020	

Table 4: For the 1991, 1994 and the pooled samples we estimated the PLAM model with log-AGE and log-INCOME as additive components and log-DRIVERS, log-SIZE, residence, lifecycle and regional dummy variables in the linear part. For the subsample of households that reported price information, we estimate the PLAM specification in Equation (12) with log-PRICE as additive component but excluding the regional dummy variables (that are used as instruments in the first-stage regression to account for endogeneity of the price variable). Standard errors for the density-weighted average derivative obtained by bootstrap. Significance at 1% is denoted by ** and at 5% by *. N indicates the sample size.



(a) PRICE

(b) AGE

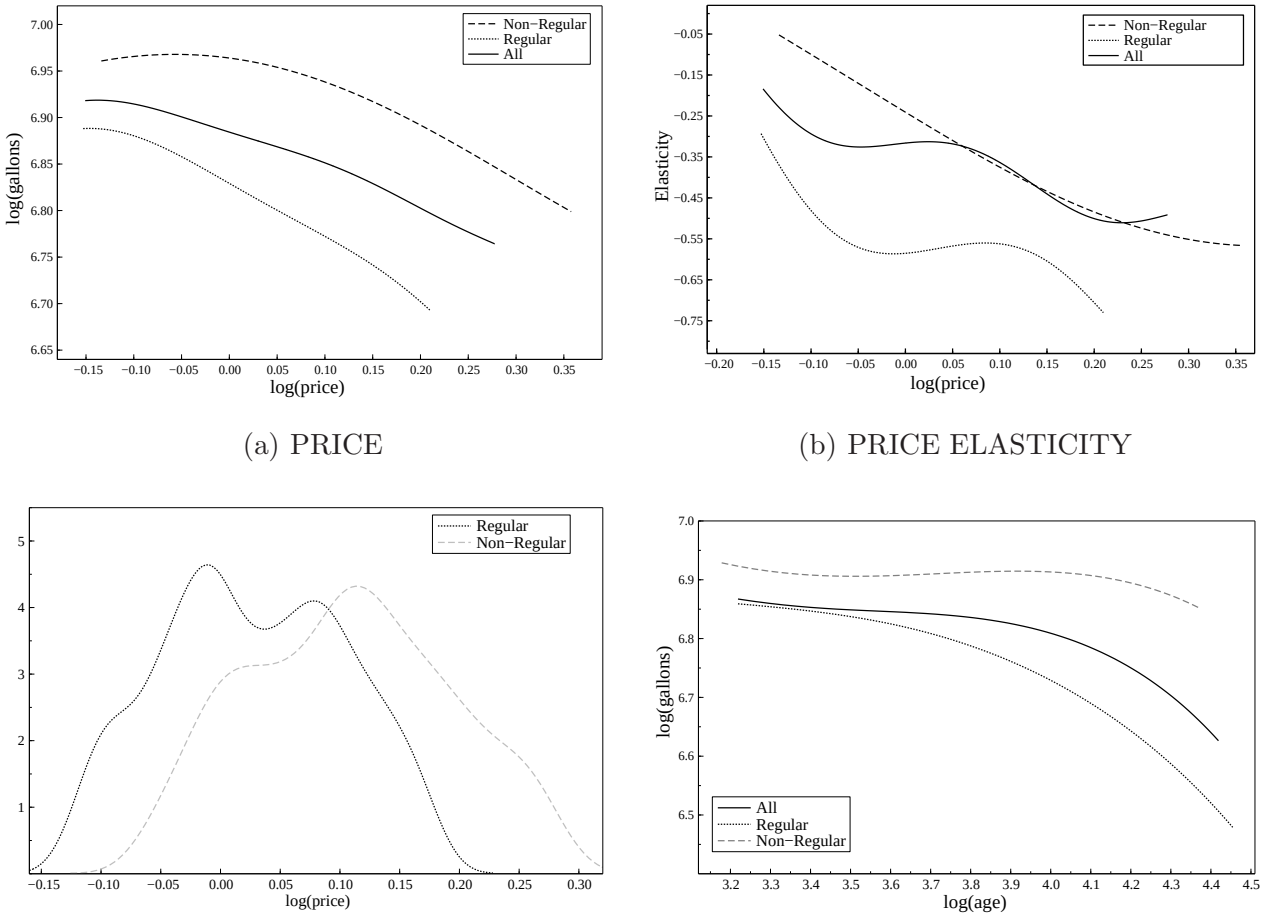
(c) PRICE ELASTICITY

(d) INCOME

Figure 4: Estimated nonparametric components for PRICE, AGE and INCOME of the PLAM specification in Equation (12) with 95% bootstrap confidence intervals. Panel (c) is the nonparametric estimate of the price elasticity.

	Regular		Non-Regular	
	Av. Der.	Std. Err.	Av. Der.	Std. Err.
log(price)	-0.545	0.209	-0.331	0.144
log(age)	-0.344	0.076	0.018	0.053
log(income)	0.135	0.024	0.20	0.023
	Coeff.	Std. Err.	Coeff.	Std. Err.
log(drivers)	0.783*	0.0635	0.546*	0.0621
log(hld size)	0.066	0.0805	0.094	0.0736
Residence Dummy Variables:				
urban	-0.108*	0.0342	-0.137*	0.0324
rural	0.177*	0.0357	0.171*	0.0392
Lifecycle Dummy Variables:				
7<child<15	0.118*	0.0597	0.102	0.0537
16<child<17	0.007	0.076	0.129	0.0697
2+adlts<35	0.092	0.087	0.018	0.0745
35<2+adlts<59	0.168*	0.076	0.087	0.0687
2+adlts>60	0.114	0.093	-0.153	0.0833
1adlt<35	0.22	0.13	0.161	0.118
35<1adlt<59	0.125	0.117	-0.055	0.111
1adlt>60	-0.073	0.129	-0.495**	0.125
First-stage Residuals	0.069	0.26	-0.118	0.194
R^2	0.394		0.413	
N	1682		1338	

Table 5: Estimation results for the PLAM specification in Equation (12) for regular and non-regular households. Standard errors for the density-weighted average derivative obtained by bootstrap. Significance at 1% is denoted by ** and at 5% by *. N indicates the sample size.



(a) PRICE

(b) PRICE ELASTICITY

(c) SMOOTHED PRICE DENSITY

(d) AGE

Figure 5: Estimated nonparametric components for PRICE and AGE (Panels (a) and (d)) of the PLAM specification for regular, non-regular and all households. Panel (b) shows the estimated price elasticities for regular and non-regular households and Panel (c) the smoothed price density for the two groups.