An Evaluation of the Soda Tax with Multivariate Nonparametric Regressions

Eric J. Belasco
Assistant Professor
Texas Tech University
Department of Agricultural and Applied Economics
eric.belasco@ttu.edu

Sujit K. Ghosh
Professor
North Carolina State University
Department of Statistics

Benaissa Chidmi
Assistant Professor
Texas Tech University
Department of Agricultural and Applied Economics

Selected Paper prepared for presentation at the Agricultural & Applied Economics Associations 2010 AAEA, CAES & WAEA Joint Annual Meeting, Denver, Colorado, July 25-27, 2010.

An Evaluation of the Soda Tax with Multivariate Nonparametric Regressions

Abstract:

This research extends past work by Shonkwiler and Yen (1999) by allowing for distributional flexibility and nonlinear responses in the form of established semiparametric and nonparametric regressions. The proposed models are shown to outperform the parametric version typically used in demand analysis to characterize a system of censored equations in terms of model fit and prediction power. Using the developed models, we derive elasticities associated with different individual-specific scenarios with regard to the recently proposed "penny-an-ounce" tax on soft drinks sweetened with sugar.

Keywords: censoring, health taxes, nonparametric regressions

An Evaluation of the Soda Tax with Multivariate Nonparametric Regressions

The expanded use and availability of micro-level data sets has led to an increased demand for methods that model limited dependent variables accurately and efficiently. This is of particular interest in the area of consumer demand, where modeling systems of censored equations are often used. Further, the residuals are often highly correlated across equations which leads to an increased efficiency gain using a systems approach. The accurate identification of price elasticities are a crucial component to the evaluation of health policy aimed at adjusting prices in order to change human behavior regarding healthy or unhealthy consumption. The most common approach to dealing with a censored system of equations is a two-step approach originally described by Shonkwiler and Yen (1999) (henceforth SY) which uses Seemingly Unrelated Regressions (SURs) with an updating procedure to correct the standard errors. The largest limiting factor of this approach includes the need for an *a priori* assumption regarding the functional form of the relationship between the dependent variable and covariates as well as distributional identity.

For these reasons, this research expands the SY two-step approach in order to accommodate the use of semi- and nonparametric regressions that allow for functional form and distributional flexibility. Although, this method is aimed at consumer demand applications, it also may be used in applications such as modeling disease spread, animal growth dynamics, multivariate risks, and ecological measures. While the use of univariate models that account for censoring are well-developed, those concerned with dimensions of censoring larger than 3 are sparse. This research looks to add an important component to existing research by developing a model that deals with censoring of high-dimensions, but still allows for distributional and functional flexibility.

We assume a two-step process where the dependent variable, Y, is the product of a binary variable, W, and a positive valued value, V, such that $Y = WV$. SY assume W is derived from a probit model, while V is based on SURs. For semiparametric regressions, we assume W is derived from a single-index conditional probability based on Klein and Spady (1993) and V is based on the regression developed by Ichimura (1993). For nonparametric analogs, we assume a conditional probability based on Hall, Racine, and Li (2004), while the continuous regression is based on Racine and Li (2004). Conditional error terms from censored equations are updated based on that from noncensored outcomes within an observation. At the same time, observations without censoring are modeled based on a multivariate regression. In this way, the correlation between observations is explicitly modeled in a single step, which is different from the method used in SY.

In order to evaluate the relative merits of the developed model we use simulated and actual scanner-level consumer data to assess the out-of-sample predictive power and in-sample model fit, relative to the estimates derived based on SY. Predictive power will be assessed by randomly excluding data and computing the Mean Squared Prediction Error (MSPE) which evaluates the squared deviation between predicted and actual values, while model fit will be assessed using an appropriate R-square measure.

The data consist of weekly scans of carbonated beverages purchases for approximately 1,400 households during the years 2006, provided by Information Resources, Inc. for the BehaviorScan market Eau Claire, Wisconsin. The data is aggregated over one year and three categories of products: regular sugar-sweetened soda, diet soda (not sweetened with sugar), and club soda. The data is comprised of two components. The first component includes the quantities and expenditure for food and beverage purchases. The second component provides information on household demographics, such as income, age, family size, and education. The demand analysis will evaluate the relationship between individual purchases of sodas sweetened with sugar, diet soda, and club soda. The developed multivariate nonparametric model will be applied to this demand system and compared the SY method.

A demand evaluation into sugar-sweetened sodas is of particular importance at a time when policy makers and academic research are considering a tax on soda (Brownell and Frieden, 2009). Recent proposals have included a penny-an-ounce tax on sugar-sweetened beverages in health research (Brownell et al., 2009) and policy settings (New York Times, 2010), in an effort to curb consumption of sugar-sweetened sodas that have been linked to obesity. Further, this tax is further defended with the tax revenues that can be specifically targeted to fund programs aimed at fighting obesity and in order to pay for added social costs associated with consumption of beverages high in sugar with links to obesity.

This paper provides three distinct contributions to existing research. First, the development of a multivariate semi- and nonparametric regression model provides more flexible methods to evaluate many rich and disaggregated applied microeconomic datasets. Second, the developed model explicitly examines the commonly used linear assumptions regarding covariates, such as food prices, income, and age, with regarding to consumer demand. Without making any *a priori* assumptions, the more accurate functional relationship can be captured to inform food marketers and policy makers regarding price and income elasticities. Third, we evaluate the relationship between purchasing patterns of different beverages that have different implications for health and obesity studies. Results may provide results relevant to the development of policy where taxes and subsidies of beverages are considered, particularly in the case of sugar-sweetened soda. This research allows for an evaluation into the impact of a tax on sugar-sweetened beverages.

Traditionally, the main concern of censored demand studies has been to account for censoring by using maximum likelihood models to account for positive probability of observing zero consumption (see for example, Wales and Woodlan (1983); Lee and Pitt (1986); Chiang and Lee (1992); and Cornick, Cox, and Gould (1994)). For instance, SY, which is based on Heien and Wessells (1990), proposed a consistent two-step estimation procedure for system of demand equations. In the first step, the consumer's decision to consume the product is modeled as a dichotomous choice using a probit model. In a second step, a system of demand equations augmented by a selectivity regressor derived from probit estimates in the first step is estimated. A common feature in the two-step estimation is the use of a parametric estimation procedure that uses either maximum likelihood or Zellner (1962)'s seemingly unrelated regression (SUR).

The increased popularity of SY in demand analysis finds its roots in the ability to accommodate the zero consumption as well as its ease of implementation. In food demand, numerous studies have used SY framework to analyze censored demand. For instance, Yen (2005) extends SY model to multivariate sample selection model in the case of linear equations, while Yen and Lin (2006) extends the SY in the case of non linear equations and partial sample selection. Malaga, Pan, and Duch-Carvallo (2009) combine the two step estimation of SY with the nonlinear quadratic Almost Ideal Demand System (NQUAIDS) model of Banks, Blundell, and Lewbel (1997) to estimate meat demand in Mexico.

While the SY approach to estimating a system of demand equations provides a rather straightforward way of estimating a censored system of equations, it is not without its own strict assumptions. First, the binary and continuous components are both assumed to follow a specified parametric distribution. Consistent estimation of either component relies on the correct parametric assumption. To provide more flexibility regarding this restriction, a recent study by Sam and Zheng (2010) use a semiparametric approach characterize the binary component and assume a parametric form for the positive observations. Their approach is similar to the SY approach with the notable exception that the binary component is modeled according to the semiparametric approach of Klein and Spady (1993) as opposed to a probit. As pointed out in Cameron and Trivedi (2005), semiparametric models are often used in place of nonparametric methods because often the multiple dimensions of slicing used in nonparametric methods often allows too few data points for each slice. Further, single index models assume a linear index function in order to reduce the dimensionality associated with nonparametric estimation leading to computational advantages. However, as pointed out by Racine (2008), the curse-of-dimensionality is functionally related to the number of continuous variables and the number of values taken by those variables. Recent methods that include the use of categorical variables are not as prone to this issue, given the number of values is relatively small.

The second assumption made by the SY approach, as well as any other parametric approach, is regarding the functional relationship between the independent and dependent variables. The approach taken in this paper allows for the functional relationship to be determined through nonparametrically estimating the bandwidth or kernel, followed by a nonparametric regression. This allows for a more flexible functional relationship that the traditional linear function, which only reports an average estimate across the values of the independent variables. One hypothesis that is posited here is that the relationship between prices and quantities are not linear, which has its origins in past research. In fact, there is a growing empirical evidence that shows nonlinear relationship in the budget share equations. For instance, Banks, Blundell, and Lewbel (1997) extend the AIDS model to allow for quadratic logarithmic expenditure share and therefore, nonlinear relationship between prices and quantities.

The use of semiparametric methods to estimate a censored regression was first detailed in Powell (1984), who suggested the use of a censored least absolute deviation (CLAD) estimator. This estimator is based on the finding that censored observations can be characterized with a median regression model, leading to the use of an estimator that minimizes the absolute deviations between y_i and $max(X_i\beta, 0)$. Extensions of this model are documented in Pagan and Ullah (1999). A common theme from these models is that the conditional median is a linear function of X_i, such that $Med(y_i^*) = X_i\beta$. An exception to this is Lewbel and Linton (2002), who derive a nonparametric censored regression model without assuming the described linear relationship.

The remainder of this paper will progress as follows. Next, we develop a framework to model univariate and multivariate censored systems. This method, similar to the SY approach, will consist of a distinct binary and continuous component. The notable difference between these approaches will be that this approach will make minimal assumptions on the functional and parametric forms by using nonparametric techniques. Then, we apply this model to scanner-level data regarding carbonated beverage purchases. We also apply the SY method to this same data and compare model fit by way of R^2 and correct prediction of censoring. Additionally, one-third of the data are withheld in order to assess the ability of each model to predict. The final section, then estimates the predicted impact on quantity from the proposed tax using 4 different individual-specific scenarios and utilizing each derived estimator.

Methodology

This section begins by developing the univariate framework for estimation, which is then extended to the more general multivariate setting. A censored variable, Y, with a discontinuous distribution at $y = 0$ can be expressed as the product of two variables, $Y = VW$. First, the binary valued V measures the probability of a censored outcome. Then, a continuous variable, W, measures the positive valued outcome for noncensored observations. A recent study by Belasco and Ghosh (2008) assumed V to be modeled as a logistic CDF, while W was assumed to generated according to a lognormal distribution. This paper makes the notable distinction, in the univariate case, of not making any parametric assumptions regarding the binary or continuous components.

In deriving the nonparametric censored regression model, we begin with

$$\begin{aligned} Y_i &= W_i * (g(X_i, \beta) + \varepsilon_i) & (1) \\ W_i &= I\,(h(Z_i, \gamma) + \nu_i > 0) & (2) \end{aligned}$$

where $I(.)$ is an indicator function, $\varepsilon_i \sim iid(0, \sigma_\varepsilon^2)$ and $\nu_i \sim iid(0, \sigma_\nu^2)$. Notice that a parametric analog to this specification might include $h(Z_i, \gamma)$ being distributed according to a probit model, such that $Pr(W_i = 1|Z_i) = \Phi(Z_i\gamma)$ as in Belasco and Ghosh (2008). Additionally, $g(X_i, \beta)$ might be distributed according to a lognormal distribution, such that $log(V_i) \sim N\left(X_i\beta, \sigma_\varepsilon^2\right)$. However, each of these assumptions makes a bold assumption that the parametric distribution is known to the researcher.

In order to develop a tractable method to evaluate a system of equations, while preserving cross-equation correlation, we use the following method. First, it is important to note that the joint distribution of k random variables $Y_1, Y_2, ..., Y_k$ can be characterized as

$$f(Y_1, Y_2, ...Y_k) = f(Y_1|Y_2, ..., Y_k) * f(Y_2|Y_3, ..., Y_k) * \cdots * f(Y_k). \qquad (3)$$

In the case where each dependent variable is fully continuous, incorporating this into a system of single equations could be written as

$$\begin{aligned} Y_{ij} &= g_j(\beta_j, X_i, Y_{i,j+1}, ..., Y_{ik}) + \varepsilon_{ij} \qquad \forall j < k & (4) \\ &= g_j(\beta_j, X_i) + \varepsilon_{ij} \qquad for j = k & (5) \end{aligned}$$

where $g_1(.), g_2(.), ..., g_k(.)$ are unknown distributions for individuals $i = 1, 2, ..., n$ and equations $j = 1, 2, ..., k$. This system is made more complex given the two components in each equation. The above equations are thus slightly modified in the following way

$$\begin{aligned} Y_{ij} &= W_{ij} * \left(g_j(\beta_j, X_i, Y_{i,j+1}, ..., Y_{ik}) + \varepsilon_{ij}\right) \qquad \forall j < k & (6) \\ &= W_{ij} * \left(g_j(\beta_j, X_i) + \varepsilon_{ij}\right) \qquad for j = k & (7) \end{aligned}$$

such that

$$\begin{aligned} W_{ij} &= I\left(h_j(\gamma_j, Z_i, W_{i,j+1}, ..., W_{ik}) + \nu_{ij}\right) \qquad \forall j < k & (8) \\ &= I\left(h_j(\gamma_j, Z_i) + \nu_{ij}\right) \qquad for j = k & (9) \end{aligned}$$

While nonparametric methods can be quite cumbersome in some applications, due to the well known “curse of dimensionality,” single-index models circumvent this issue by assuming that $E(y|x) = h(Z'\gamma)$, where h is an unknown link function that assumes a linear index relationship with the dependent variable. This allows for a reduction of the dimensionality that plagues nonparametric methods (Sam and Zheng, 2010).

Klein and Spady (1993) suggest the maximization of the following log-likelihood function

$$LL(\gamma, h) = \sum_{i=1}^{n} \left(w_i log(h(Z_i'\gamma)) + (1 - w_i) log(1 - h(Z_i'\gamma))\right) \qquad (10)$$

This method is essentially the same as minimizing the sum of squared errors for

$$S_n(\gamma,h) = \sum_{i=1}^{n} \left(y_i - h(Z_i'\beta)\right)^2 . \tag{11}$$

If h were known, this would be relatively straight-forward to estimate using well-established least squares methods. Ichimura (1993) proposed the use of the leave-one-out estimator, also known as the Nadaraya-Watson (or local constant) estimator. The size of the subsets to "leave-out" largely depends on the selection of the bandwidth. In this study, we use the cross-validation approach explained in more detail by Racine (2008). The advantage is that this process is data-driven, which largely fits with the parametric flexibility of semi- and nonparametric modeling strategies.

In order to utilize nonparametric methods, we assign the binary component, $g(X_i,\beta)$ to be computed based on the approach by Hall, Racine, and Li (2004) who use a cross-validation approach to select the smoothing parameters. In any nonparametric regression, estimates are obtained by cutting the data into slices, then estimating the local behavior within a slice. The size of the slices are computed using a kernel density estimator. The main advantage to the cross-validation employed in this setting is that the method simultaneously determines the smoothing parameters and identifies irrelevant variables in X.

In order to characterize the continuous portion of the Y distribution, we use the method developed by Racine and Li (2004) who again use a cross-validation approach that is able to handle interactions between categorical and continuous variables in a natural manner. The main advantage to this approach is that it allows data points to determined dependencies and interactions between mixed data types.[1]

[1] While the approach by Lewbel and Linton (2002) is very flexible and implementable, it assumes all variables are continuous and therefore has obvious limitations regarding implementation in applied research.

Empirical Application

In this section, we apply the previously developed methods to scanner-level consumption data for carbonated beverages. The estimation of demand elasticities are commonly used to assess the impact of a tax. In this application, health officials have proposed the penny-an-ounce tax on sugar sweetened beverages. This implies that the tax itself is proportional to size and not the amount of sugar in the beverage. Brownell et al. (2009) point out three favorable outcomes that may come from the tax. First, households consume less sugar-sweetened drinks. Past demand studies have estimated this elasticity to be between -0.8 to -1.0, meaning that an increase to prices by 10% results in an 8-10% decrease in the amount consumed. To put this into perspective, a 20-oz soft drink that costs $1.00 would have an increased cost of 20% with the new tax. Based on the stated elasticities, the predicted impact would be large. However, these elasticities often are computed in ways that do not recognize the nonlinear demand relationship.

The second advantage of the tax would be to shift consumption toward drinks not sweetened by sugar, which so far have no demonstrated negative health effects. Third, the extra tax revenue can be used to internalize the social cost associated with consuming goods that are not healthy and lead to negative health outcomes and higher health expenditures for all individuals. While, we do not examine the third advantage here, we do compare the impact of a tax on sugar-sweetened soda on itself as well as non-sugar-sweetened soda to evaluate the own and cross price elasticities. Further, we demonstrate that these elasticities are more accurately determined when they account for nonlinearities and allow for parametric flexibility.

The next sections describes the data we use, which is followed by the reported results of each estimation method. These three methods (parametric, semi-parametric, and non-parametric) are compared in terms of goodness-of-fit tests and predictive power tests. This section then concludes with a discussion regarding the relative marginal impacts resulting from each model.

Data

The model described above is estimated using scanner-level data provided by Information Resource Inc., for the year 2006. Weekly data of carbonated beverages were aggregated to give the annual consumption of 1,374 households in Eau Claire, Wisconsin for three beverage categories: regular soft drinks, low calorie soft drinks, and club soda. The data provide information on quantities and expenditure of these three categories as well as socio-demographic information about households. Table 1 presents summary statistics of the data used in the analysis. During 2006, Out of 1374 households, 1113 households(81%) bought regular soft drinks, 956 households (70%) bought low calorie soft drinks, and 296 households (21%) bought club soda.

Table 1: Summary Statistics and Variable Definitions (N=1,374)

Variable	Mean	Std. Dev.
Quantities (gallons per year)		
Regular Soda (Consuming households: 81.0%)	6.78	10.62
Diet Soda (Consuming households: 69.6%)	6.59	10.04
Club Soda (Consuming households: 21.5%)	2.37	4.76
Expenditures ($ per year)		
Regular Soda (Consuming households: 81.0%)	25.28	45.01
Diet Soda (Consuming households: 69.6%)	26.23	41.52
Club Soda (Consuming households: 21.5%)	10.05	17.91
Prices ($ per gallon)		
Regular Soda	9.23	8.15
Diet Soda	8.63	5.25
Club Soda	10.01	6.38
Wage Earners	1.07	0.83
Family Size	2.59	1.24
Proportion of Sample		
Inc1 (HH Income < 20k)	0.15	
Inc2 (HH Income 20k to 65k)	0.57	
Inc3 (HH Income > 65k)	0.29	

Quantities were converted from a volume equivalent measure (per 192 ounces) into a gallon equivalent measure (per 128 ounces) in order to allow for meaningful interpretations. Based on the data described above, the average amount of regular and diet soft drinks consumed at the household level was equal to 6.78 and 6.59 gallons, respectively when we include only the households that consumed a positive amount of that drink. This amount is equivalent to just over 70 cans of soda or just under 13 two-liter bottles of soda. This amount is substantially less for the fewer households that consume club soda.

Prices were established by dividing the total expenditures by the gallon equivalent measure. For households that do not purchase anything within a particular category, the prices faced by consumers are latent. These missing prices are augmented to the data assuming the log of prices are distributed as a lognormal distribution. Given this specification, we simulate from a lognormal distribution with the mean and standard deviation equal to the sample estimates for the log of the observed prices. This method guarantees positive prices and preserved the sample characteristics of the observed variables when the latent observations are included.

This data also contains demographic factors associated with each household. Income and family size appear to be the most important variables in this set of equations. Other variables, such as race, education, and marital status were originally included but found to be insignificant. Income is comprised of two parts since the income variable is categorical. First, pre-tax head of household income (*HH Income*) is placed into one of the three categories listed above. Further, the number of wage earners (*Wage Earners*) is also included in order to control for families with dual-incomes. Wage earners is a count of the number of individuals in the household who work (full or part-time) for wages. This also controls for the habits of retired individuals, living on a fixed income who report no head of household income.

An obvious attraction to the nonparametric and semiparametric forms given this setting is the interaction and nonlinear relationship between some of these variables with any consumptive good. For example, in a parametric model, we would need to specify these interactions and add nonlinear parameters to allow for that response. However, in order to allow for a nonlinear relationship to be determined by the data, we would need an *a prior* assumption regarding the functional form before we could evaluate the relevancy of our assumption. The methods used in this study allow us to make no assumption about the function form or interaction between variables, but we capture the movements that are characterized in the data. This allow for a richer characterization of the data.

Estimation Results

As previously described, estimation is conducted on two components of the censored distribution. First, we focus on the binary component, which determines the likelihood of censoring, which is followed by estimation of the continuous component.

In order to compare the ability of each model to characterize the in-sample data, we use two metrics for binary outcome models. First, we compute the overall percentage predictions that correctly predict the binary outcome. More specifically, if the $Pr(\hat{W}_{ij}|Z_i) > 0.5$ corresponds with $W_{ij} = 1$, or $Pr(\hat{W}_{ij}|Z_i) < 0.5$ corresponds with $W_{ij} = 0$, then the observation is correctly predicted. While this does provide some insight into how well the model fits the data, it isn't a complete picture, which is why R-squared is also included. As described in Hayfield and Racine (2008), comparing R-squared for parametric and non-parametric models can be done with the following:

$$R^2 = \frac{[\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} \tag{12}$$

Because this measure is bounded by $[0,1]$ and it is exactly the same as the standard R-squared measure used in OLS estimation, it is used here to compare the in-sample model fit and accounts for the distance between predicted and actual values.

We also asses the predictive power of each model concerning a randomly determined out-of-sample portion of the data. For this analysis we evaluate the percentage that are correctly predicted, as well as commonly used Root Mean Squared Error (RMSE), which is computed as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

The results for each model are shown below in Table 2. The in-sample results demonstrate the semi-

and nonparametric models ability to improve the model fit of binary component both in terms of higher prediction power and R-squared. This is not surprising given the smoothing techniques that are used in order to better capture movements that vary across the range of variables.

Table 2: Measures of Model Fit and Predictive Power for Binary Component

	Regular Soda			Diet Soda			Club Soda		
	Probit	KS	NP	Probit	KS	NP	Probit	KS	NP
In-Sample Fit									
% Correctly Predicted	0.798	0.802	0.805	0.695	0.729	0.733	0.780	0.780	0.806
R2	0.034	0.065	0.729	0.065	0.127	0.448	0.041	0.029	0.142
Out-of-Sample Prediction									
% Correctly Predicted	0.818	0.820	0.993	0.712	0.714	0.947	0.783	0.783	0.021
RMSE	0.374	0.383	0.220	0.440	0.460	0.345	0.412	0.409	0.786

However, in the binary component, this model fit is a trade-off with weakened prediction power for out-of-sample observations. This is a significant finding given that the hypothesized predictive power between parametric and semi- or nonparametric models was ambiguous. This is because the more flexible methods allow for a better characterization of the data that allow for nonlinear and interactive impacts to be more accurately characterized, which leads to improved prediction power. However, this improvement comes at a cost, in the form of overspecification of the model. For example, if bandwidths are computed to be tighter than they should be, the researcher experiences an increased ability to fit the model and decreased ability to fit out-of-sample data. For this reason, the selection of bandwidth is crucial for the performance in optimizing both component. However, bandwidth is not selected based on prediction power, it is usually determined based on in-sample fit.

After estimating the binary component, the continuous portion of the distribution is estimated using the proposed methods. Table 3 reports the associated in-sample R-squared measures associated with positive values and all values, respectively.

Since only positive values are used to estimate the semi- and nonparametric estimators, it is not surprising that these estimators fit the positive values substantially better. It is also notable that when we include censored values, the prediction is also improved over parametric methods when

Table 3: R-Squared Associated with In-sample data

	Regular Soda	Diet Soda	Club Soda
In-Sample (Pos. values only)			
Par	0.072	0.101	0.137
Semi	0.377	0.257	0.914
NP	0.225	0.384	0.868
In-Sample (All Obs)			
Par	0.056	0.021	0.013
Semi	0.278	0.144	0.047
NP	0.198	0.213	0.001

the product of both components are incorporated. Semi-parametric methods appear to fit the data particularly well. This finding is interesting given the fact that single-index models are often used for computational efficiency, but can also be thought of as a compromise between parametric and nonparametric models. This added structure with flexibility appear to perform well within-sample and out-of-sample.

This out-of-sample fit is demonstrated in table 4 where the omitted final third of the data is evaluated in terms of prediction power. The results demonstrate ambiguous results in the sense that each method is found to be the best predictor of only one beverage category. As explained above, if nonparametric models simply over-specify the data and do not add a more vivid picture of the relationship, then prediction will suffer relative to parametric models. Given that nonparametric and semiparametric models do not under-perform parametric version, and that the parametric version is clearly out-performed by the more flexible models, nonparametric and semiparametric models are shown to outperform, overall, the parametric counterpart.

Table 4: RMSE Associated with Out-of-sample data

	Regular Soda	Diet Soda	Club Soda
In-Sample (Pos. values only)			
Par	8.035	6.986	1.611
Semi	8.478	7.157	1.386
NP	7.985	8.588	4.988

Estimated Impacts from Tax

The average amount of annual household consumption of sugar-sweetened soft drinks during 2006 in our data is 698.6 ounces (or 5.5 gallons[2]) per year, where 19% of the households did not record any purchases of sugar-sweetened soft drinks. With the proposed tax, this amounts to an average tax of $6.99 per household per year. Within our data, the 90th percentile of sugar-sweetened soft drink consumers, purchased almost three times as much as the average household with 1,769.3 ounces (or 13.8 gallons), leading to an annual tax of $17.69.

With the average price per ounce for regular soda at 0.0319, a one-cent tax per ounce is set at a rate of 31.37%. The important question here is what impact would such a tax have on consumption and to what extent to these impacts change for different individuals and price levels. Table 5 shows the amount of tax that is based on volume, and the active percentage rate changes based on the price.

Table 5: Structure of the Proposed Penny-An-Ounce Tax

Price		Tax		Tax /
Per Gallon	Per Ounce	Per Gallon	Per Ounce	Price
2.000	0.016	1.280	0.010	0.640
3.000	0.023	1.280	0.010	0.427
4.000	0.031	1.280	0.010	0.320
5.000	0.039	1.280	0.010	0.256
6.000	0.047	1.280	0.010	0.213
7.000	0.055	1.280	0.010	0.183

As with any tax that is tied to volume, the tax as a percentage of price decreases for higher priced regular sodas.[3] While parametric models with fixed estimates have a long history in econometrics, their ability to identify heterogeneity within a population is limited. For example, one hypothesis posited in this research is that the marginal response to increased prices is different based on demographic factors as well as the level of prices.

[2] Assumes 1 gallon = 128 ounces

[3] For example, a 20-oz regular soda that is $1.00, the per gallon price is $6.34, while the per ounce price is $.05. With the tax, the new price would be $1.20, a 20% increase.

While past studies have evaluated the elasticity associated with non-alcoholic consumption (Yen et al., 2004; Zheng and Kaiser, 2008), this paper compares parametric marginal impacts with that of nonparametric and semiparametric versions of demand. The developed model is hypothesized to outperform parametric estimators due to its less rigid assumptions that allow for more flexibility with regard to distributional assumptions and nonlinear marginal effects, both of which are identified without *a priori* assumptions. However, these gains typically come at a cost in terms of using fewer observations at each point to derive a marginal impact.

Table 6: Marginal Impacts on different consumer types

	Scenarios			
Variable	A	B	C	D
Family Size	5	2	3	2
Income	2	2	1	2
Wage Earners	2	2	1	0
n	28	75	11	171
med(P_1)	2.66	3.01	3.91	3.22
Post-Tax P_1	3.94	4.29	5.19	4.5
P_1 Increase (%)	48.12	42.53	32.74	39.75
med(P_2)	2.63	3.11	2.82	3.37
med(P_3)	3.75	3.75	2.59	3.75
med(Q_1)	5.59	2.9	3.375	2.25
med(Q_2)	1.92	2.18	2.25	2.25
med(Q_3)	0.53	0.53	0.40	1.00
Marginal Increase in Q_1 (%)				
Par	-2.56	-4.19	-4.06	-5.48
SP	0.13	35.14	-95.44	155.64
NP	4.78	-3.03	-5.36	-3.78
Marginal Increase in Q_2 (%)				
Par	-15.01	-12.25	-11.02	-11.29
SP	51.20	80.32	-52.49	44.76
NP	-34.84	-14.95	-5.38	-41.60
Marginal Increase in Q_3 (%)				
Par	0.93	2.41	0.88	0.47
SP	1.32	-180.00	-36.25	3.70
NP	-8.11	-9.43	105.75	-12.80

Note: P_1, P_2, P_3,Q_1, Q_2, Q_3 are prices and quantities associated with regular soda, diet soda, and club soda, respectively.

To illustrate, we compare four different scenarios with conditioning variables as shown in Table 6. Scenario A is a large family with middle-class income and two wage earners; scenario B is a middle class married couple, both wage earners, without kids; scenario C is a low-income family of 3 with only one wage earner; and scenario D is a middle-class retired couple.

Rather than compute elasticities at the mean of all variables, we take the median price and volume associated with each group from the data. For example, 28 observations fit the characteristics in Scenario A, of which the median prices and quantities are based on. This is intended to more accurately capture the types of choices made by that group of individuals. For example, individuals in scenario A tend to pay lower prices for all beverages, although tend to consume in higher quantities.

At each point, we assume the tax is imposed and increases that median price by $1.28 per gallon, after taxes. With different prices in each scenario, this assumes a different percentage increase in prices that range from 39.75% to 48.12%. Because scenario A has the lowest median price, the percentage increase is largest on that group.

While semiparametric methods appear to strike a good balance between fitting and predicting the data, the marginal impacts are not in line with economic intuition as the tax leads to increased consumption of regular soda in scenarios A, B, and D. Further, the volatility with which the elasticities are estimated has a large range of values. Signs also appear unchanged in other commodities, which does not match with the expectations of Brownell et al. (2009), where regular soda drinkers were thought to switch toward untaxed diet soda alternatives.

Nonparametric results appear to be more in line with expectations in the sense that most show a decreased consumption of regular soda from the tax, with the exception of scenario A. Further, diet soda and club soda consumption is also expected to decrease by marginal amounts. Based on these results, groups would substitute to other commodities not included in this study.

Parametric results imply more theoretically consistent results, although still far from the expectations of Brownell et al. (2009). For example, a 48% increase in price, for scenario A, leads to a minor 3% decrease in consumption. Larger consumption decreases are found in diet soda where the two are shown to be complementary products. The impact on club soda drinkers is not statistically significant.

These results shed some light on a few shortcomings of this research that are areas that need to be addressed. First, other alternatives such as juices and energy drinks are likely substitutes for soda and should be included a study of this nature in order to more accurately assess the impact of a soda tax on this market. Second, while we use a fairly large data set, nonparametric and semiparametric estimators with categorical variables are limited in power by the amount of individuals within each set of covariates. One suggestion would be to include more years to this study in order to observe more variability in prices and more observations within each set of categorical variables.

Given the well-documented issue of the curse in dimensionality for data with a large amount of data, we employ a method that is equivalent to the method described in Racine (1993). This method is based on the finding that scaling factors are independent of sample size, meaning computing bandwidth for subsets of the data recursively is equivalent to computing the optimal bandwidth of the entire sample.This method saves an incredible amount of time given that computation time can often increase exponentially in nonparametric method when large data sets are used.

Conclusion

This study focused on developing a framework to extend the SY method by incorporating distribution flexibility, in the form of semiparametric and nonparametric, into characterizing a system of censored equations. Model fit and predictive power tests demonstrate the ability of the developed models to outperform parametric counterparts. At the same time, this study identifies some potential shortcomings of the more flexible methods in their inability to estimate marginal outcomes at a point. While nonparametric methods have the potential to more accurately characterize heterogeneous effects, this identification is limited to sufficient data within each set independent variable pairings. Future work in this area includes augmenting existing data in order to include new products that are within this market segment, such as juice and energy drinks, as well as, include more years of data in order to more confidently estimate at any particular point.

References

Banks, J., R. Blundell, and A. Lewbel. 1997. "Quadratic Engel Curves and Consumer Demand." *Review of Economics and Statistics* 79:527–539.

Belasco, E., and S. Ghosh. 2008. "Modeling Semi-continuous Data Using Mixture Regression Models with an Application to Cattle Production Yields." Unpublished, Selected Paper. American Agricultural Economics Association Annual Meeting, Orlando, FL.

Brownell, K., T. Farley, W. Willett, B. Popkin, F. Chaloupka, J. Thompson, and D. Ludwid. 2009. "The Public Health and Economic Benefits of Taxing Sugar-Sweetened Beverages." *The New England Journal of Medicine* 361:1599–1605.

Brownell, K., and T. Frieden. 2009. "Ounces of Prevention - The Public Policy Case for Taxes on Sugared Beverages." *The New England Journal of Medicine* 360:1805–1808.

Cameron, A.C., and P.K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.

Chiang, J., and L. Lee. 1992. "Discrete/Continuous Models of Consumer Demand with Binding Nonnegativity Constraints." *Journal of Econometrics* 54:79–93.

Cornick, J., T.L. Cox, and B.W. Gould. 1994. "Fluid Milk Purchase: A Multivariate Tobit Analysis." *American Journal of Agricultural Economics* 76:74–82.

Hall, P., J. Racine, and Q. Li. 2004. "Cross-validation and the Estimation of Conditional Probability Densities." *Journal of the American Statistical Association* 99:1015–1026.

Hayfield, R., and J. Racine. 2008. "Nonparametric Econometrics: The np Package." *Journal of Statistical Software* 27:1–32.

Heien, D., and C. Wessells. 1990. "Demand Systems Estimation with Microdata: A Censored Regression Approach." *Journal of Business and Economic Statistics* 8:365–371.

Ichimura, H. 1993. “Semiparametric least squares (SLS) and weighted SLS estimation of single-index models.” *Journal of Econometrics* 58:71–120.

Klein, R., and R. Spady. 1993. “An Efficient Semiparametric Estimator for Binary Response Models.” *Econometrica* 61:387–421.

Lee, L., and M. Pitt. 1986. “Microeconometric Demand Systems with Binding Non Negativity Constraints: The Dual Approach.” *Econemetrica* 54:1237–1242.

Lewbel, A., and O. Linton. 2002. “Nonparametric Censored and Truncated Regression.” *Econometrica* 70:765–779.

Malaga, J., S. Pan, and T. Duch-Carvallo. 2009. “Did Mexican demand Change under NAFTA?” Contributed Paper prepared for presentation at the International Association of Agricultural Economists Conference, Beijing, China, August 16-22, 2009.

New York Times. 2010. “New York Health Official Champions the Soda Tax.” `http://www.nytimes.com/2010/04/05/health/policy/05daines.html`, April 4.

Pagan, A., and A. Ullah. 1999. *Nonparametric Econometrics*. Cambridge University Press.

Powell, J. 1984. “Least Absolute Deviations Estimation for Censored Regression Model.” *Journal of Econometrics* 25:303–325.

Racine, J. 1993. “An Efficient Cross-Validation Algorithm for Window Width Selection for Nonparametric Kernel Regression.” *Communications in Statistics - Simulation and Computation* 22:1107–1114.

—. 2008. “Nonparametric Econometrics: A Primer.” *Foundations and Trends in Econometrics* 3:1–88.

Racine, J., and Q. Li. 2004. “Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data.” *Journal of Econometrics* 119:99–130.

Sam, A., and Y. Zheng. 2010. “Semiparametric Estimation of Consumer Demand Systems with Micro Data.” *American Journal of Agricultural Economics* 92:246–257.

Shonkwiler, J.S., and S.T. Yen. 1999. “Two-Step Estimation of a Censored System of Equations.” *American Journal of Agricultural Economics* 81:972–982.

Wales, T., and A. Woodlan. 1983. “Estimation of Consumer Demand Systems with Binding Non-negativity Constraints.” *Journal of Econometrics* 21:263–285.

Yen, S. 2005. “A Multivariate Sample Selection Model: Estimating Cigarette and Alcohol Demands with Zero Observations.” *American Journal of Agricultural Economics* 87:453–466.

Yen, S., and B. Lin. 2006. “A Sample Selection Approach to Censored Demand Systems.” *American Journal of Agricultural Economics* 88:742–749.

Yen, S., B. Lin, D. Smallwood, and M. Andrews. 2004. “Demand for Nonalcoholic Beverages: The Case of Low-Income Households.” *Agribusiness* 20:309–321.

Zellner, A. 1962. “An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias.” *Journal of the American Statistical Association* 57:348–367.

Zheng, Y., and H. Kaiser. 2008. “Advertising and U.S. Nonalcoholic Beverage Demand.” *Agricultural and Resource Economics Review* 37:147–159.