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

Sebastiano Manzan\textsuperscript{a} and Dawit Zerom\textsuperscript{b}

This version: December 1, 2005

\textsuperscript{a}Department of Economics, University of Leicester, Leicester, United Kingdom
\textsuperscript{b}School of Business, University of Alberta, Edmonton, Canada

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. Methodologically, we propose a root-$n$ consistent estimator for the vector of parameters of the linear part of the PLAM that is semi-parametrically efficient. The estimator is also computationally more convenient compared to recently proposed alternative kernel smoothers. 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. 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 $-0.3$, at low prices, to $-0.45$, at high prices, suggesting that households might respond differently to price changes depending on the level of price. In addition, users of regular gasoline seem to be more sensitive to price changes compared to users of nonregular (premium and mid-range) gasoline.

Keywords: semiparametric methods, partially linear additive model, gasoline demand.
JEL codes: C14, D12
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. 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 before 1990, but much lower after the 1990’s. A study by the US Department of Energy (1996), for example, 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. 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.

By carefully addressing some data issues as well as using new econometric methodology, 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 where 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

\footnote{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.}
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 available in the RTECS is unreliable. As a proxy for the price variable (per household), EIA 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 EIA 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 the country and across grades of gasoline.

Second, we implement the partially linear additive model (hereinafter PLAM) as a reduced model for the gasoline demand. The PLAM set-up is a semiparametric one in the sense that it involves both a nonparametric and a linear part. The nonparametric component of PLAM is additive. In our gasoline demand analysis, the linear part includes 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. The PLAM set-up also allows interactions among the variables of the nonparametric part by incorporating them within the linear part. Thus the PLAM is a more general specification than the linear regression model while retaining ease of interpretability. Furthermore, a practical application of the PLAM specification to real data is scarce in the applied econometrics literature and hence our contribution can also be of a general interest for applied researchers.

Third, we propose a method to estimate the PLAM using kernel-type smoothers. In particular, we introduce an estimator of the linear parameter vector that is root-n consistent with a normal asymptotic distribution. Interestingly, when the true specification
is PLAM, the proposed estimator is asymptotically more efficient than an estimator that ignores the additive structure of the nonparametric part, such as that of Robinson (1988). While expanding the currently available approaches, our estimator also offers the following advantages. First, unlike the other kernel-based estimators (e.g., Fan et al., 1998; Fan and Li, 2003; and Moral and Rodriguez-Poo 2004), our estimator is semiparametrically efficient in the sense of Chamberlain (1992) when the error has a constant variance. Second, our estimator is computationally feasible as it reduces the computational cost of previous kernel methods by order of \( n \). From a practical standpoint, this advantage can be very significant when \( n \) is large and/or when implementing computer-intensive methods such as bootstrap or simulation.

This paper differs from that of Schmalensee and Stoker (1999) in several important dimensions. First, we construct price data based on vehicles information that permits the estimation of valid price elasticities using the RTECS database. In this sense, our study addresses the empirical problem raised by Schmalensee and Stoker (1999) with respect to the estimation of price elasticity. Second, we model gasoline demand using PLAM which is more efficient than the partially linear model when additivity is a reasonable approximation. Furthermore, the PLAM circumvents the dimensionality problem that arises from adopting the partially linear model (e.g., Robinson, 1988). Third, we propose a new feasible econometric methodology to estimate the PLAM.

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 \(-0.3\), at low prices, and \(-0.45\), at high price values. This result suggests that households might respond differently to price changes depending on the level of the fuel price. We further investigate this issue by considering separately the households that consume only ”regular” gasoline and those that purchase “non-regular” grades of gasoline\(^2\). The estimation results for the two groups show that regular users are more sensitive to price changes (estimated elasticity of \(-0.5\)) compared to non-regular users (that have an elasticity of \(-0.35\)). The price elasticity of regular gasoline is quite similar for both low and high prices. Instead, the demand for “non-regular” fuel is quite inelastic at low prices and becomes increasingly reactive at high prices. Separate analysis of the two groups also shows some significant differences in the effects of income, age and number of driver.

The remainder of the paper is organized as follows. In Section (2), we propose a semiparametric method for the estimation of the PLAM. We also outline computational details to facilitate practical implementation of the method. 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.

\(^2\)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.
2 A Semiparametric Method

In this section, we summarize the methodological contributions of the paper. In particular, we discuss the PLAM set-up and the associated estimation procedure.

2.1 Model

The PLAM takes the following form,

\[ Y_i = \beta_0 + X_i' \beta + m_1(Z_{1i}) + \ldots + m_q(Z_{qi}) + u_i \quad (i = 1, \ldots, n), \]  

(1)

where \( Y_i \) is a scalar dependent variable, \( X_i \) is a \( p \times 1 \) vector of explanatory variables, \( \beta = (\beta_1, \ldots, \beta_p)' \) is a \( p \times 1 \) vector of unknown parameters, \( \beta_0 \) is a scalar parameter, \( Z_i = (Z_{1i}, \ldots, Z_{qi})' \) is a \( q \times 1 \) vector of explanatory variables, \( m_1(\cdot), \ldots, m_q(\cdot) \) 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 distribution of the regressors \((X, Z)\) is left completely unspecified.

The PLAM is particularly attractive for the following reasons. On the one hand, with respect to the pure additive model,

\[ Y_i = \beta_0 + m_1(Z_{1i}) + \ldots + m_q(Z_{qi}) + u_i, \]  

(2)

the PLAM provides considerable flexibility by allowing interaction terms among the elements of \( Z \) enter as the linear part of the model. This is possible as the PLAM permits \( X_i \) to be a deterministic, but non-additive, function of \((Z_{1i}, \ldots, Z_{qi})\). Furthermore, the PLAM allows a subset or all of the variables in \( X \) to be discrete, while pure additive models admit continuous variables only. On the other hand, compared to the partially linear model with non-structured nonparametric component, i.e.

\[ Y_i = X_i' \beta + m(Z_{1i}, \ldots, Z_{qi}) + u_i, \]  

(3)

the PLAM has explicit nonparametric components that can be estimated with a one-dimensional nonparametric rate and hence avoid the so-called curse of dimensionality (Stone, 1985 and 1986). Furthermore, when the true data generating model is PLAM, simply using model (3) (i.e. ignore the additivity of \( m(\cdot) \)) to estimate \( \beta \) can lead to an inefficient estimate of \( \beta \). Note also that model (3) does not allow the intercept \( \beta_0 \) but only “slope” coefficients to be estimated.

2.2 Estimation of the parametric part

Here we introduce a semiparametrically efficient and computationally feasible approach to estimate the parametric part of the PLAM. Our method expands the currently available approaches, notably the works of Fan et al. (1998), Fan and Li (2003) and Moral and Rodriguez-Poo (2004) that are kernel-based and that of Li (2000) which is a series-based method. Below we outline the construction of the estimator.
Consider the model $A_i = \theta^{(A)}(Z_i) + V_i$ where $A$ is a vector- or real valued dependent variable, $Z$ is a $q \times 1$ vector of explanatory variables, $\theta^{(A)}(z) \equiv E[A_i|Z_i = z]$ is an unknown vector- or real valued smooth function and $V_i$ is an unobservable noise component that satisfies $E[V_i|Z_i] = 0$. Let $\theta^{(A)}(z) = \theta_1^{(A)}(z_1) + \ldots + \theta_q^{(A)}(z_q)$ be a function chosen subject to the constraints $E[\theta_1^{(A)}(z_1)] = 0$ (for all $j = 1, \ldots, q$) to minimize

$$E[(\theta^{(A)} - \theta^{(A)})(\theta^{(A)} - \theta^{(A)})'].$$

Then, we say $\theta^{(A)}(\cdot)$ is the closest (best) additive approximation to $\theta^{(A)}(\cdot)$ in $L_2^3$. Stone (1985) proved the existence of $\theta^{(A)}(\cdot)$ that satisfies (4) when $\theta^{(A)}(\cdot)$ is real valued.

Now, for $j = 1, \ldots, q$, let the vector $W_j$ denote the set of all $Z$ variables excluding $Z_j$, i.e. $W_j = (Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots, Z_q)'$. Following the idea of Kim et al. (1999), let’s define a 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 $p_{w}(\cdot)$ are the density functions of $Z_j$ and $W_j$, respectively and $p(\cdot)$ is the joint probability function of $Z = (Z_j, W_j)$. Let $h_j^{(A)}(z_j) = E[\phi(Z_j, W_j)A|Z_j = z_j]$. Noting that $E[\phi(Z_j, W_j)V|Z_j = z_j] = 0$, it is easy to see that

$$h_j^{(A)}(z_j) = \int \theta^{(A)}(z)p_{w}(W_j)dW_j \quad (j = 1, \ldots, q).$$

We can observe from (5) that $h_1^{(A)}(z_1), \ldots, h_q^{(A)}(z_q)$ are the $L_2(\mathcal{Q})$ (with $\mathcal{Q}$ being a product probability measure) projections of $\theta^{(A)}(z)$ onto the space functions of $z_1, \ldots, z_q$, respectively. Further, let $h_j^{(A)}(z_j) = h_j^{(A)}(z_j) - E[h_j^{(A)}(z_j)]$. Then, as per definition of additivity, we may define the sum

$$h^{(A)}(z) = \sum_{j=1}^q h_j^{(A)}(z_j),$$

as the best additive approximation to the function $\theta^{(A)}(z)$, i.e.

$$\inf_{\theta^{(A)} \in \mathcal{J}} E[\{\theta^{(A)}(z) - h^{(A)}(z)[\theta^{(A)}(z) - h^{(A)}(z)]\}^2].$$

When $\theta^{(A)}(\cdot)$ is in fact additive ($\theta^{(A)}(\cdot) \in \mathcal{J}$), for all $j = 1, \ldots, q$, $h_j^{(A)}(z_j) = \theta_j^{(A)}(z_j)$. Hence, $h^{(A)}(z) = \theta^{(A)}(z)$. Thus, we can use $h_j^{(A)}(z_j)$ to estimate the additive components of the pure additive model as given in (2).
2.2.2 An estimator for $\beta$

Let’s assume that the additive components in (1) satisfy $E[m_j(Z_{ji})] = 0$ for all $j = 1, \ldots, q$. We set the intercept $\beta_0$ to zero in the sequel without loss of generality. Note that in the foregoing discussions we used $\phi(\cdot, \cdot)$ to help derive an additive approximation to the unknown smooth function $\theta^{(A)}(Z)$. Here, we use this function to help reduce the PLAM into a linear-like model that facilitates the estimation of $\beta$. It can be shown that $\phi(\cdot, \cdot)$ satisfies

$i) \quad E[\phi(Z_j, W_j)|Z_j = z_j] = 1$ and $ii) \quad E[\phi(Z_j, W_j)m_k(Z_j)|Z_j = z_j] = 0$ for $k \neq j$.

Now, applying these properties on model (1),

$$h_j^{(Y)}(Z_{ji}) = m_j(Z_{ji}) + (h_j^{(X)}(Z_{ji}))' \beta \quad (j = 1, \ldots, q). \quad (8)$$

To arrive at the above equation, we use notations from Section (2.2.1) and replace $A$ by $Y$ or $X$. Observe also that $E[\phi(Z_{ji}, W_{ji})Y_i|Z_{ji}]$ and $E[\phi(Z_{ji}, W_{ji})X_i|Z_{ji}]$ correspond to $h_j^{(Y)}(Z_{ji})$ and $h_j^{(X)}(Z_{ji})$, respectively. Now, adding the $q$-equations in (8) and subtracting the result from (1), we obtain

$$Y_i - h^{(Y)}(Z_i) = (X_i - h^{(X)}(Z_i))' \beta + u_i, \quad (9)$$

where we have used the definition in (6).

Notice from (9) that we have reduced the PLAM (i.e. Equation (1)) to a linear model where the dependent and independent variables are expressed in deviations form around the best additive approximations, $h^{(Y)}(\cdot)$ and $h^{(X)}(\cdot)$, respectively. In this way, an estimator of $\beta$ can be defined as the vector of OLS coefficients $\hat{\beta}$ of the deviation $Y_i - h^{(Y)}(Z_i)$ on $X_i - h^{(X)}(Z_i)$. If we were to ignore the additive structure of $m(z_1, z_2, \ldots, z_q)$ and assumed instead the partial linear model (3), we would have estimated $\beta$ by regressing $Y_i - \theta^Y(Z_i)$ on $X_i - \theta^X(Z_i)$ (using the definitions in Section (2.2.1)) as proposed by Robinson (1988).

Unfortunately, the estimation of $\beta$ given above is unfeasible since the functions $h^{(Y)}(Z_i)$ and $h^{(X)}(Z_i)$ are unknown quantities. One solution is to replace these quantities by their kernel estimators. Let $\tilde{A}_i$ denote an estimator of $h^{(A)}(Z_i)$ where $A$ can be $Y$ or $X$. Based on (6), we define $\tilde{A}_i$ as

$$\tilde{A}_i = \sum_{j=1}^{q} \tilde{A}_i^j \quad (10)$$

where $\tilde{A}_i^j$ denotes an estimate of $h^{(A)}_j(Z_{ji}) \equiv E[\phi(Z_{ji}, W_{ji})A_i|Z_{ji}]$. We compute $\tilde{A}_i^j$ by

$$\tilde{A}_i^j = \frac{1}{(n - 1)b} \sum_{\ell \neq i}^n K\left(\frac{Z_{j\ell} - Z_{ji}}{b}\right) \frac{\hat{p}_w(W_{j\ell})}{\hat{p}(Z_{j\ell}, W_{j\ell})} A_{\ell} \quad (i = 1, \ldots, n; \quad j = 1, \ldots, q), \quad (11)$$

where $K(\cdot)$ is a kernel function, $b$ is a bandwidth (or smoothing parameter), and $\hat{p}_w(\cdot)$
and \( \hat{\rho}(\cdot) \) are kernel-smoothers of the corresponding densities. It should be noted that \( \hat{A}_i^j \) is a leave-out estimator in the sense that the \( i \)th observation \((A_i, Z_i)\) is not used in the computation. Finally, using the kernel smoothers defined above, an estimator of \( \beta \) for the PLAM can be defined as the vector of OLS coefficients \( \hat{\beta} \) of the deviation \((Y_i - \hat{Y}_i)\) on \((X_i - \hat{X}_i)^4\).

In a short communication, Manzan and Zerom (2005) derives the asymptotic properties \( \hat{\beta} \). Let \( \Phi = E[\varepsilon_i \varepsilon_i'] \) where \( \varepsilon_i = X_i - \hat{X}_i \). Under some regularity conditions, and provided \( \Phi \) is positive definite, we prove that

\[
n^{1/2}(\hat{\beta} - \beta) \to N(0, \Sigma), \tag{12}\]

where \( \Sigma = \Phi^{-1} \Omega \Phi^{-1} \) and \( \Omega = E[\sigma_i^2(X_i, Z_i, W_i)\varepsilon_i \varepsilon_i'] \). The variance-covariance matrix \( \Sigma \) can also be consistently estimated by \( \hat{\Sigma} = \hat{\Phi}^{-1} \hat{\Omega} \hat{\Phi}^{-1} \) where \( \hat{\Phi} = n^{-1} \sum_i (X_i - \bar{X}_i)(X_i - \bar{X}_i)' \), \( \hat{\Omega} = n^{-1} \sum_i \hat{u}_i^2(X - \bar{X}_i)(X - \bar{X}_i)' \), and \( \hat{u}_i = Y_i - \hat{Y}_i - (X_i - \bar{X}_i)\hat{\beta} \).

Using the above result and the discussion in Section (2.2.1), Manzan and Zerom (2005) show that the proposed estimator has the following advantages compared to existing kernel-based estimators. With respect to Fan et al. (1998) and Fan and Li (2003), our estimator attains the semiparametric efficiency bound (Chamberlain, 1992) of the partially linear additive model under homoskedastic errors. Furthermore, when the true specification is PLM, the proposed estimator is asymptotically more efficient than an estimator that ignores the additive structure of the nonparametric part, such as that of Robinson (1988).

### 2.2.3 Computation and bandwidth choice

As noted before, the implementation of \( \hat{\beta} \) requires the smoothed estimates \( \hat{Y} \) and \( \hat{X} \). Below we outline a computationally convenient procedure that facilitates the implementation of \( \hat{Y} \) and \( \hat{X} \) in matrix-oriented statistical software packages such as Gauss, Matlab, Ox or R. First, define the following \( n \times n \) smoother matrices

\[
S_j^z = \left[ \frac{1}{(n-1)b} k \left( \frac{Z_{ji} - Z_{j\ell}}{b} \right) \right]_{i,\ell}; \quad S_j^w = \left[ \frac{1}{nb_{\beta-1}} k \left( \frac{W_{ji} - W_{j\ell}}{b} \right) \right]_{i,\ell}.
\]

To make sure that \( \ell \neq i \), also define \( S_{j*}^z = S_j^z \odot T \) where \( T \) is an \( n \times n \) symmetric matrix such that when multiplied element by element with \( S_j^z \) makes the diagonal elements of \( S_j^z \) zero. The operators \( \odot \) denotes matrix Hadamard product. Then, one can compute the \( n \times 1 \) vector estimates \( \hat{A}_j = (\hat{A}_1^j, \ldots, \hat{A}_n^j)' \) (\( A \) can be \( X \) or \( Y \)) in a single step by

\[
\hat{A}_j = S_{j*}^z \{ A \odot (S_j^w e)/((n - 1)(S_j^z \odot S_j^w e)) \}
\]

\(^4\)In empirical work, one may also be interested in estimating the intercept \( \beta_0 \). It is easy to see that when \( \beta_0 \neq 0 \), equation (9) would become \( Y_i - h(Y)(Z_i) = (1 - g)\beta_0 + (X_i - h(X)(Z_i))'\beta + u_i \). Hence, we would instead regress \( Y_i - \hat{Y}_i \) on \((1, (X_i - \bar{X}_i)') \) so as to incorporate the estimation of the intercept.
where ./ denote matrix Hadamard division, e = (1, \ldots, 1)' and A = (A_1, \ldots, A_n)'. Notice that to compute \( \tilde{A} = (\tilde{A}_1, \ldots, \tilde{A}_n)' \) we only need \( O(n^2) \) operations, while by comparison marginal integration based procedures (e.g., Fan et al., 1998) involve \( O(n^3) \) operations. Thus, there is a reduction in computation by order of the sample size \( n \). From a practical standpoint, this computational advantage can be very significant when \( n \) is large and/or when implementing computer-intensive methods such as bootstrap or cross-validation.

Obviously the implementation of the above smoothers \( \tilde{Y} \) and \( \tilde{X} \) requires choices to be made on both the bandwidth \( b \) and the type of kernel function \( K(\cdot) \). We consider bandwidths \( b \) that decrease to 0 at the rate \( n^{-2/7} \), i.e., \( b = an^{-2/7} \) 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.

Using the above argument, the problem reduces to the choice of \( a \). To select this, we consider a cross-validation (CV) procedure. Based on the formulation in (9), we select the CV constant \( (a) \) as follows:

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

where, \( \tilde{A}_i \) (A can be \( X \) or \( Y \)) is an estimator of \( h(A)_i \) at \( (A_i, Z_i) \) without using the observation \( (A_i, Z_i) \). Then, we choose the bandwidth as \( \hat{b} = \hat{a}n^{-2/7} \).

### 2.3 Estimation of the nonparametric components

Based on (8) and using the estimator \( \hat{\beta} \), we can compute \( m_j(\cdot) \) as

\[
\hat{m}_j(Z_{ji}) = \tilde{Y}_j - (\tilde{X}_j)' \hat{\beta} \quad (j = 1, \ldots, q),
\]

where \( \tilde{A}_j \) (A can be \( Y \) or \( X \)) is defined in (11). 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 \( \tilde{Y}_j \) and \( \tilde{X}_j \), the asymptotic distribution of the additive components \( \hat{m}_j(\cdot) \) will remain unaffected by the estimation of \( \beta \) and follows from the distribution of \( \tilde{Y}_j - \tilde{X}_j \). In this way, the estimation of \( \beta \) 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 (13) 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

\[
m_1(z_1), m_2(z_2), \ldots, m_{j-1}(z_{j-1}), m_{j+1}(z_{j+1}), \ldots, m_q(z_q)
\]

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 accuracy of the nonparametric components, ensuring their efficiency is warranted. For example, the price effect (the main
focus of the paper) will be modeled as being nonparametric. Following the approach of Kim et al. (1999), we implement a one-step backfitting procedure. First, use \( \hat{\beta} \) to compute \( \hat{Y}_i = Y_i - X'_i \hat{\beta} \). Second, for each \( j \in [1, 2, \ldots, q] \), compute the partial residuals \( \hat{\varepsilon}_i^j = \hat{Y}_i - \sum_{k \neq j} \hat{m}_k(Z_{ki}) \) where the \( \hat{m}_k(\cdot) \) estimates are obtained from (13). Finally, apply a local linear smoothing of \( \hat{\varepsilon}_i^j \) on \( Z_{ji} \). Let’s denote the resulting nonparametric components estimators by \( \hat{m}_j^v(\cdot) \). Theoretical results on consistency and asymptotic normality of the oracle estimators within the pure additive set-up have already been established by Kim et al. (1999). Since the estimation of \( \beta \) does not affect the nonparametric components due to their faster rate of convergence, we do not need a new asymptotic theory for \( \hat{m}_j^v(\cdot) \).

It should be noted that in the implementation of the one-step backfitting, one needs to choose a different bandwidth for the local linear step. The asymptotic theory suggests that the bandwidth be chosen as \( \sim cn^{-1/5} \). Following this, the bandwidth are chosen as \( \hat{\sigma}_e n^{-1/5} \) where \( \hat{\sigma}_e \) is the standard deviation of \( \hat{\varepsilon} \). Our experience with the one-step estimator shows that results are not sensitive to bandwidth values so long as they are chosen at \( n^{-1/5} \) rate.

Finally, we outline a procedure for calculating point-wise confidence intervals of the nonparametric estimates \( \hat{m}_j^v(\cdot) \). Because the asymptotic variance of \( \hat{m}_j^v(\cdot) \) 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 \( \hat{m}_j^v(\cdot) \), the residuals of the PLAM in Equation (1) are given by

\[
\hat{u}_i = \hat{Y}_i - \sum_{j=1}^q m_j^v(Z_{ji}).
\]  

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

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

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

\[
Y_{i,s}^* = X'_i \hat{\beta} + \sum_{j=1}^q m_j^v(Z_{j,i}) + \tilde{u}_{i,s}^*.
\]

For each replicate \( (X, Z, Y^*_i) \), we compute the nonparametric component (denoted by \( m_j^{v,s}(z_j) \)) at fixed values \( Z_j = z_j \). Then, bootstrap confidence interval for \( m_j(z_j) \) is simply calculated using the appropriate percentiles of \( \{m_j^{v,s}(z_j)\}_{s=1}^S \).

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 (3) for the 1988 and 1991 RTECS data 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 households observations 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 model gasoline demand by considering log \( \text{price} \), log \( \text{age} \) and log \( \text{income} \) as additive nonparametric components in the following PLAM specification

\[
\log \text{gals}_i = m_P(\log \text{price}_i) + m_A(\log \text{age}_i) + m_I(\log \text{income}_i) + X'_i \beta + u_i
\]  

where \( m_P(\cdot) \), \( m_A(\cdot) \) and \( m_I(\cdot) \) are unspecified smooth functions, \( \text{gals}_i \) is gasoline consumption of household \( i \) measured in gallons, \( \text{price}_i \) is the average cost per gallon, \( \text{age}_i \) is the age of the household \( i \) head, \( \text{income}_i \) is the annual income of a household and \( X_i \) is a vector of household characteristics: log \( \text{number of drivers} \), log \( \text{household size} \), and dummy
variables for residence (urban, suburban and rural) and dummy variables for the lifecycle categories.

The above model differs from that in Schmalensee and Stoker (1999) in two important aspects. First, their model does not consider the gasoline price variable (due to data problem). Instead they include location (region) variables in order to ensure that the other parameter estimates do not suffer from possible bias due to the omission of the price variable. Our study complements Schmalensee and Stoker (1999) by providing valid price elasticities that are crucial in the analysis of gasoline demand. Second, unlike model (15) where log income and log age are treated additively, Schmalensee and Stoker (1999) model these variables as a jointly nonparametric function $m_{A,I}(\cdot, \cdot)$. We address the additivity assumption below.

The additive specification of (15) may seem restrictive. But, as we mentioned in Section (2), we can augment the model by allowing interactions among log price, log age and log income. These interactions can be added to the linear part of PLAM and tested with standard tests. These interactions terms can not be considered in both pure additive models (Equation 2) and partial linear models (Equation 3). We tested the significance of the interaction terms: $(\log \text{price} \times \log \text{age})$, $(\log \text{price} \times \log \text{income})$, and $(\log \text{age} \times \log \text{income})$. Both individual and joint-tests strongly indicate that none of the interactions are significant at the 10% level. By slicing the estimated surface $\hat{m}_{A,I}(\cdot, \cdot)$, Schmalensee and Stoker (1999) also noticed that log age and log income could be modeled additively (see Figures (1) and (2) on pages 651 and 652). Thus, in addition to being convenient for the computation of the elasticity and its ease of interpretation, the PLAM seems also to be supported by the data.

As argued by Yatchew and No (2001) in their conclusion section (for Canadian data), we also consider the possible bias in our estimate due to the possible endogeneity of the price variable. This problem arises when $E(u_i | \log \text{price}_i) \neq 0$. In this case the nonparametric estimator is not consistent due to the correlation between the error term in Equation (15) and the log price variable. We follow Blundell et al. (1998) to account for the possible endogeneity of the price variable. Assume there is a set of instrumental variables $Z_i$ such that

$$
 \log \text{price}_i = Z_i' \pi + v_i
$$

with $E(v_i|Z_i) = 0$. We can then include the residuals $v_i$ in Equation (15), that is,

$$
 \log \text{gals}_i = m_P(\log \text{price}_i) + m_A(\log \text{age}_i) + m_I(\log \text{income}_i) + \rho v_i + X_i' \beta + u_i \tag{17}
$$

where we assume that $E(u_i | \log \text{price}_i) = 0$. Under these assumption, the resulting estimator of $m_P(\cdot)$ is consistent. The null hypothesis of exogeneity of the price variable can be easily tested using the least squares estimator of $\rho$. As suggested in Schmalensee and Stoker (1999), a natural candidate as instrumental variables in Equation (16) are the regional dummy variables. Equation (17) is estimated by including in the PLAM specification the fitted residuals $\hat{v}_i$ from the first-stage regression in Equation (16).
3.2 Results and Discussion

We begin by discussing the method that EIA 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 the EIA. They computed the price effect from the nonparametric estimate of the function $m_{PI}(.,.)$ by slicing the curve along the income dimension. The $m_{PI}(.,.)$ was estimated in the framework of the partial linear model in Equation (3) using the approach of Robinson (1988).

The EIA-RTECS does not collect fuel purchase diaries\(^5\). 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\(^6\)) 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 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 EIA 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

---

\(^5\)The EIA stopped collecting purchase diaries starting from the 1988 survey 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 the EIA.

\(^6\)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.
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\(^7\).

**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 the EIA. 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 (17) using the average cost (the price variable) calculated by the EIA. 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\(^8\). 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**

### 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, the EIA-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 by the EIA procedure.

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

---

\(^7\)The sample includes also 1398 households that have more than one vehicle and purchase different gasoline grades.

\(^8\)The 1994 data has not been investigated by Schmalensee and Stoker (1999).
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\(^9\). 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\(^{10}\) of the different gasoline grades for each of the 9 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 EIA methodology.

Table (3) here

We assign an average cost to each household which is defined as total expenditure (using the last fuel price) divided by the total gallons consumed. For the households that reported prices for only part of their cars, we input 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.

---

\(^9\) 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.

\(^{10}\) We deflated prices in 1994 to 1991 levels using the CPI Index.
compared to the EIA procedure (at the division level instead of regional). The average cost, \( price_{i} \), for household \( i \) is given by

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

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) where 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 EIA measure. This is due to the effect of accounting for the *intra-divisional* dispersion of prices\(^{11}\). The bi-modality that was apparent for the EIA price measure has now 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 (17) 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 (15) 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 \( \log \text{age} \) and \( \log \text{income} \) 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 \( \log \text{price} \), \( \log \text{age} \) and \( \log \text{income} \) along with the estimated price elasticity\(^{12}\). 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.

\(^{11}\) 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\$.

\(^{12}\) 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.
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 as follows. The first interesting result of the analysis is that the log price component is negatively sloped in the complete range of the variable. Panel (c) of Figure (4) shows the nonparametric estimate of the price elasticity. For low prices it is close to -0.30 and increases toward -0.45 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 for all their vehicles regular gasoline (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 what 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 number of drivers variable is highly significant with an estimated elasticity of 0.669. 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.

Table (4) here

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\textsuperscript{13}: those consuming (for

\textsuperscript{13}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” and “non-regular” in order to have a large number of observations in each group. The households that
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 (17) 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 density-weighted average price derivative for regular users is equal to -0.52 and for non-regular to -0.35. Households that buy exclusively regular gasoline are more sensitive to price changes compared to households that purchase non-regular grades. Panels (a)-(c) of Figure (5) shows the estimated price component, the price elasticity of the groups and the smoothed price density for the two groups. The demand for regular gasoline is slightly flat for low prices and then decreases in a linear fashion. The plot of the price elasticity suggests it is approximately constant on the interval of price variation. Instead, the price component for non-regular users is flat for prices below $1 and then slopes downward. The price elasticity starts at about -0.25 for low prices and increases toward -0.48 for high prices. The price component for the complete sample lies between the regular and non-regular price components. The distribution of prices faced by the two groups (see Panel (c) in the Figure) implies that the aggregate curve is close to the regular one for low prices (where most households consume regular gasoline) and gradually shifts toward the non-regular price component at high prices (where most households purchase non-regular grades). The finding for the full sample that the price elasticity increases (in absolute value) at high prices can be explained as the result of the larger sensitivity of both regular and non-regular gasoline demand when prices are large (approximately above $1.1).

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 -0.263). 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 reported prices for their vehicles is composed of 3020 observations (about half the sample of Yatchew and No (2001)) of which 1682 consumed regular for all their vehicles, 319 purchased exclusively midgrade, 190 only premium, and the remaining 829 bought different grades.
compared to regular (0.20 and 0.13, respectively) while the opposite effect holds for the drivers effect (0.52 and 0.77, 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 illustrate a novel approach to the estimation of partially linear models with an application to gasoline demand in the United States. We assume that the non-parametric part of the model is additive and refer to the model as the Partially Linear Additive Model (PLAM). The model is flexible in the specification because it includes both a parametric and nonparametric part. In addition, for each variable treated non-parametrically we estimate a component that allows interpreting its relationship with the dependent variable. We show that the estimator of the vector of parameters is consistent and achieves the semi-parametric efficiency bounds of Chamberlain (1992). The proposed estimator has other advantages, such as better finite sample properties and computational efficiency compared to alternative non-parametric estimators.

In an extended application of the PLAM, we reexamine the issue of price elasticity as raised 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 the EIA. In particular, we show that there is significant dispersion in gasoline prices across the US and across grades of fuel. By estimating a PLAM with the 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; suggesting 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 \(-0.5\)) compared to non-regular users (that have an elasticity of \(-0.35\)). The price elasticity of regular gasoline is quite similar for both low and high prices. Instead, the demand for “non-regular” fuel is quite inelastic at low prices and becomes increasingly reactive at high prices. Separate analysis of the two groups also shows some 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 is a new empirical finding. In light of this result, further empirical investigation with more recent data (when available) is warranted.
**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 Gallon) estimate\(^\text{14}\) to calculate the amount of gallons consumed 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

\(^{14}\)The estimate is provided by the Environmental Protection Agency (EPA) and is specific to the type of vehicles considered and the fuel type purchased.
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


<table>
<thead>
<tr>
<th>Variables</th>
<th>1991</th>
<th>1994</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St. Dev.</td>
</tr>
<tr>
<td>log(gallons)</td>
<td>6.75</td>
<td>0.718</td>
</tr>
<tr>
<td>log(income)</td>
<td>3.31</td>
<td>0.794</td>
</tr>
<tr>
<td>log(drivers)</td>
<td>0.548</td>
<td>0.4</td>
</tr>
<tr>
<td>log(hld size)</td>
<td>0.882</td>
<td>0.536</td>
</tr>
<tr>
<td>log(age)</td>
<td>3.76</td>
<td>0.363</td>
</tr>
</tbody>
</table>

**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 35-39 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 35-39 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.
Figure 1: EIA 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 a 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.
Figure 2: The estimated price component $m_p[\log(price_i)]$ for the PLAM specification in Equation 17 when the EIA price measure is considered. The estimate is based on the pooled 1991 and 1994 surveys, 5260 households. 95% confidence intervals obtained by bootstrap.
<table>
<thead>
<tr>
<th></th>
<th>1991 Valid</th>
<th>1994 Valid</th>
<th>Pooled Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>log(gallons)</td>
<td>6.86</td>
<td>6.75</td>
<td>6.85</td>
</tr>
<tr>
<td></td>
<td>(0.68)</td>
<td>(0.72)</td>
<td>(0.72)</td>
</tr>
<tr>
<td>log(age)</td>
<td>3.78</td>
<td>3.76</td>
<td>3.80</td>
</tr>
<tr>
<td></td>
<td>(0.34)</td>
<td>(0.36)</td>
<td>(0.34)</td>
</tr>
<tr>
<td>log(income)</td>
<td>3.45</td>
<td>3.31</td>
<td>3.31</td>
</tr>
<tr>
<td></td>
<td>(0.73)</td>
<td>(0.79)</td>
<td>(0.72)</td>
</tr>
<tr>
<td>log(drivers)</td>
<td>0.61</td>
<td>0.55</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>(0.39)</td>
<td>(0.40)</td>
<td>(0.39)</td>
</tr>
<tr>
<td>log(hld size)</td>
<td>0.91</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>(0.52)</td>
<td>(0.53)</td>
<td>(0.52)</td>
</tr>
<tr>
<td>log(gallons) by gasoline grade:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular</td>
<td>6.78</td>
<td>6.69</td>
<td>6.80</td>
</tr>
<tr>
<td></td>
<td>(0.72)</td>
<td>(0.74)</td>
<td>(0.75)</td>
</tr>
<tr>
<td>Midgrade</td>
<td>6.62</td>
<td>6.48</td>
<td>6.58</td>
</tr>
<tr>
<td></td>
<td>(0.71)</td>
<td>(0.69)</td>
<td>(0.80)</td>
</tr>
<tr>
<td>Premium</td>
<td>6.67</td>
<td>6.50</td>
<td>6.61</td>
</tr>
<tr>
<td></td>
<td>(0.63)</td>
<td>(0.69)</td>
<td>(0.69)</td>
</tr>
<tr>
<td>More grades</td>
<td>7.12</td>
<td>7.07</td>
<td>7.13</td>
</tr>
<tr>
<td></td>
<td>(0.49)</td>
<td>(0.50)</td>
<td>(0.52)</td>
</tr>
</tbody>
</table>

Gasoline Grade (in % of total):

<table>
<thead>
<tr>
<th></th>
<th>1991</th>
<th>1994</th>
<th>Pooled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular</td>
<td>0.55</td>
<td>0.53</td>
<td>0.56</td>
</tr>
<tr>
<td>Midgrade</td>
<td>0.11</td>
<td>0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>Premium</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>More Grades</td>
<td>0.28</td>
<td>0.28</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Number of Vehicles (in % of total):

<table>
<thead>
<tr>
<th></th>
<th>1991</th>
<th>1994</th>
<th>Pooled</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>0.21</td>
<td>0.28</td>
<td>0.22</td>
</tr>
<tr>
<td>Two</td>
<td>0.40</td>
<td>0.39</td>
<td>0.40</td>
</tr>
<tr>
<td>Three</td>
<td>0.24</td>
<td>0.20</td>
<td>0.23</td>
</tr>
<tr>
<td>More</td>
<td>0.15</td>
<td>0.12</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Total          | 1571     | 2697     | 3020      |

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.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>New England</td>
<td>115.92</td>
<td>126.53</td>
<td>135.05</td>
<td>109.05</td>
<td>114.21</td>
<td>125.01</td>
</tr>
<tr>
<td></td>
<td>(7.71),[118]</td>
<td>(7.86),[32]</td>
<td>(9.82),[40]</td>
<td>(9.58),[69]</td>
<td>(9.44),[18]</td>
<td>(10.11),[28]</td>
</tr>
<tr>
<td>Mid Atlantic</td>
<td>110.31</td>
<td>119.19</td>
<td>131.85</td>
<td>105.31</td>
<td>113.14</td>
<td>121.32</td>
</tr>
<tr>
<td></td>
<td>(9.30),[254]</td>
<td>(13.02),[31]</td>
<td>(12.89),[88]</td>
<td>(8.95),[217]</td>
<td>(7.33),[48]</td>
<td>(10.23),[75]</td>
</tr>
<tr>
<td>E/N Central</td>
<td>101.97</td>
<td>110.55</td>
<td>117.38</td>
<td>96.72</td>
<td>102.16</td>
<td>109.43</td>
</tr>
<tr>
<td></td>
<td>(9.32),[320]</td>
<td>(11.77),[33]</td>
<td>(17.62),[52]</td>
<td>(7.65),[346]</td>
<td>(9.75),[67]</td>
<td>(12.02),[76]</td>
</tr>
<tr>
<td>W/N Central</td>
<td>100.4</td>
<td>99.08</td>
<td>108.72</td>
<td>95.12</td>
<td>98.81</td>
<td>104.25</td>
</tr>
<tr>
<td></td>
<td>(10.6),[344]</td>
<td>(9.61),[39]</td>
<td>(12.86),[61]</td>
<td>(9.28),[190]</td>
<td>(8.16),[26]</td>
<td>(7.14),[23]</td>
</tr>
<tr>
<td>South Atlantic</td>
<td>103.36</td>
<td>112</td>
<td>121.69</td>
<td>96.72</td>
<td>103.88</td>
<td>113.99</td>
</tr>
<tr>
<td></td>
<td>(9.19),[182]</td>
<td>(8.11),[48]</td>
<td>(8.68),[65]</td>
<td>(8.95),[270]</td>
<td>(13.99),[82]</td>
<td>(8.93),[84]</td>
</tr>
<tr>
<td>E/S Atlantic</td>
<td>102.34</td>
<td>109.25</td>
<td>115.33</td>
<td>96.48</td>
<td>104.49</td>
<td>113.45</td>
</tr>
<tr>
<td></td>
<td>(7.85),[151]</td>
<td>(8.69),[24]</td>
<td>(9.68),[46]</td>
<td>(6.75),[117]</td>
<td>(5.68),[23]</td>
<td>(113.45),[53]</td>
</tr>
<tr>
<td>W/S Atlantic</td>
<td>102.77</td>
<td>112.55</td>
<td>117.07</td>
<td>96.91</td>
<td>106.01</td>
<td>109.45</td>
</tr>
<tr>
<td></td>
<td>(8.93),[137]</td>
<td>(12.75),[29]</td>
<td>(11.56),[59]</td>
<td>(7.06),[181]</td>
<td>(5.38),[39]</td>
<td>(8.95),[63]</td>
</tr>
<tr>
<td>Mountain</td>
<td>102.71</td>
<td>103.5</td>
<td>111.65</td>
<td>107.69</td>
<td>112.1</td>
<td>117.24</td>
</tr>
<tr>
<td></td>
<td>(8.56),[198]</td>
<td>(10.95),[12]</td>
<td>(10.70),[26]</td>
<td>(8.40),[131]</td>
<td>(5.69),[13]</td>
<td>(9.65),[22]</td>
</tr>
<tr>
<td>Pacific</td>
<td>111.40</td>
<td>113.62</td>
<td>129.63</td>
<td>111.82</td>
<td>120.04</td>
<td>127.61</td>
</tr>
<tr>
<td></td>
<td>(11.94),[258]</td>
<td>(14.23),[29]</td>
<td>(17.88),[95]</td>
<td>(7.48),[198]</td>
<td>(8.84),[36]</td>
<td>(9.55),[60]</td>
</tr>
</tbody>
</table>

Table 3: Average Real Prices in $ cents per gallon based on vehicles data. The number in (·) is the standard deviation of the price per division and per grade of gasoline and [·] 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.
Figure 4: Estimated nonparametric components for PRICE, AGE and INCOME of the PLAM specification in Equation (17) with bootstrap standard errors. Panel (c) is the nonparametric estimate of the price elasticity.
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 (17) 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). Significance at 1% is denoted by ** and at 5% by *. N indicates the sample size.
Table 5: Estimation results for the PLAM specification in Equation (17) for regular and non-regular households. Significance at 1% is denoted by ** and at 5% by *. N indicates the sample size.

<table>
<thead>
<tr>
<th></th>
<th>Regular Average Derivative</th>
<th>Non-Regular Average Derivative</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(price)</td>
<td>-0.525</td>
<td>-0.356</td>
</tr>
<tr>
<td>log(age)</td>
<td>-0.263</td>
<td>-0.009</td>
</tr>
<tr>
<td>log(income)</td>
<td>0.129</td>
<td>0.201</td>
</tr>
<tr>
<td>log(drivers)</td>
<td>0.796* 0.0636</td>
<td>0.543* 0.0624</td>
</tr>
<tr>
<td>log(hld size)</td>
<td>0.062 0.0808</td>
<td>0.096 0.0736</td>
</tr>
<tr>
<td><strong>Residence Dummy Variables:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>urban</td>
<td>-0.103* 0.0344</td>
<td>-0.132* 0.0324</td>
</tr>
<tr>
<td>rural</td>
<td>0.182* 0.0358</td>
<td>0.178* 0.0392</td>
</tr>
<tr>
<td><strong>Lifecycle Dummy Variables:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7&lt;child&lt;15</td>
<td>0.116 0.0594</td>
<td>0.112 0.0538</td>
</tr>
<tr>
<td>16&lt;child&lt;17</td>
<td>-0.005 0.076</td>
<td>0.152 0.0695</td>
</tr>
<tr>
<td>2+adlts&lt;35</td>
<td>0.095 0.0865</td>
<td>0.012 0.0749</td>
</tr>
<tr>
<td>35&lt;2+adlts&lt;59</td>
<td>0.153 0.0756</td>
<td>0.118 0.0686</td>
</tr>
<tr>
<td>2+adlts&gt;60</td>
<td>0.034 0.0909</td>
<td>-0.135* 0.0836</td>
</tr>
<tr>
<td>1adlt&lt;35</td>
<td>0.224 0.131</td>
<td>0.157 0.118</td>
</tr>
<tr>
<td>35&lt;1adlt&lt;59</td>
<td>0.122 0.117</td>
<td>-0.021 0.11</td>
</tr>
<tr>
<td>1adlt&gt;60</td>
<td>-0.171* 0.127</td>
<td>-0.473* 0.125</td>
</tr>
<tr>
<td>First-stage Residuals</td>
<td>-0.017 0.25</td>
<td>-0.152 0.2</td>
</tr>
<tr>
<td>R²</td>
<td>0.388</td>
<td>0.417</td>
</tr>
<tr>
<td>N</td>
<td>1682</td>
<td>1338</td>
</tr>
</tbody>
</table>
Figure 5: Estimate nonparametric components for PRICE and AGE of the PLAM specification with bootstrap standard errors. Panel (b) shows the estimated price elasticities for regular and non-regular households and Panel (c) the smoothed price density for the two groups.