Assessing the Utility of Top-Coded Consumer Expenditure Survey Data October 2015

Daniel K. Yang* Daniell Toth†‡

Abstract

The Consumer Expenditure Survey implements a statistical disclosure limitation process known as "top-coding" in the public used microdata release to conceal sensitive and identifiable information in order to protect the households confidentiality. This process replaces, for example, the high (low) end households annual income by the average of all high (low) end households annual income in the microdata for public users. Top-coding can numerically affect the utility of the microdata, especially for analyses that are sensitive to the high (low) end of the distribution. For instance, parameter estimates and confidence intervals can both be biased by this process. In this study, we investigate the impact of top-coding on CE microdata utility for multiple regression and logistic regression models used to analyze the relationship between certain expenditures and household income after adjusting demographic characteristics. We employ a multiple integration approach to estimate the empirical cumulative distribution function (ECDF) and an Anderson–Darling distance (A-DD) measurement to investigate the effects of top-coding on the utility of the CE microdata. We then evaluate A-DD under the background of a two-stage economics model and explore a robust logistic regression method on the propensity of expenditure reporting to offset the impacts of top-coding.

Key Words: confidentiality, disclosure limitation, utility measures, empirical CDF, survey data, top-coding

1. Introduction

An essential mission of Bureau of Labor Statistics (BLS) is to collect and disseminate public data on labor market activity, working conditions, and price changes. When releasing microdata to the public, BLS typically alters or withholds some of the original data to protect the confidentiality of respondents' identities or other sensitive data. However, this withholding or alterations may negatively impact the utility of the released data. With every method used to protect the private data of each respondent, there is a trade off between risk and utility: more security, less utility. Statistical agencies attempt to strike a balance to achieve adequate protection while still providing useful data to the public.

The Consumer Expenditure Survey (CE) aims to collect and publish data on the spending activities as well as family income, and other demographic and social-economic characteristics of U.S. families and single consumers. One way this data is collected is through the Quarterly Interview Survey. Microdata obtained from this panel survey is provided to the public annually. In the microdata release, CE implements a statistical disclosure limitation (SDL) method called top-coding to mask respondents' identifiable and sensitive information.

While one of the most important uses of the CE data is to regularly revise the Consumer Price Index market basket of goods and services and their relative importance, there are many other important uses of the publicly released microdata. This is the only national

*U.S. Bureau of Labor Statistics Office of Survey Methods Research, 2 Massachusetts Avenue Suite 1950, NE Washington, DC 20212

†U.S. Bureau of Labor Statistics Office of Survey Methods Research, 2 Massachusetts Avenue Suite 1950, NE Washington, DC 20212

‡Disclaimer: Any opinions expressed in this paper are those of the author(s) and do not constitute policy of the Bureau of Labor Statistics.

survey to cover the full spectrum of consumer's spending, household income, demographic and social-economic characteristics. As such, it is heavily relied on by economic policymakers examining the impact of policy changes on economic groups, by other Federal agencies, such as the Bureau of Economic Analysis for benchmarking annual growth rates and the Census Bureau as the source of thresholds for the Supplemental Poverty Measure, as well as businesses and academic researchers studying consumers' spending habits and trends. Regardless of the type of analysis, a variety of statistical models have been applied to CE data to meet the needs of each individual data user (Yang and Gonzalez 2013).

Given the important role of CE in the academic and research areas, it is imperative for the program office to periodically assess the utility of the publicly released microdata. There are generally two approaches to evaluating the utility of data affected by a SDL method (Woo, Reiter, Oganian and Karr 2009). The first approach is to use an analysis specific measure which requires knowledge of how the data will be used in analysis. For example, one can compare regression results from the original data with results achieved using the data set after the SDL method has been applied. Karr, Kohnen, Oganian, Reiter and Sanil (2006) proposed measuring the overlap of confidence intervals for model parameters obtained using the original and those using the protected data, where greater overlap indicates higher utility.

The second approach is to use a global measure which requires knowledge of how the data are distributed. Three methods were proposed for global measure approach: propensity scores, cluster analysis, and empirical cumulative distribution function (CDF), by Woo, Reiter, Oganian and Karr (2009). The propensity scores measure is the average squared deviance between the ratio of estimated propensity scores of a unit being altered and the percentage of SDL altered units over the combined data set of the original and SDL altered data sets. However, this measure is not robust with respect to the model specification. The cluster analysis first classifies the combined data set of the original and the altered data sets into a predetermined number of groups, then it computes the average squared differences of the within-cluster ratio of the size of the original data over the size of the SDL altered data minus the overall ratio of the size of the original data over the size of the altered data. The empirical CDF method computes the Kolmorgorov-Smirnov statistic (maximum absolute difference) and average squared difference between the original data empirical CDF and the altered data empirical CDF. The Kullback-Liebler divergence between the empirical distributions of the original data set and the SDL altered data set had also been introduced as a global measure (Karr, Kohnen, Oganian, Reiter and Sanil 2006), however, its reliance on the multivariate normal assumption makes it unattractive for implementation. Yang and Toth (2014) conducted a bootstrap re-sampling study to evaluate the effects of top-coding on the utility of the CE microdata and found a data utility measurement based on a modified form of Kullback-Liebler divergence to provided useful indication.

In this article, we propose using multiple integration to estimate the empirical CDF (ECDF) of the original and the altered data then to obtain an Anderson–Darling Distance (A-DD) measurement between those ECDFs. We investigate A-DD in terms of a two-stage economics model scenario. We also implement a robust logistic regression approach for more stable coefficient estimates of expenditure reporting propensity. The organization of the paper is the following: Section 2 introduces the CE top-coding process and illustrates its impacts in distribution. In Section 3, we describe the CE data sets and introduce the two-stage Cragg model, ECDF and A-DD. Section 4 contains the results of our analysis on CE data and Section 5 draws conclusions from the analysis.

2. CE Top-coding Process

The release of CE Survey microdata requires use of an SDL to conceal any sensitive and personally identifiable information (PII) in order to protect the household's confidentiality and anonymity. Though the CE collects data from an anonymous sample of the population, some consumer units have characteristics so far outside the norm, such as a very high income or unusual expenses (e.g. extremely high utility bills), that release of this information would make identification possible. In order to conceal any identifiable characteristics, CE implements a SDL process called top-coding before releasing the microdata.

The idea behind top-coding is to replace all values that are above the top or below the bottom $\alpha\%$ with the average of all values above or below this threshold. This SDL only affects outliers and is guaranteed to provide accurate means (first moment estimates). Suppose variable y_i is top-coded for all $y_i > y_{1-\alpha}$, where α is a percentile level, e.g. $\alpha = 0.05$. Then

$$\bar{y} = \frac{1}{N}\sum_i y_i = \frac{1}{N'_A + N_A}\left(\sum_{j\in A'} y_i + \sum_{i\in A} y_i\right) = \frac{1}{N}\left(\sum_{j\in A'} y_i + N_A\bar{y}_A\right) = \tilde{\bar{y}},$$

where $\bar{y}_A = \frac{1}{N_A}\sum_{i\in A} y_i$, $A = \{y_i | y_i > y_{1-\alpha}\}$, $\bar{y}$–confidential mean, $\tilde{\bar{y}}$–top-coded mean (public released). Therefore, the top-coded mean, $\tilde{\bar{y}}$, is the same as the confidential mean, $\bar{y}$.

Despite the guarantee of accurate first order estimates, this process can still have a negative impact on data quality by distorting higher order associations. Below is an example of CE 2011 household income sampling distribution (Figure 1).

In Figure 1, we can see that the confidential household income distribution is wider than the top-coded one and that very high income households are scattered to the right. The question becomes: what impact will this distortion of the distribution have on estimates involving higher order moments, such as regression coefficients?

3. Methodology

3.1 CE Quarterly Interview Data

In our study, we considered four different years (2008 to 2011) of CE household interview data and their corresponding publicly released microdata. We chose four expenditure variables: property taxes, utilities, health care and domestic services as examples of variables often used by economists when analyzing CE data. Those four expenditures also provide us with three different types of association with household income: 1) highly correlated with income and highly top-coded (property taxes), 2) not highly correlated with income but highly top-coded (utilities, health care), and 3) highly correlated with income but not highly top-coded (domestic services).

Beside household income, the following covariates were also used in the analysis of the expenditures: housing tenure (owner or not); geographical region (Northeast, Midwest, South, West); number of members in the household; number of persons over 64 in the household; number of members under age 2 in the household; reference person's age, ethnicity, education attainment, and gender (Male, Female).

3.2 Two-stage Economics Cragg Model

Cragg (1971) introduced an elemental scenario of economics phenomenon: consumers buy a specific type product or service in a specific interval of time. If we consider the

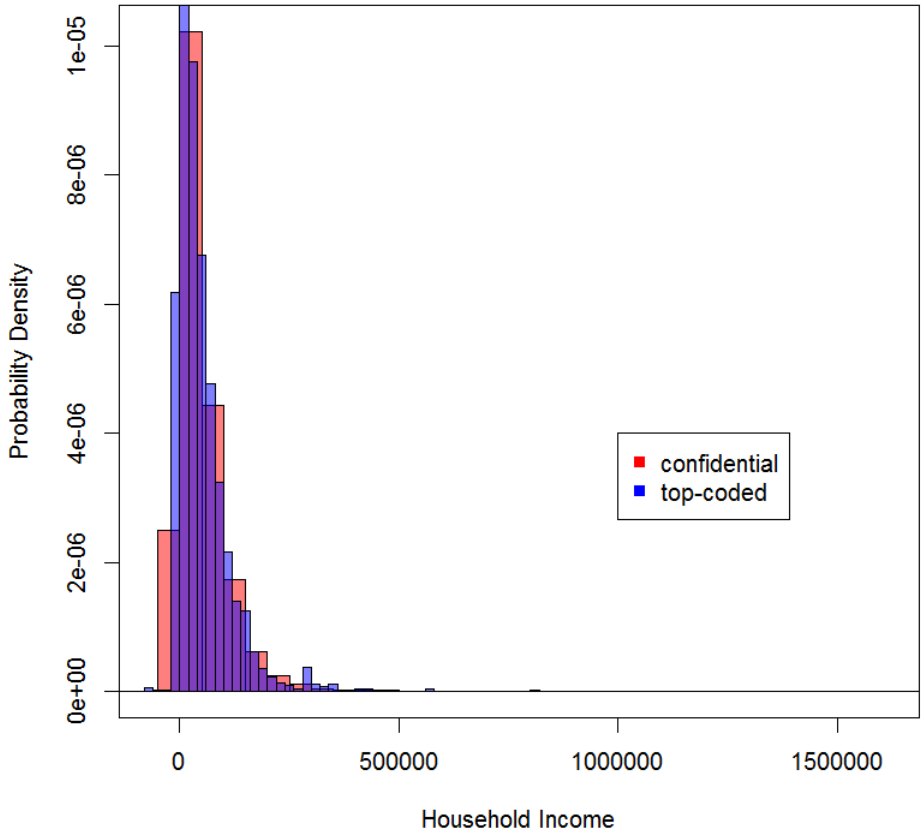

Figure 1: CE 2011 Household Income Sampling Distribution: Confidential vs. Top-coded

purchase as an event and it did take place, then a non-zero expenditure (assuming from a continuous, positive random variable) is observed. If not, then a zero expenditure is occurred. The CE Quarterly Interview data collection is a perfect example of Cragg (1971) scenario. Now we have two measurements: one is the propensity of non-zero expenditure reporting (a.k.a consumption), the other is the reported non-zero expenditure. In addition, the natural association between expenditure and income requires us to bring the non-zero income into the model. In the economics literature, the natural log of reported expenditures and non-zero income had been commonly used in economics model because the original distributions were closer to log-normal (Dippo 1984), right skewed with heavy tail (Omori 2010) or positively skewed (Family Spending 2009, Chiswick 1974, Pritchett 2010).

Cragg (1971) proposed a two-stage model to accommodate the economics scenario by estimating the propensity scores to consume and modeling non-zero expenditures. Here, let X denote the covariates matrix including characteristic variables.

3.2.1 The First Stage

3.2.1.1 Predicted Propensity of Purchase: Logistic Regression Model

Let us suppose a dichotomous indicator $Z_i = 1$ if a non-zero expenditure is reported, and $Z_i = 0$ otherwise, then a logistic regression model can be represented as:

$$logit\big(P\big(Z_i = 1\big)\big) = \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} + \ldots + \theta_q X_{qi},$$

where X_{1i} is household income, and $X_{2i}, \ldots, X_{qi}$ contain the demographic variables (housing tenure, geographical region, etc.). The coefficients of each logistic regression model will be estimated using both the confidential and the top-coded data for each of the four years, 2008 – 2011.

3.2.1.2 Predicted Expenditures (Buyers Only): Multiple Linear Regression Model

Assuming a linear relationship between the natural log of non-zero expenditure and the natural log of non-zero household income plus other demographic variables, we fit a "double-log" linear regression models between the non-zero expenditures and the non-zero household income after adjusting for the demographic variables for each expenditure and each year of data. The regression model can be represented as:

$$E\big(ln(y_i)\big) = \beta_0 + \beta_1 ln(X_{1i}) + \beta_2 X_{2i} + \ldots + \beta_p X_{pi},$$

where y_i is the non-zero expenditure (one of: property taxes, utilities, health care and domestic services),X_{1i} is the non-zero household income, and $X_{2i}, \ldots, X_{pi}$ contain the demographic variables (housing tenure, geographical region, etc.). The coefficients of each regression model will be estimated using both the confidential and the top-coded data for each of the four years, 2008 – 2011.

3.2.2 The Second Stage

3.2.2.1 Marginal Propensity to Consume (MPC)

At the second stage, the Marginal Propensity to Consume (MPC) is defined as the change in expenditure given a unit change in income (Paulin and Duly 2002 pp. 57).

$$MPC = \frac{\partial E(y)}{\partial X_1}.$$

For an economist, the approximation such as $E(\hat{y}) \approx \hat{P} \exp\left[E\big(ln\hat{y}\big)\right]$ would be a preferred outcome to estimate the MPC.

3.2.2.2 Income Elasticity

The income Elasticity (η) is defined as the percent change in expenditure for a specific good given a 1-percent increase in income (Paulin and Duly 2002 pp. 57):

$$\eta = MPC \frac{X_1}{E(y)}.$$

"The double-log transformation offers a two-fold benefit: one is stabilizing the variance, the additional benefit enables translating the coefficient of the natural logged covariate, e.g. income into an elasticity estimate. That is, in the model of Section 3.2.1.2, if the natural log of income coefficient $\hat{\beta}_1 = 2.0$ with statistically significant p-value, then $\hat{\beta}_1$ can be interpreted as 1% increase in the household income is associated with a $\beta_1 = 2\%$ increase in expenditure (Paulin and Lee 2002 pp. 32).

3.3 Empirical CDF

As illustrated in Section 3.2.2, the departure between confidential and top-coded estimates in terms of predicted propensity of purchase and buyers-only expenditures, will be reflected in the bias of MPC and income Elasticity. To evaluate the top-coding effect, one option is to studying the distribution of empirical CDF (ECDF) from a multiple integration perspective by conditioning on each variable once at a time to estimate the Anderson-Darling Distance (A-DD) between confidential and top-coded ECDF. The motivation for estimating the ECDF of a joint distribution of outcomes and covariates is that it would allow us to compare the discrepancies of distributions between the confidential and top-coded data. Anderson-Darling Distance is a statistical distance between two distributions in term of ECDF, and it places more weight on observations in the tails of the distribution. This will allow the program office to gauge the differences that had been made by top-coding and how they had been reflected into economics models, e.g., in terms of non-zero expenditure reporting propensity scores (PS) model logistic regression coefficient and its 95% CI overlapping.

The joint distribution of $(Y, X_1, X_2, \ldots, X_p)$ forms an ECDF as the following:

$$
\begin{aligned}
F_{Y,X_1,X_2,\ldots,X_p}\Big(y, x_1, x_2, \ldots, x_p\Big) &= P\Big(Y \le y, X_1 \le x_1, X_2 \le x_2, \ldots, X_p \le x_p\Big) \\
&= P\Big(Y \le y, X_1 \le x_1, X_2 \le x_2, \ldots | X_p \le x_p\Big) P\Big(X_p \le x_p\Big) \\
&= P\Big(Y \le y, X_1 \le x_1 | X_2 \le x_2, \ldots, X_p \le x_p\Big) \\
&\quad \times P\Big(X_2 \le x_2 | \ldots, X_p \le x_p\Big) \ldots P\Big(X_p \le x_p\Big) \\
&= P\Big(Y \le y | X_1 \le x_1, X_2 \le x_2, \ldots, X_p \le x_p\Big) \\
&\quad \times P\Big(X_1 \le x_1 | X_2 \le x_2, \ldots, X_p \le x_p\Big) \\
&\quad \times P\Big(X_2 \le x_2 | \ldots, X_p \le x_p\Big) \ldots P\Big(X_p \le x_p\Big).
\end{aligned}
$$

Let w_i demote the final weights of the i^{th} household to account for the CE complex design, this ECDF can be estimated by

$$
\begin{aligned}
\hat{F}_{Y,X_1,X_2,\ldots,X_p}(a, b, c, \ldots, d) \;=\; & \frac{1}{w}\sum_i w_i I\Big(Y_i \le a | X_{1i} \le b, X_{2i} \le c, \ldots, X_{pi} \le d\Big) \times \\
& \frac{1}{w}\sum_i w_i I\Big(X_{1i} \le b | X_{2i} \le c, \ldots, X_{pi} \le d\Big) \times \\
& \frac{1}{w}\sum_i w_i I\Big(X_{2i} \le c | \ldots, X_{pi} \le d\Big) \times \\
& \ldots \frac{1}{w}\sum_i w_i I\Big(X_p \le d\Big),
\end{aligned}
$$

where $w = \sum_i w_i$ is the sum of final weights of a subset, e.g. formed by cross tables of housing tenure (owner or not), geographical region (Northeast, Midwest, South, West), reference person's ethnicity (white or non-white) and gender (Male, Female).

3.4 Anderson–Darling Distance (A-DD) Measure

Let $F_c(x)$ be the confidential ECDF and let $F_t(x)$ be the top-coded ECDF, and let j index a subset based on a cross table cell of housing tenure, geographical region, reference person's

ethnicity and gender and let subset sample proportion be p_j. Then the A-DD is

$$A = \sum_j p_j n_j \frac{\Big(F_t(x) - F_c(x)\Big)^2}{F_c(x)\Big(1 - F_c(x)\Big)} dF_c(x)$$

4. Comparing Confidential and Top-coded Distributions and Parameter Estimates

4.1 Comparing Confidential and Top-coded ECDFs

Since the ECDF function is obtained by a series of terms of conditional probability through summation, we can first take a closer look at a simple ECDF term involving household income and property tax conditional on other demographics under the subset of 2008 Renter, Northeast region and White Male (Figure 2 (a)), we can see a clear departure at the tail end of the distribution between the confidential and top-coded data, and the rest of distributions are identical. However, if we look further into a triple ECDF term involving age, household income and property tax conditional on other demographics under the same subset (Figure 2 (b)), then the tail-end departure became less visible at each step, and so forth. Hence, in the ECDF of all demographics under the same subset (Figure 2 (c)), the distribution between the confidential and top-coded data are almost identical. The same pattern appeared across 2009 – 2011 for the identical subset (Figure 3 - 5).

Therefore, the ECDF plots indicated the distribution departure between confidential and top-coded data occurred at the tail end. This observation made us to wonder how would a distributional measure, e.g. A-DD reflect those tail end deviations. And nonetheless, those tail end departures played the role as outliers and/or influential observations to alter the household income coefficient estimates and 95% CIs as the consequence of top-coding, e.g. in the propensity model of non-zero expenditure reporting.

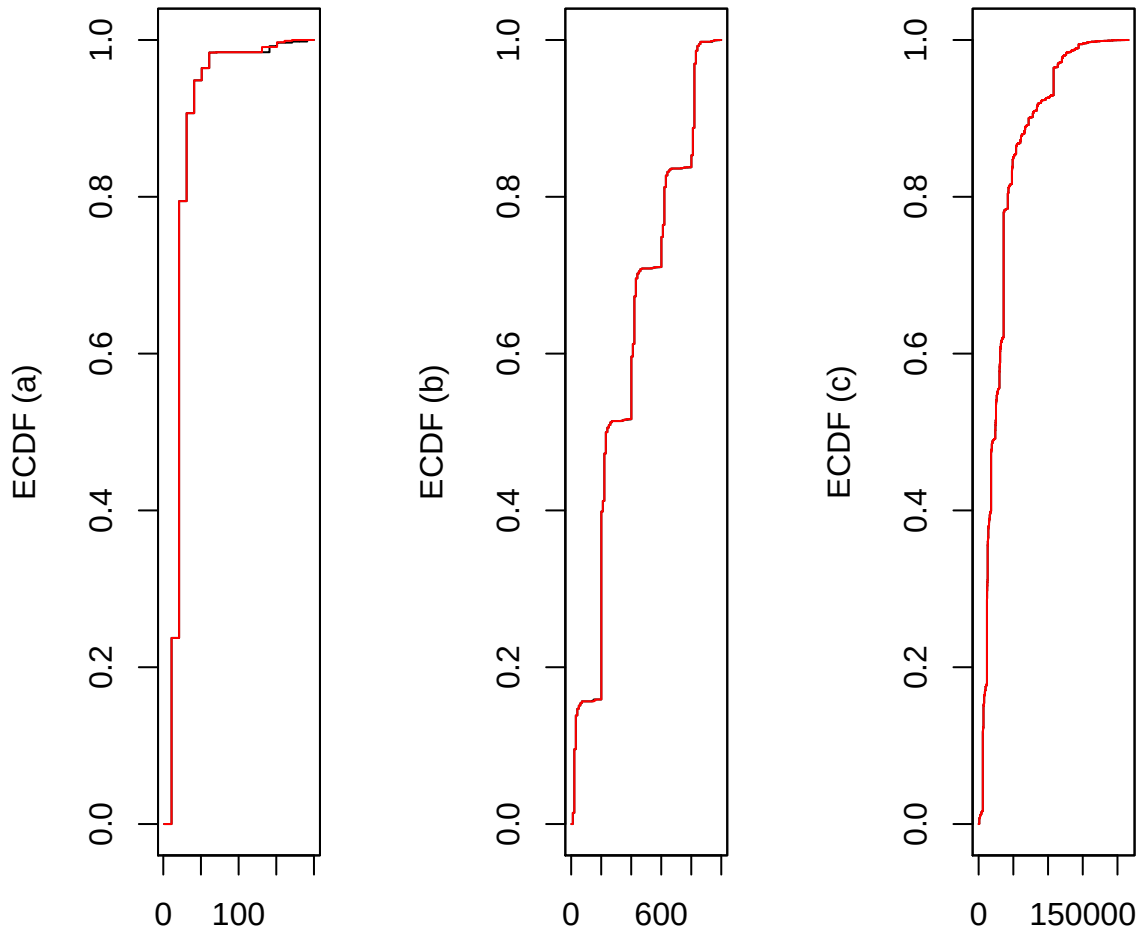

Figure 2: Empirical CDF 2008 Subgroup (Renter, Northeast region, White, Male):
(a) $F_{(Income,\ Property\ Tax)}$, (b) $F_{(Age,\ Income,\ Property\ Tax)}$,
(c) $F_{(Child\ \#,\ Senior\ \#,\ Education,\ FamilySize,\ Age,\ Income,\ Property\ Tax)}$.

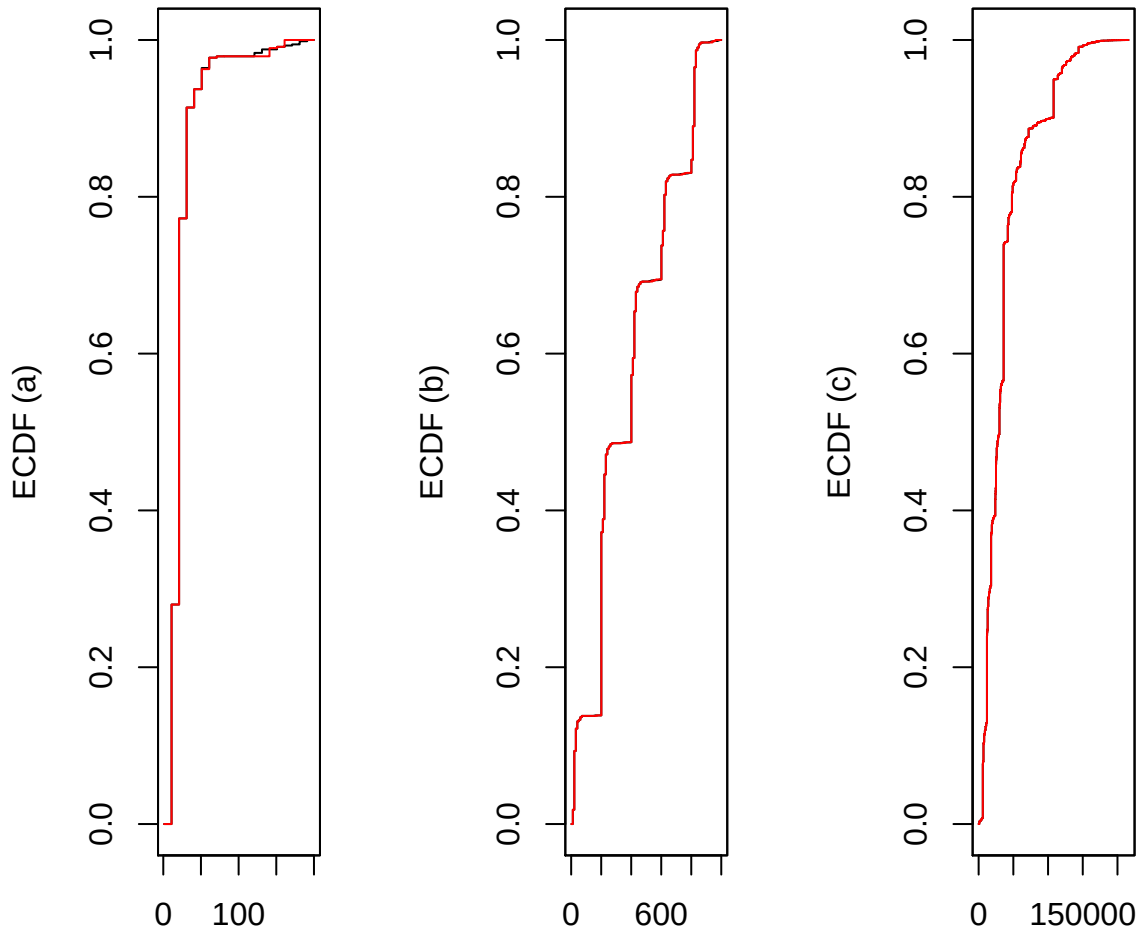

Figure 3: Empirical CDF 2009 Subgroup (Renter, Northeast region, White, Male):
(a) $F_{(Income,\ Property\ Tax)}$, (b) $F_{(Age,\ Income,\ Property\ Tax)}$,
(c) $F_{(Child\ \#,\ Senior\ \#,\ Education,\ FamilySize,\ Age,\ Income,\ Property\ Tax)}$.

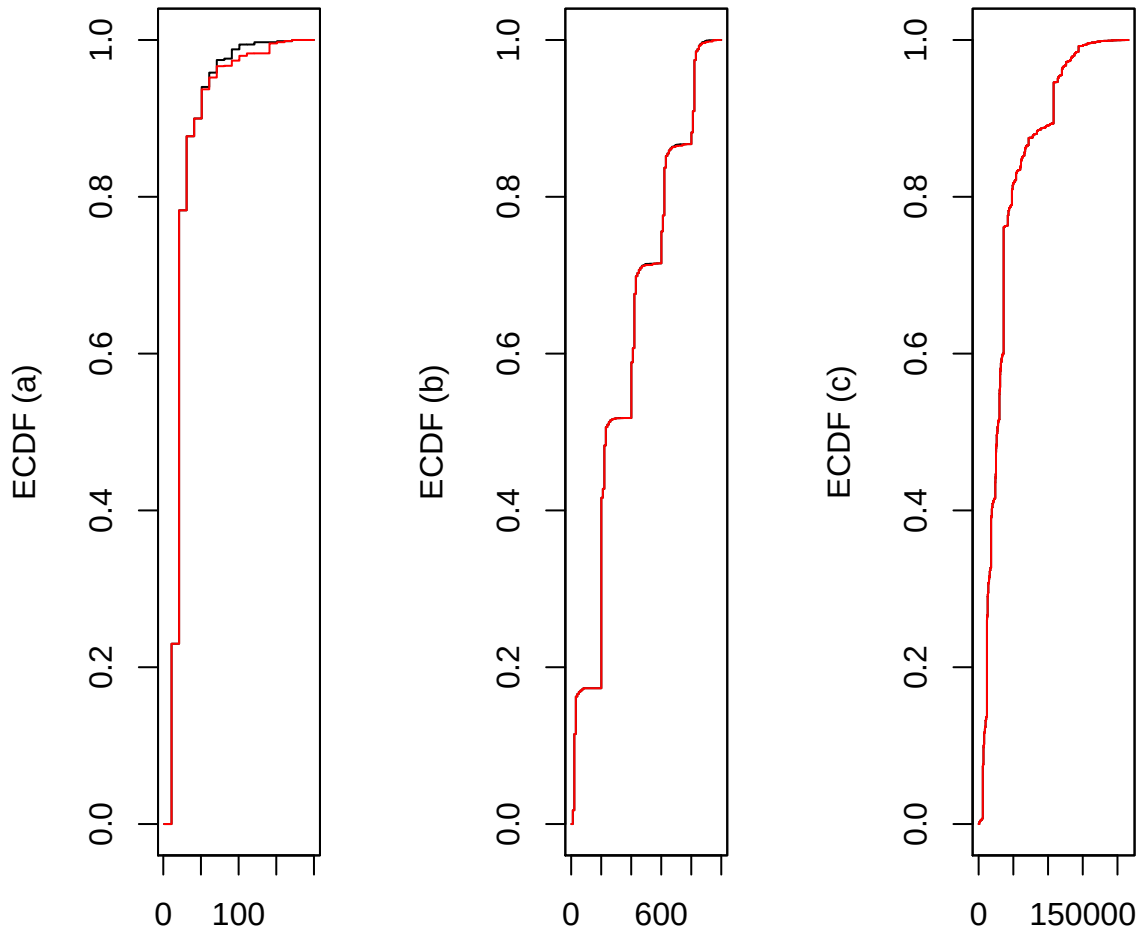

Figure 4: Empirical CDF 2010 Subgroup (Renter, Northeast region, White, Male):
(a) $F_{(Income,\ Property\ Tax)}$, (b) $F_{(Age,\ Income,\ Property\ Tax)}$,
(c) $F_{(Child\ \#,\ Senior\ \#,\ Education,\ FamilySize,\ Age,\ Income,\ Property\ Tax)}$.

4.2 Model of Specifications of Non-zero Expenditures and Propensity of Purchase

First, in order to specify the terms for modeling the natural logarithm of non-zero expenditures, we employed a regression tree to provide a glimpse of the natural logarithm of households income, its potential quadratic terms, characteristic variables and possible interactions. Then, we applied the Bayesian information criterion (BIC) backward elimination to determine the final model specification of the natural log of non-zero expenditures of property taxes, utilities, health care and domestic services. And then we applied these models for 4 years (2008-2011).

Second, in order to specify the terms for modeling the propensity of purchase (non-zero expenditure reporting), we employed a regression tree to provide a glimpse of households income, characteristic variables and possible interactions. Then, we applied the Bayesian information criterion (BIC) backward elimination to determine the final model specification of the propensity of non-zero reporting of property taxes, utilities and health care and domestic services. And then we applied these models for 4 years (2008-2011).

4.3 Quadratic Term of the Natural Logarithm of Households Income

Here are the scatter plots of the "double-log" of 2008 non-zero property taxes vs. household (HH) income by region (Northeast, Midwest) and by household size < 3 $vs.$ ≥ 3 (Figure 6). We can see that the "double-log" transformation did reduce the deviation, and the curvature indeed indicated a quadratic term of the natural log of household income. We also see the similar pattern (of "double-log" of non-zero property taxes vs. household income) by region (South, West) and by household size < 3 $vs.$ ≥ 3 (Figure 7).

4.4 Results of BIC Selected Coefficient Estimates of the Natural Log of Household Income for Modeling Natural Log of Non-zero Expenditures, 95% CI

Figure 8 shows the estimated coefficient of the natural log of household income and 95% CI over years of the BIC selected multiple linear regression (MLR) model where the natural log of non-zero expenditures are the response variables. The 95% CIs seems to overlap very well for the "double-log" regression models across 2008-2011.

4.5 Results of A-DD Estimates and Propensity Models of Non-zero Expenditure Reporting Coefficient Estimates of Household Income, 95% CI

We investigated A-DD from two aspects, one is from the empirical CDF based on multiple integration by conditioning on each variable once at a time (CECDF), another is from the empirical CDF only based on each expenditure variable.

For the estimated household income coefficient and its 95% CI over the years of the BIC selected logistic propensity model where non-zero property taxes reporting indicator is the response variable (Figure 9), we can see that the 95% CIs seems to overlap $< 50\%$ for 2009 and 2010, and there is no-overlapping in 2011.

For the estimated household income coefficient and its 95% CI over the years of the BIC selected logistic propensity model where non-zero utilities reporting indicator is the response variable (Figure 10), the 95% CIs appears to overlap pretty good for non-zero utilities reporting propensity from 2008 to 2010, but overlap $< 50\%$ for 2011.

For the estimated household income coefficient and its 95% CI over the years of the BIC selected logistic propensity model where non-zero health care reporting indicator is the response variable (Figure 11), we can see that the 95% CIs seems to overlap pretty good for non-zero health care reporting propensity from 2008 to 2011.

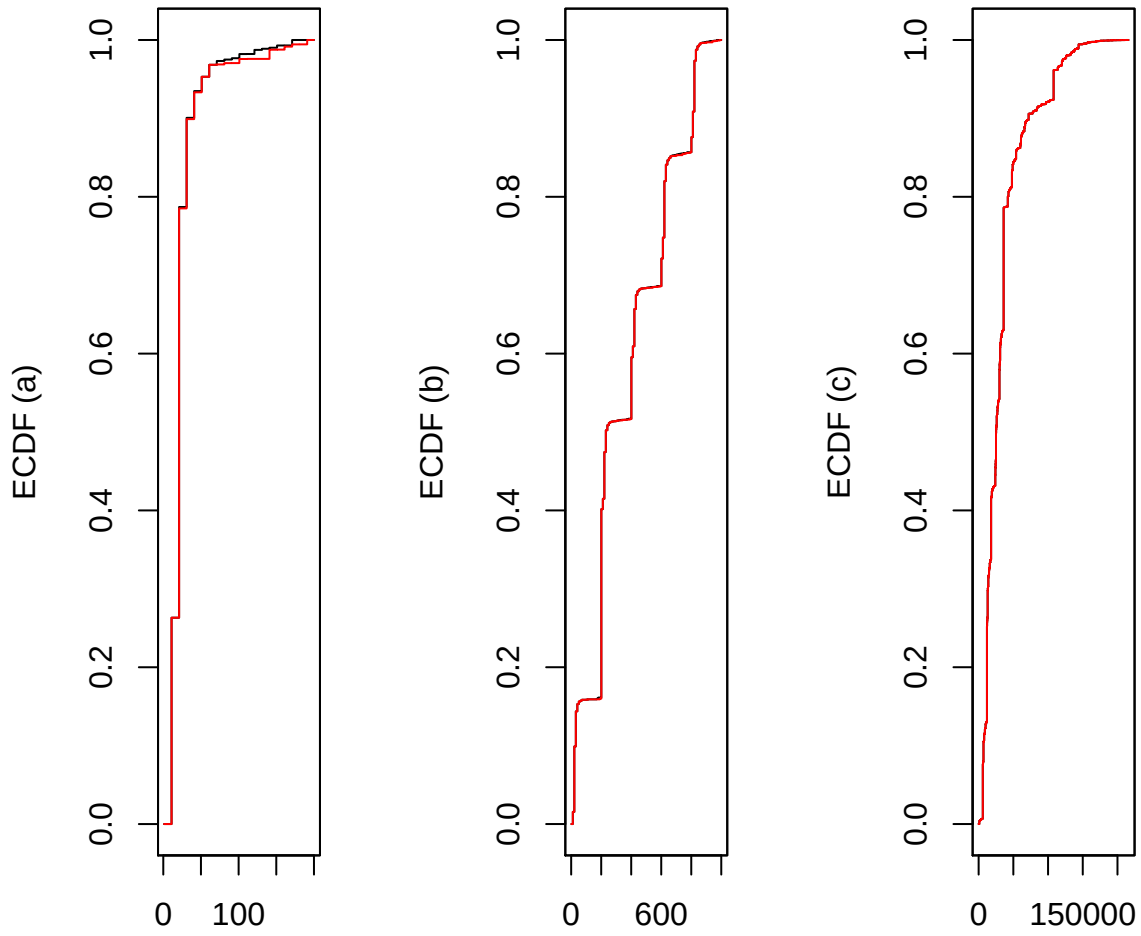

Figure 5: Empirical CDF 2011 Subgroup (Renter, Northeast region, White, Male):
(a) $F_{(Income,\ Property\ Tax)}$, (b) $F_{(Age,\ Income,\ Property\ Tax)}$,
(c) $F_{(Child\ \#,\ Senior\ \#,\ Education,\ FamilySize,\ Age,\ Income,\ Property\ Tax)}$.

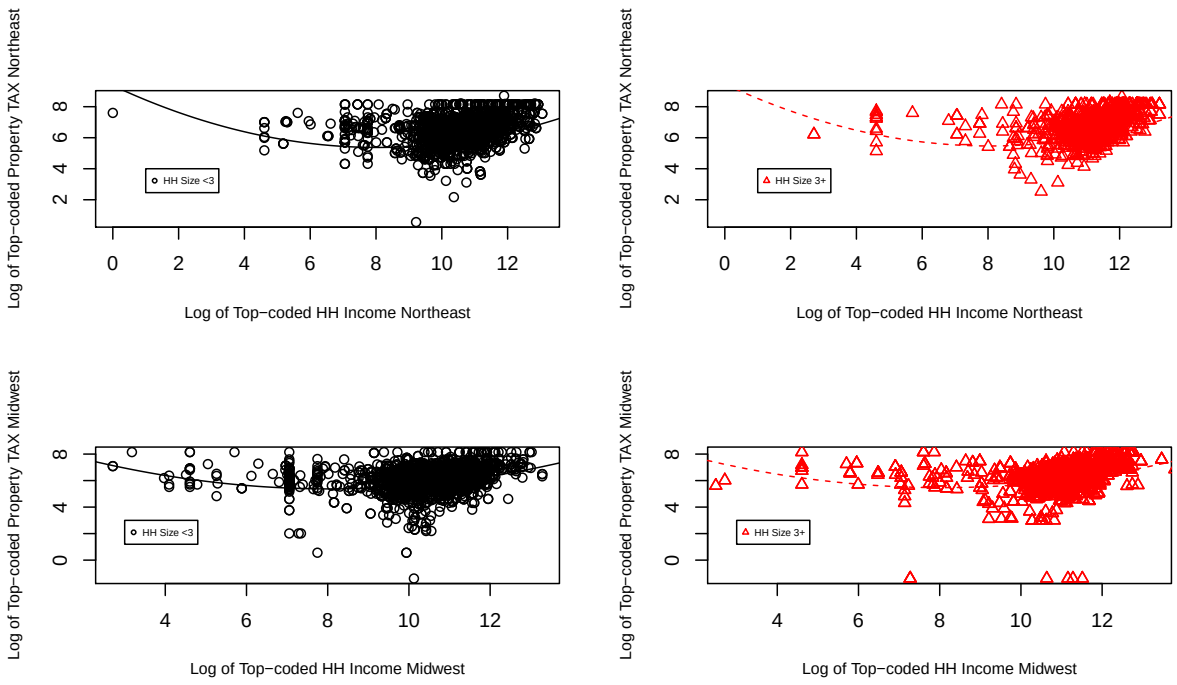

Figure 6: 2008 Double-log of Top-coded Non-zero Property Taxes vs. Household (HH) Income by Region (Northeast, Midwest) and HH size ($< 3, \geq 3$)

For the estimated household income coefficient and its 95% CI over the years of the BIC selected logistic propensity model where non-zero domestic services reporting indicator is the response variable (Figure 12), the 95% CIs appears to overlap $< 50\%$ for 2008, and no-overlapping in 2010, but seems to overlap very well for non-zero domestic services reporting propensity in 2009 and 2011.

Anderson-Darling Distance will reflect variously of Empirical CDF differences depend on specific models.

The estimated A-DD reflected differently depends on specific models, for the top-coded propensity of non-zero expenditure reporting logistic coefficient of household income and its 95% CI, comparing to the confidential one across 2088 – 2011 (Figure 9 - 12). This would suggest the difficulty of a distributional measure like A-DD to capture the differences of specific model between confidential and top-coded coefficient estimates, e.g. propensity of consumption.

4.6 A Robust Logistic Regression Approach

As the consequence of influential observations in Section 4.5, we observed the tipping point, which is the proportion of extreme data that the logistic regression estimator can sustain before being altered to a swiftly excessive amount. For an example, in a linear regression scenario, parallel to the sample mean which can be easily shifted by a handful of outliers, a few extreme observations can disrupt the ordinary least squares (OLS) estimate (Fox and Weisberg 2010). One solution is to adopt an estimator that offsets the impact from influential observations and outliers to counter the extreme data, and hence, to provide robust estimates. Künsch et. al. (1989) proposed the conditionally unbiased bounded-influence estimator (CUBIF), where the "influence function" quantifies the proximate impact of adding or removing an influential observation. The purpose is to confine the

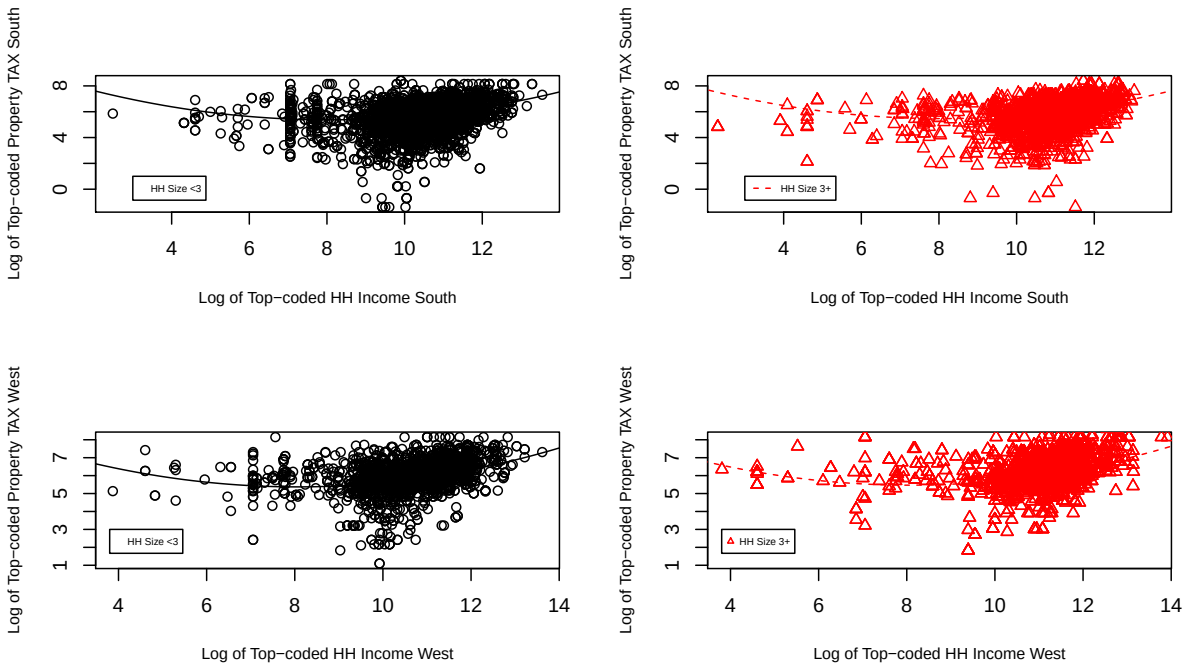

Figure 7: 2008 Double-log of Top-coded Non-zero Property Taxes vs. Household (HH) Income by Region (South, West) and HH size ($< 3, \geq 3$)

influence function by setting up a limit on it and to solve the estimating equation with lower variation in term of "lower-weighting" influential, leveraging and outlying observations.

Carroll and Pederson (1993) simplified a CUBIF approach for a logistic model:

$$P(Y = 1|x) = \big[1 + exp\big(-x^T\theta\big)\big]^{-1}.$$

A robust estimator solves the estimating equation:

$$\sum_{i=1}^{n} w_i x_i \big(Y_i - \big[1 + exp\big(-x_i^T\theta\big)\big]^{-1} - c\big(x_i, \theta\big)\big) = 0,$$

where $w_i = w\big(x_i, x_i^T\theta, Y_i\big)$ is in "Schweppe" class and the $c\big(x_i, \theta\big)$ regularity is in Künsch, et. al (1989). The estimating equation is conditionally unbiased given x:

$$E\big[w\big(x, x^T\theta, Y\big)\big(Y - \big[1 + exp\big(-x^T\theta\big)\big]^{-1} - c\big(x, \theta\big)\big)\big] = 0.$$

The $robust$ package in R provides a numerical implementation of CUBIF approach. However, we found that the singularity would occur because of the factor variables would cause the covariance matrix not to have an inverse (became unsolvable), e.g. the covariance matrix became singular during the iterations, etc. Hence, in our work, we further reduced categorical terms and interactions in all CE non-zero expenditure reporting propensity model so $glmRob$ function could produce results. Figure 13 shows that the estimated coefficient of the household income and 95% CI over years of the robust logistic propensity model where non-zero expenditure reporting indicators are the response variables. The 95% CIs seems to overlap very well for the robust logistic propensity models across 2008-2011.

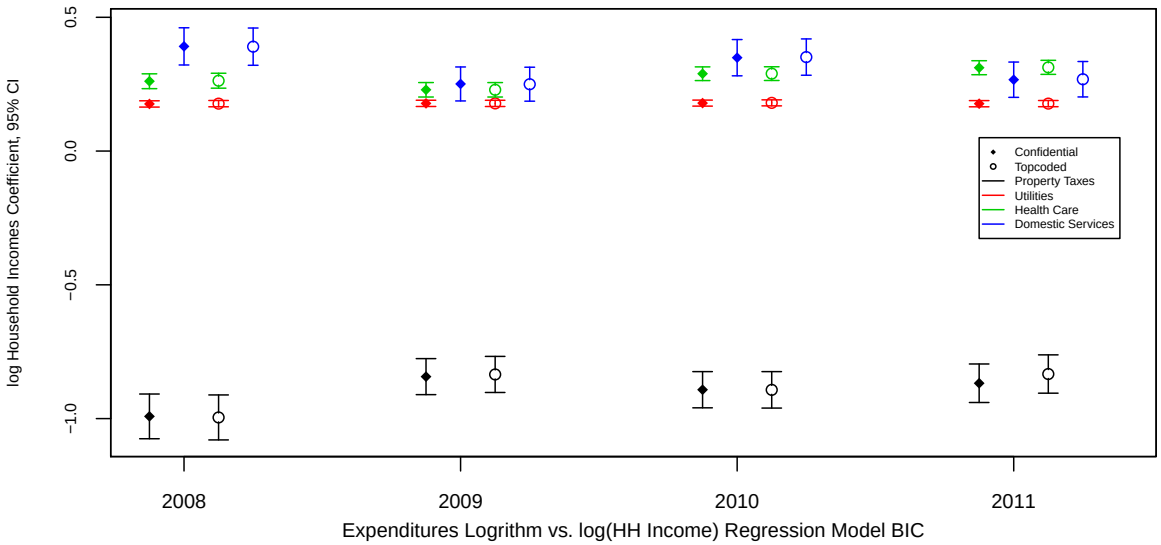

Figure 8: 2008–2011 BIC Selected MLR: Natural Log of Non-zero Expenditures vs. Natural Log of Non-zero Household (HH) Income

4.7 Departure Between BIC Selected and Robust Propensity Curves of Predicted Values

We provided the following visual comparisons on how far the BIC selected propensity curves and robust propensity curves of predicted values would apart from each other.

Figure 14 shows the comparison between non-overlapping BIC selected and robust non-zero property taxes (2011) and domestic services (2010) reporting propensity (predicted values) for household income only of confidential and top-coded data. We can see a big departure in 2011 non-zero property taxes reporting propensity.

Figure 15 shows the comparison between overlapping $< 50\%$ BIC selected and robust non-zero property taxes (2009, 2010) reporting propensity (predicted values) for household income only of confidential and top-coded data. There seems to be a visible non-parallel pattern in 2009 non-zero property taxes reporting propensity.

Figure 16 shows the comparison between overlapping $< 50\%$ BIC selected and robust non-zero utilities (2011) and domestic services (2008) reporting propensity (predicted values) for household income only of confidential and top-coded data. We can see a non-parallel departure in 2008 non-zero domestic services reporting propensity.

5. Conclusion

In this article, we investigated the regression, logistic and robust logistic coefficients, computed using the top-coded data set compared to those computed using the confidential data set, of CE expenditures and household income after adjusting demographic characteristics. We studied the distribution of empirical CDF (ECDF) from a multiple integration perspective by conditioning on each variable once at a time and adopted the Anderson-Darling Distance (A-DD) between confidential and top-coded ECDF as a measure to evaluate the

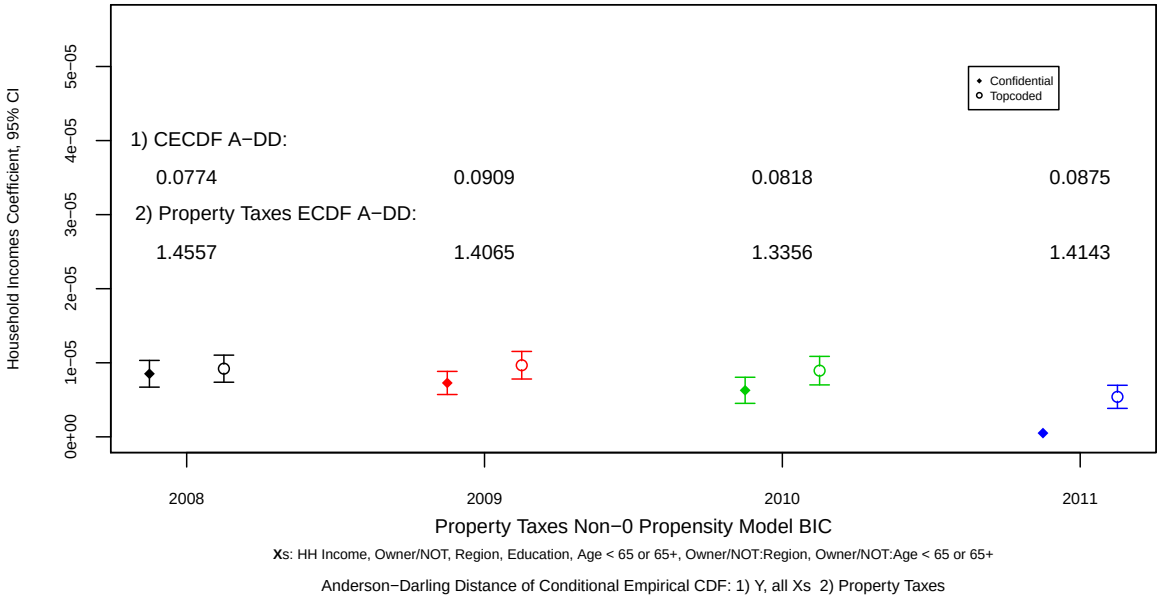

Figure 9: Aggregated Propensity of Non-zero Property Taxes Reporting, Logistic Regression Household Incomes Coefficient, 95% CI and Anderson-Darling Distance of Empirical CDF: 0.077 (2008), 0.091 (2009), 0.082 (2010), 0.0889 (2011).

impact of top-coding. Our results showed that:

1. Empirical CDF based on conditioning multiple integration provides an alternative to obtain estimated joint distribution for both confidential and top-coded data.

2. Anderson-Darling Distance will reflect variously of Empirical CDF differences depend on specific models.

3. Regression model: ln(Non-0 Expenditures) $\sim$ ln(HH Income) + Characteristics, stabilizes influential observations and produces good overlapping of coefficient 95% CI between confidential and top-coded data.

4. For the non-zero expenditures reporting propensity model, departures and poor overlapping between confidential and top-coded coefficients will have an impact in terms of model specific economic measurements, e.g. MPC and income elasticity under a Cragg two-stage model.

5. Conditionally Unbiased Bounded-Influence (CUBIF) estimation provides an alternative robust logistic approach, nonetheless, this approach removes categorical interactions and factors if necessary.

For future research steps, we would like to explore the following:

1. Further exploration of $glmRobrobust$ is needed, especially on categorical terms, e.g. robust logistic regression on subsets formed by categorical cross tables, then summarize those subset coefficients by weighting proportionally to size vs. down-weighting the influential subset?

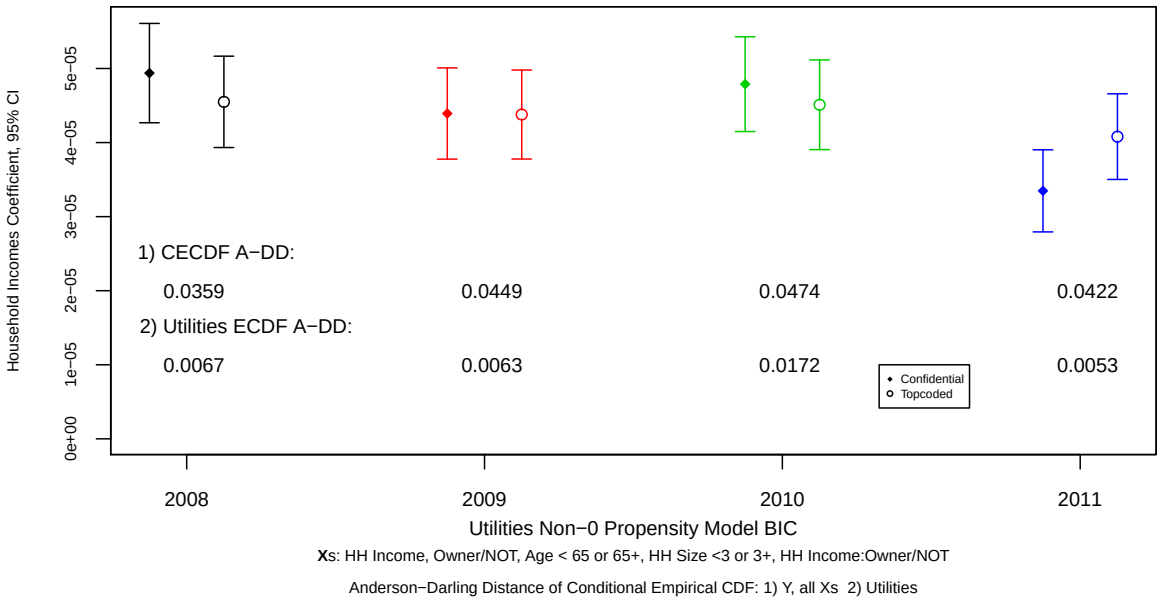

Figure 10: Aggregated Propensity of Non-zero Utilities Reporting, Logistic Regression Household Incomes Coefficient, 95% CI and Anderson-Darling Distance of Empirical CDF: 0.036 (2008), 0.045 (2009), 0.047 (2010), 0.042 (2011).

2. Bootstrap simulation: produce distributions of gini index, MPC and income elasticity for comparing confidential vs. top-coded data, then to evaluate and compare those distributions?

3. How would the Anderson-Darling Distance reflect the distributions of those "2^{nd} stage" estimates between confidential vs. top-coded data?

REFERENCES

Bureau of Labor Statistics. *Consumer Expenditure Survey (CE) Program*. Retrieved from http://www.bls.gov/cex/.

Bureau of Labor Statistics. (2009). *BLS Handbook of Methods* (Chapter 16 Consumer Expenditures and Income). Retrieved from http://www.bls.gov/opub/hom/pdf/homch16.pdf.

Yang, D. K. and Gonzalez J. M. (2013). "Impact of Design Changes on Economic Analyses Project Report," Bureau of Labor Statistics Technical Report.

Yang, D. K. and Toth, D. (2014). "Measuring Impact of Top-Coding on the Utility of Consumer Expenditure Microdata," *Proceedings of the Survey Research Methods Section*, Joint Statistical Meetings 2014, Boston, MA.

Karr, A. F., Kohnen, C. N., Oganian A., Reiter, J. P. and Sanil, A. P. (2006). "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality," *The American Statistician*, Vol. 60, No. 3, 1–9.

Karr, A.F., Oganian, A., Reiter, J.P. and Woo, Mi-Ja (2006). "New Measures of Data Utility," in *Workshop Manuscripts of Data Confidentiality, A Working Group in National Defense and Homeland Security*. Retrieved from http://sisla06.samsi.info/ndhs/dc/Papers/NewDataUtility-01-10-06.pdf.

Woo, M.-J., Reiter, J.P., Oganian, A. and Karr, A.F. (2009). "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation," *The Journal of Privacy and Confidentiality*, Vol. 1, Number 1, pp. 111–124.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap* (1st edition). Chapman & Hall/CRC.

Kullback, S. and Leibler, R.A. (1951), "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22 (1): 7986.

Lock, E.F. and Dunson, D. B. (2014), "Shared kernel Bayesian screen," *Cornell University Statistics Methodology*. Retrieved from http://arxiv-web3.library.cornell.edu/abs/1311.0307v2.

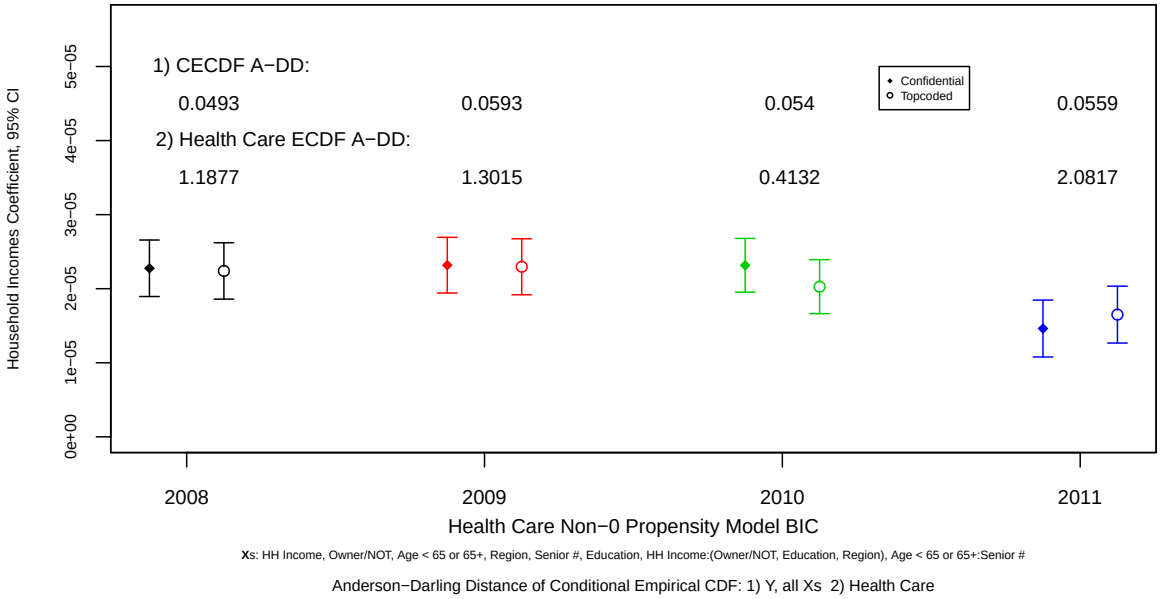

Figure 11: Aggregated Propensity of Non-zero Health Care Taxes Reporting, Logistic Regression Household Incomes Coefficient, 95% CI and Anderson-Darling Distance of Empirical CDF: 0.049 (2008), 0.059 (2009), 0.054 (2010), 0.056 (2011).

Garner, T. I. (1993), "Consumer Expenditures and Inequality: An Analysis Based on Decomposition of the Gini Coefficient," *The Review of Economics and Statistics*, Vol. 75, No. 1, pp. 134-138.

Landsburg S. E. (1999), *Price Theory and Applications* (4^{th} edition), South-Western College Publishing.

Jacoby, G. W. (2005). "Regression III: Advanced Methods" Workshop, Department of Political Science, Michigan State University. Retrieved from http://polisci.msu.edu/jacoby/icpsr/regress3/lectures/week3/12.Robust.pdf

Cragg, John G. (1971). "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods." *Econometrica*, Vol. 39, No. 5 (Sep., 1971), pp. 829-844.

Dippo, Cathryn S. (1984). "A Comparison of Variance Estimators Using the Taylor Series Approximation." *Census Bureau Statistical Research Division (SRD) Report Series*, SRD Research Report Number CENSUS/SRD/RR-84-21.

Omori, Megumi (2010). "Household Expenditures on Children." *Monthly Labor Review*, 133(9), 316.

Chiswick, Barry R. (1974). "The Level of Income." *Income Inequality: Regional Analyses within a Human Capital Framework*, Chapter 7 pp. 119 - 142, National Bureau of Economic Research, ISBN: 0-870-14264-X.

Family Spending (2009). "Chapter 5: Regression analysis of household expenditure and income", *Family Spending*, Volume 2009, Issue 1 (2009), 7278. doi:10.1057/fsp.2009.6.

Pritchett, Lant (2010). "Divergence, Big Time." *The Journal of Economic Perspectives*, Vol. 11, No. 3 (Summer, 1997), pp. 3-17. Macmillan, a division of Macmillan Publishers Limited.

Paulin, Geoffrey D. and Duly, Abby L. (2002). "Planning Ahead: Consumer Expenditure Patterns in Retirement." *Monthly Labor Review* (MLR), July 2002, 38-58.

Paulin, Geoffrey D. and Lee, Y. G. (2002), "Expenditures of single parents: how does gender figure in?" *Monthly Labor Review* (MLR), July 2002, 16-37.

Fox, J. and Weisberg, H. S. (2011). Robust Regression in R, An Appendix to An R Companion to Applied Regression, 2^{th} Edition, SAGE Publications, Inc. (Jan. 26th).

Künsch, H. R., Stefanski, L. A. and Carroll, R. J. (1989). "Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, with Applications to Generalized Linear Models." *Journal of the American Statistical Association*, Vol. 84, No. 406 (Jun., 1989), pp. 460-466.

Carroll, R. J. and Pederson, S. (1993). "On Robustness in the Logistic Regression Model", *Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 55, No. 3 (1993), pp. 693-706.

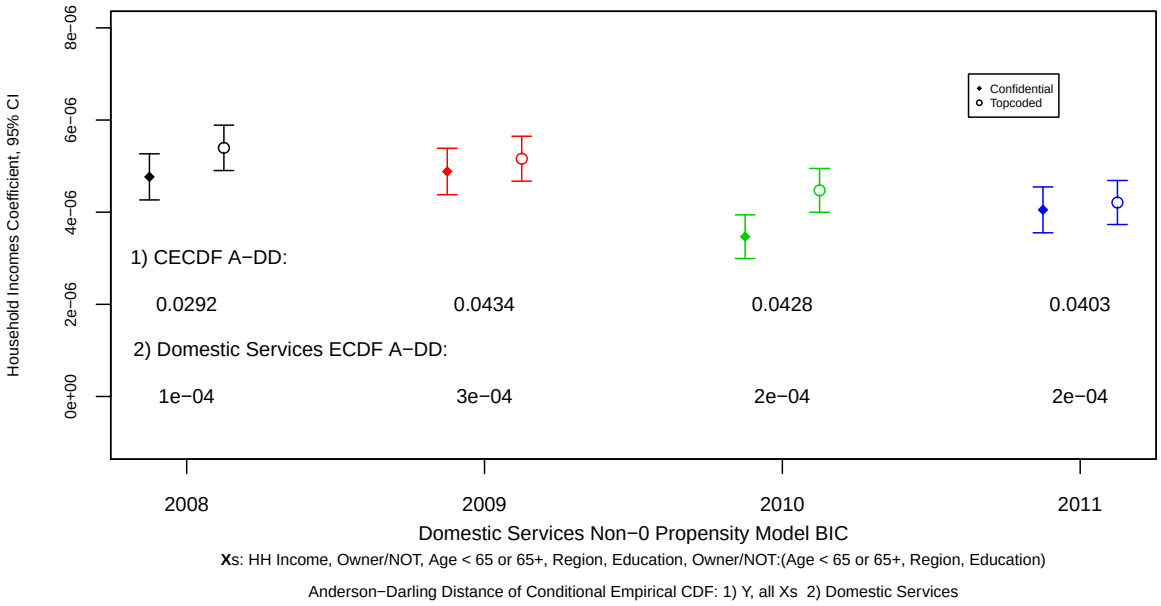

Figure 12: Aggregated Propensity of Non-zero Domestic Services Reporting, Logistic Regression Household Incomes Coefficient, 95% CI and Anderson-Darling Distance of Empirical CDF: 0.029 (2008), 0.043 (2009), 0.043 (2010), 0.040 (2011).

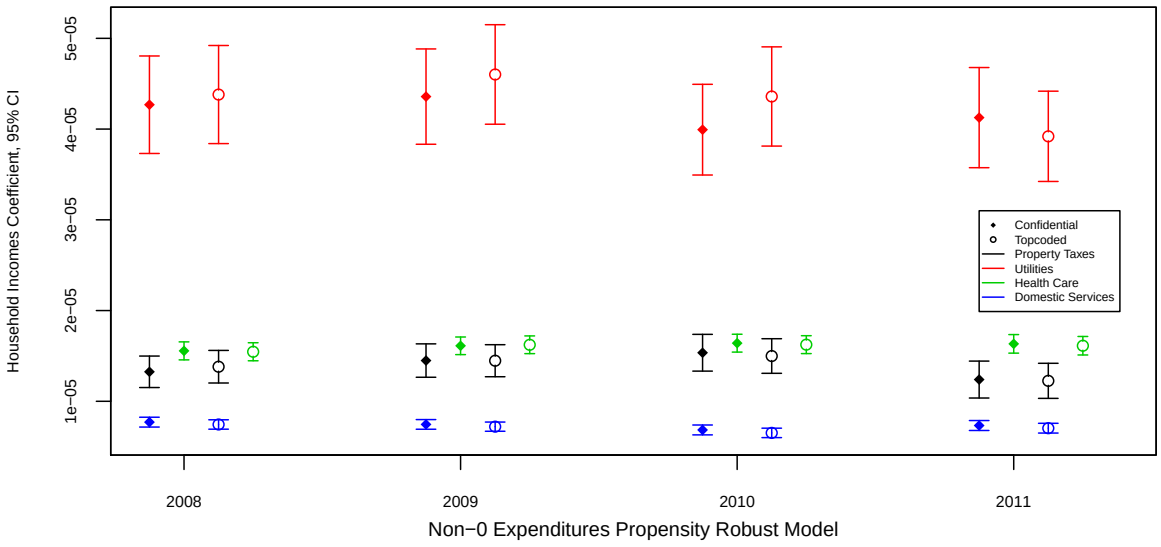

Figure 13: 2008–2011 BIC Selected MLR: Robust Propensity Model of Non-zero Expenditures Reporting vs. Household (HH) Income

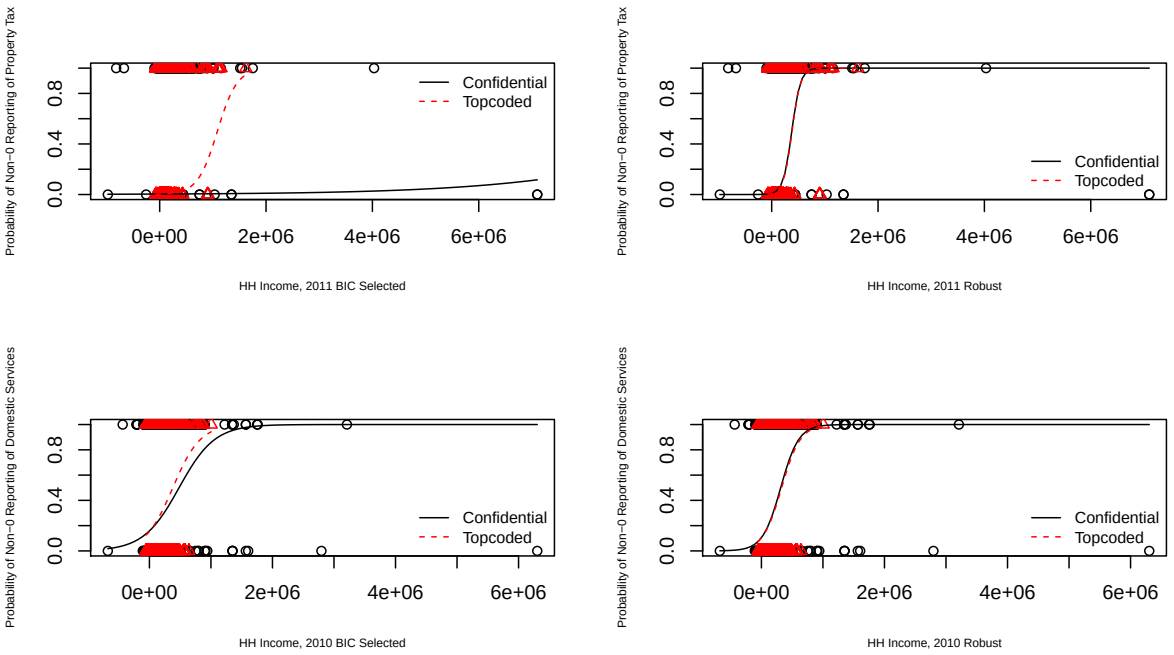

Figure 14: Non-0 Property Taxes (2011) and Domestic Services (2010) Reporting Propensity Curves of Confidential and Top-coded: BIC vs. Robust

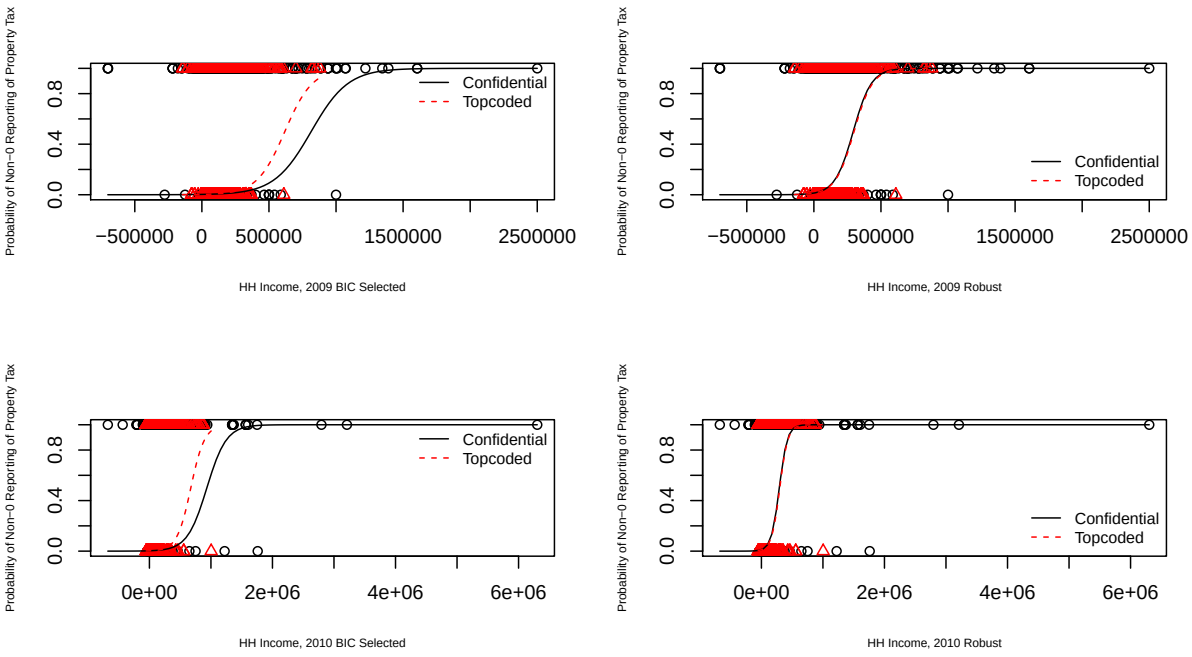

Figure 15: Non-0 Property Taxes (2009, 2010) Reporting Propensity Curves of Confidential and Top-coded: BIC vs. Robust

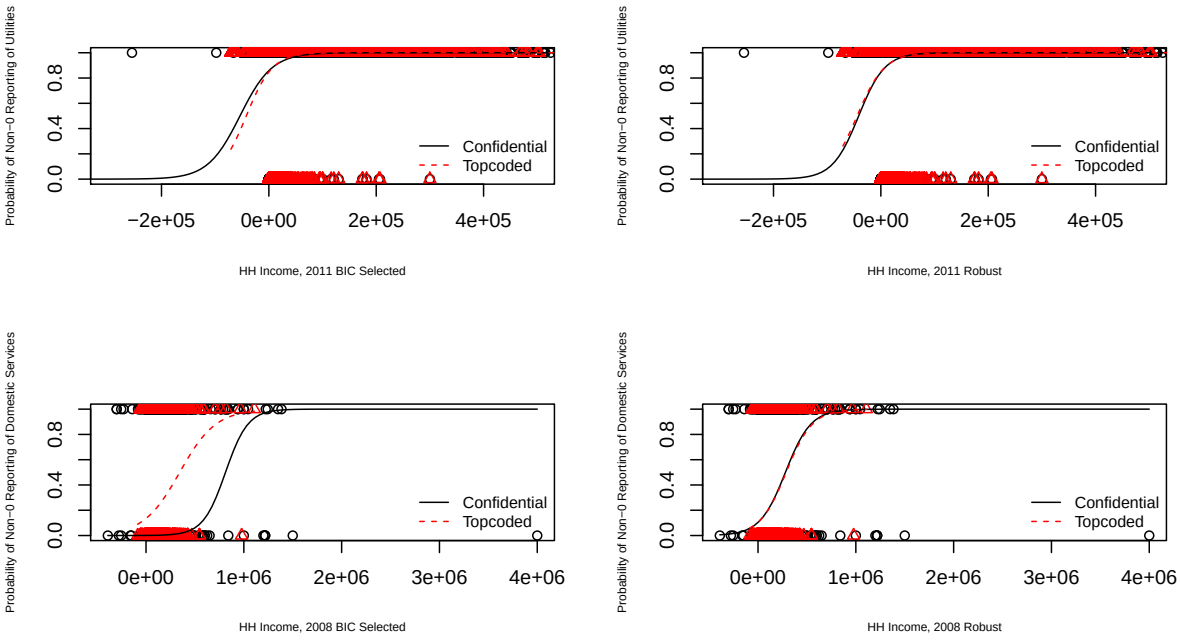

Figure 16: Non-0 Utilities (2011) and Domestic Services (2008) Reporting Propensity Curves of Confidential and Top-coded: BIC vs. Robust