Measuring Impact of Top-Coding on the Utility of Consumer Expenditure Microdata

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 models used to analyze the relationship between certain expenditures and household income after adjusting demographic characteristics. We conduct a bootstrap re-sampling study and implement a data utility measurement based on a modified form of Kullback-Liebler divergence to evaluate the effects of top-coding on the utility of the CE microdata.

Key Words: confidentiality, disclosure limitation, utility measures, confidence intervals overlap, 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 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

*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

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 the 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.

In this article, we propose using bootstrap samples to obtain the empirical distributions of the original and the altered data then to apply a modified measure of Kullback-Liebler (K-L) divergence between those empirical distributions. Our approach implements bootstrap re-sampling, hence, the empirical distributions would tend toward a normal distribution. Using a modified K-L divergence will bridge a global measure to specific analyses and models. Having a measure of data quality could benefit the CE program office by enabling them to gauge the level of change made by top-coding, e.g., under multiple linear regression (MLR). The organization of the paper is the following: Section 2 introduces the CE top-coding process and illustrates its impacts in visualizations. In Section 3, we describe the CE data sets and introduce the bootstap re-sampling and the modified K-L divergence used to evaluate the numerical impacts of top-coding. Section 4 contains the results of our analysis on CE data and Section 5 contains a discussion of conclusions drawn 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) conditional on demographic information. 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? Let us look at another example of an expenditure item, property tax, that is top-coded more often and highly associated with household income (Figure 2).

In Figure 2, we can see that most data points are top-coded for property tax only (red), few points are top-coded for household income only (green), with more points being top-coded for both property tax and household income (blue). Points that are not top-coded for either household income or property tax are in black. Higher income appears to be associated with the higher property tax, and top-coding appears to make this association increase. The pink regression fit line shows the estimated regression line between household income and property tax using the top-coded microdata, while the light blue line is line obtained using the original microdata.

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

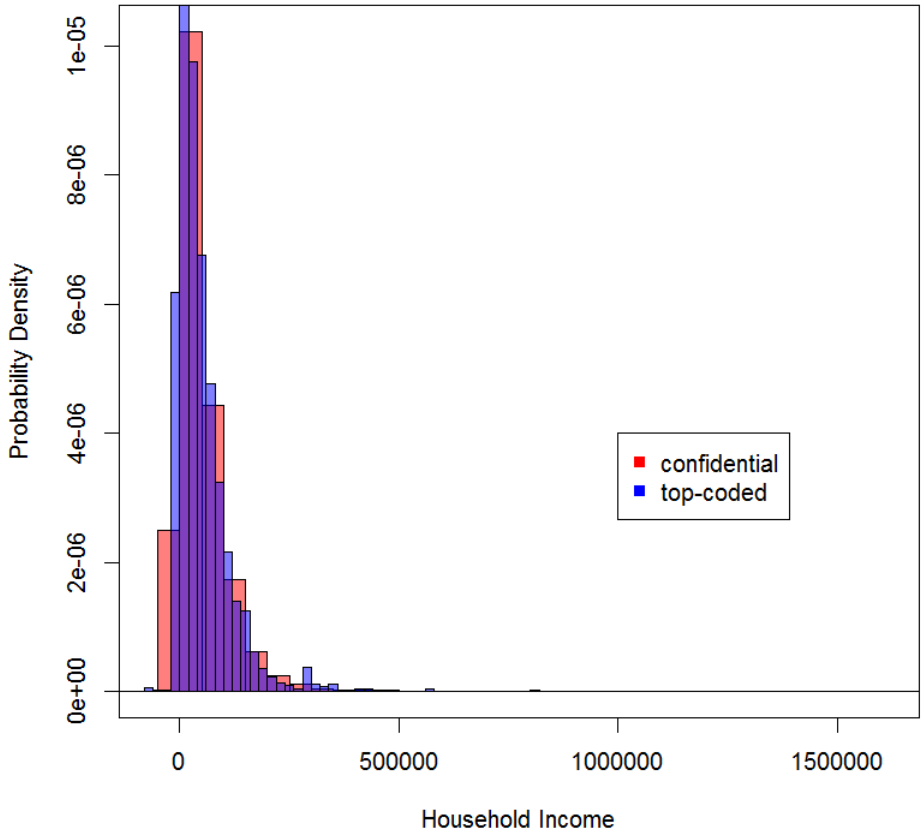

Figure 1: CE 2011 Household Income Sampling Distribution: Confidential vs. Top-coded

the household; number of members under age 2 in the household; reference person's age, ethnicity, education attainment, and gender (Male, Female).

3.2 Multiple Regression Model

Assuming a linear relationship between each expenditure and the household income and other demographic variables, we fit linear regression models between expenditures and the household income conditioned on the demographic variables for each expenditure and each year of data. The regression model can be represented as:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi},$$

where y_i is the expenditure (one of: property taxes, utilities, health care, and domestic services), X_{1i} is 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.

3.3 Bootstrap Procedure

The motivation for using a bootstrap data set is that because the sample is obtained from an unknown population sampled using a complex sample design, the sampling distribution of

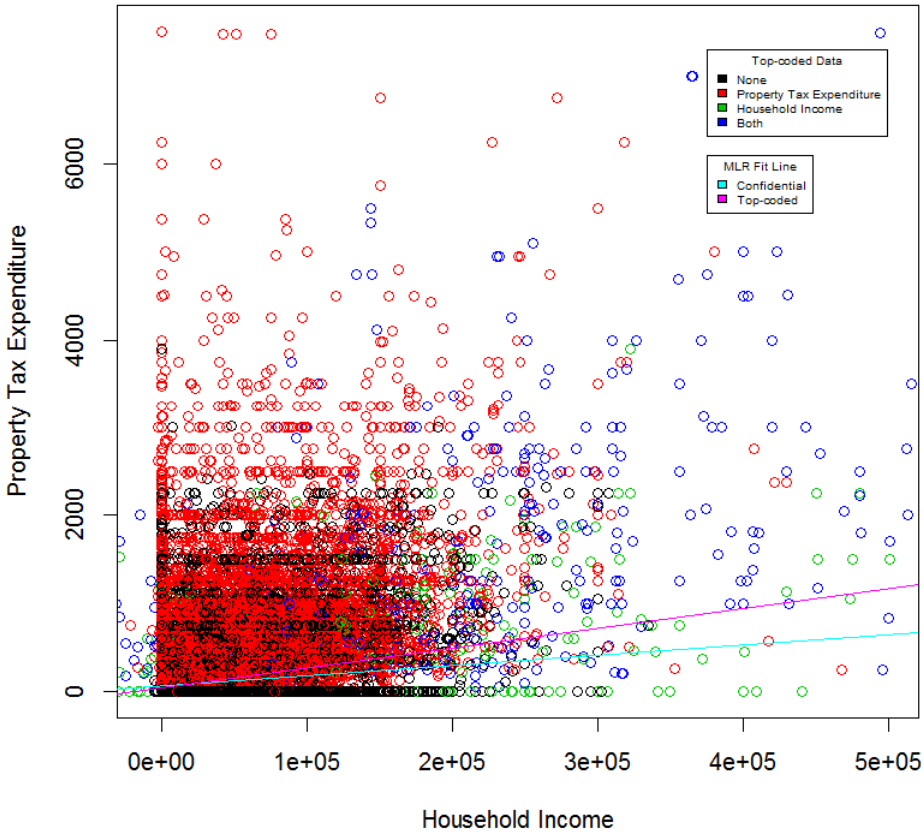

Figure 2: Property Tax Expenditure vs. Household Income

the statistics of interest is unknown. The idea of bootstrap is to use re-sampling of data to approximate the sampling distribution (Efron and Tibshirani 1993). Here are our bootstrap re-sampling steps:

1. At each year of data, we combined four quarterly interview data sets and re-sample the combined data set 1,000 times using a simple random sample with replacement (SRSWR). The bootstrap sample size is equal to the sample size of the original data.

2. For each SRSWR re-sample, we estimate the coefficient of household (HH) income $\left(\beta_1|_{\beta_2,,\ldots,\beta_p}\right)$ from both confidential data and from top-coded data.

3. The estimation of $\beta_1|_{\beta_2,,\ldots,\beta_p}$ is repeated for each of the four expenditures where property taxes, utilities, health care and domestic services, is the response variable, one at a time, within each SRSWR re-sample.

This bootstrap procedure generates 1,000 bootstrap estimates $\beta_1|_{\beta_2,,\ldots,\beta_p}$ to form the empirical distribution of the coefficient β_1 conditioned on the demographic variables. We also computed $SE\left(\beta_1|_{\beta_2,,\ldots,\beta_p}\right)$ to obtain a 95% bootstrap confidence interval (CI).

3.4 A Modified Kullback-Liebler (K-L) Divergence Measure

We divide the base of each bootstrap empirical distribution into 1000 equally spaced intervals, then use a kernel-smoothing to estimate the probability density function (pdf). Let $P(i)$ denote the estimated density of the i^{th} bootstrap coefficient estimate $\hat{\beta}_{1i}\Big|_{\beta_2,,\ldots,\beta_p}$, using the confidential data and let $Q(i)$ denote the estimated density using the bootstrapped top-coded data. The classic K-L divergence (Kullback and Leibler 1951) of Q from P,

$$D_{\mathrm{KL}}(P\|Q) = \sum_{i=1}^{n} \ln\left(\frac{P(i)}{Q(i)}\right) P(i),$$

gives us an average relative distance between confidential and top-coded empirical distributions.

In practice, it is entirely possible that for some i, $P(i) < Q(i)$, which would give a value $\ln\{P(i)/Q(i)\}P(i) < 0$ (Lock and Dunson 2014). This point would actually decrease the estimated distance even though $P(i)$ and $Q(i)$ are relatively far apart. For this reason we use a square term $\left(\ln\{P(i)/Q(i)\}\right)^2$ to compute a modified Kullback-Leibler divergence (K-L D2) estimate,

$$D_{\mathrm{KL2}}(P\|Q) = \sqrt{\sum_{i=1}^{n} \left(\ln\left(\frac{P(i)}{Q(i)}\right)\right)^2 P(i)}.$$

In the following section, we compare the empirical distributions of the bootstrap estimates of $\beta_1|_{\beta_2,,\ldots,\beta_p}$ between confidential and top-coded data in terms of visualized displays and K-L D2 estimates.

4. Comparison of Bootstrap Parameter Estimates

4.1 Scatter Cloud of $\hat{\beta}_1\Big|_{\beta_2,,\ldots,\beta_p}$ vs. $\tilde{\hat{\beta}}_1\Big|_{\beta_2,,\ldots,\beta_p}$

Consider there is a linear relationship between the top-coded coefficient and the confidential coefficient (Karr, Kohnen, Oganian, Reiter and Sanil 2006), such that:

$$\beta_{t-c} = b_0 + b_1\beta_c + \epsilon,$$

where c indexes confidential and $t-c$ indexes top-coded. In an ideal situation where top-coding has no effect on the numerical estimation, there will be $b_0 = 0, b_1 = 1$ and $\epsilon = 0$, but, obviously, the actual top-coded data will produce $b_0 \neq 0, b_1 \neq 1$ and $\epsilon \neq 0$. Therefore, we plot the bootstrap empirical distributions of regression coefficients of HH income from confidential data vs. top-coded data with respect to the response variables of the property tax, utilities, health care and domestic service expenditures, respectively for 2008-2011 (Figure 3 - Figure 6). We also add a 45° line (green) through the origin as a reference (intercept = 0), and estimated linear equations of β_{t-c} vs. β_c.

In Figure 3, we can see that for property tax models, in 2008 and 2009, the scatter cloud is through the 45° line, but 2010 and 2011 are way off (where 2011 is the worst). In Figure 4, we can see that for utility models, in 2008 the scatter cloud is close to parallel to the 45° line, in 2009, the scatter cloud is through the 45° line, but 2010 and 2011 are way off (2011 is the worst). In Figure 5, we can see that for health care models, in 2008, the scatter cloud is almost parallel to the 45° line, in 2009 the scatter cloud is close to parallel to the 45° line, but 2010 and 2011 are way off (2011 is the worst). In Figure 6, we can see that for

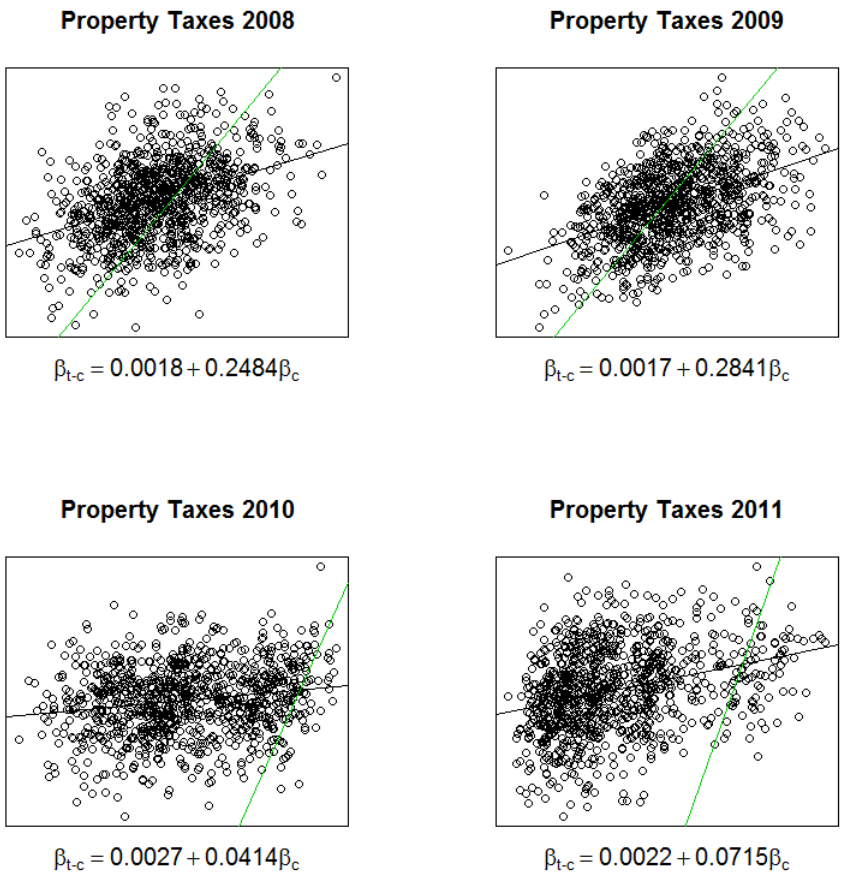

Figure 3: Property Tax Model: Bootstrap MLR $\beta_{\text{HH Income}}$

domestic service models, in 2008, the scatter cloud seems parallel to the $45°$ line, in 2009, the scatter cloud is through the $45°$ line, but in 2010 it is way off (the worst), and in 2011, the scatter cloud is through the $45°$ line again.

We have seen the slopes and variances, that is how top-coded points displace and spread, from the above scatter clouds visualizations of β_{t-c} vs. β_c. When the scatter cloud is far away from the $45°$ line, like domestic service in 2010, then the inaccuracy of top-coded coefficient estimates should raise concern. In the next subsection, we compute up K-L D2 estimates as a single measure to reflect those discrepancies observed in scatter clouds.

4.2 Results of K-L D2 Estimates and 95% Bootstrap CI

We present 95% CIs from the bootstrapped regression coefficients of confidential data vs. top-coded data and estimate K-L D2 distance between the two empirical distributions for each of the years 2008 through 2011 (Figure 7 - Figure 10). Figure 7 shows the 95% bootstrap CIs for the coefficients under multiple regression model where property tax is the response variable computed using the confidential and top-coded data sets respectively. We can see the CIs overlap in 2008 and 2009, but separate in 2010 and become further apart in 2011. In Figure 8, when utilities is the response variable, the 95% bootstrap CIs are close in 2008, almost overlap in 2009, then separate in 2010 and are further apart in 2011. When

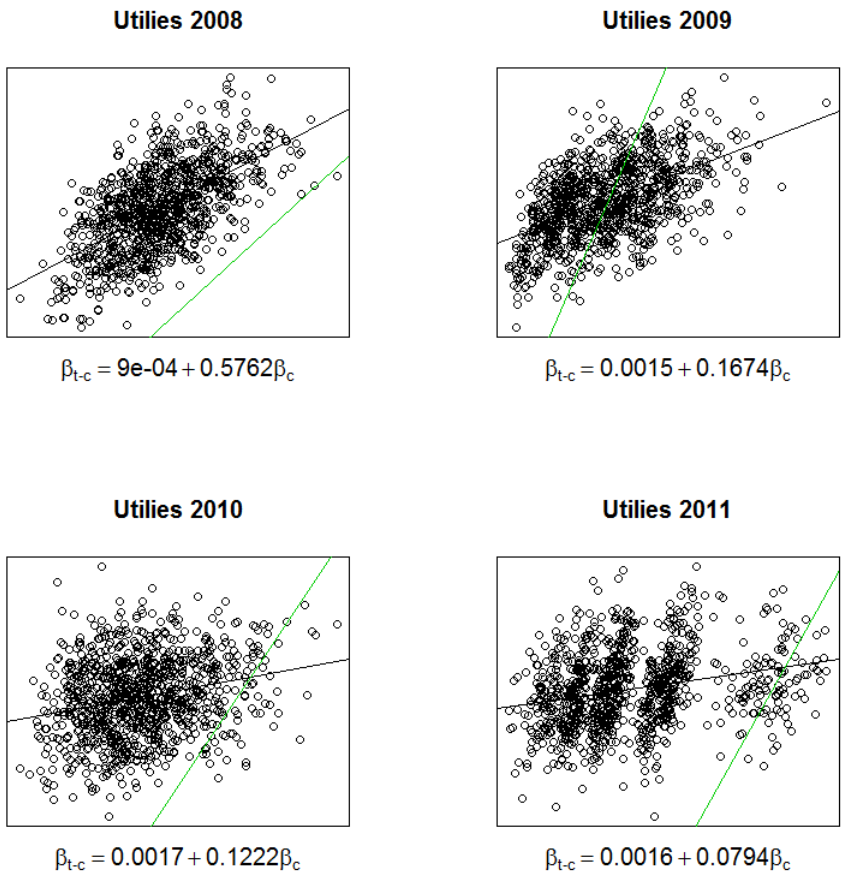

Figure 4: Utility Model: Bootstrap MLR $\beta_{\text{HH Income}}$

health care is the response variable (Figure 9) the 95% bootstrap CIs are close in 2008, start to separate in 2009, and become further apart in 2010 and 2011. In Figure 10, when domestic services is the response variable, the 95% bootstrap CIs are close in 2008, almost overlap in 2009, then become further apart in 2010, but then become close in 2011. In every case (four different utilities), the modified K-L D2 measure captured the deviation between confidential and top-coded household income coefficient for all four years, 2008-2011.

5. Conclusion

In this article, we study the relationship between regression 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 implement a kernel smoothing procedure to estimate the empirical distributions of coefficient estimates by computing the coefficients for several bootstrap samples. We adopt a modified Kullback-Liebler divergence data utility measure, K-L D2, to examine the numerical impact of top-coding on the utility of the CE microdata. Our study indicates that the modified K-L D2 measure provides promising results to reflect the effects of top-coding when compared to original data.

For future research steps, we would like to explore whether the modified K-L D2 could be scaled into a form of standardized indicator, so comparisons could be made between

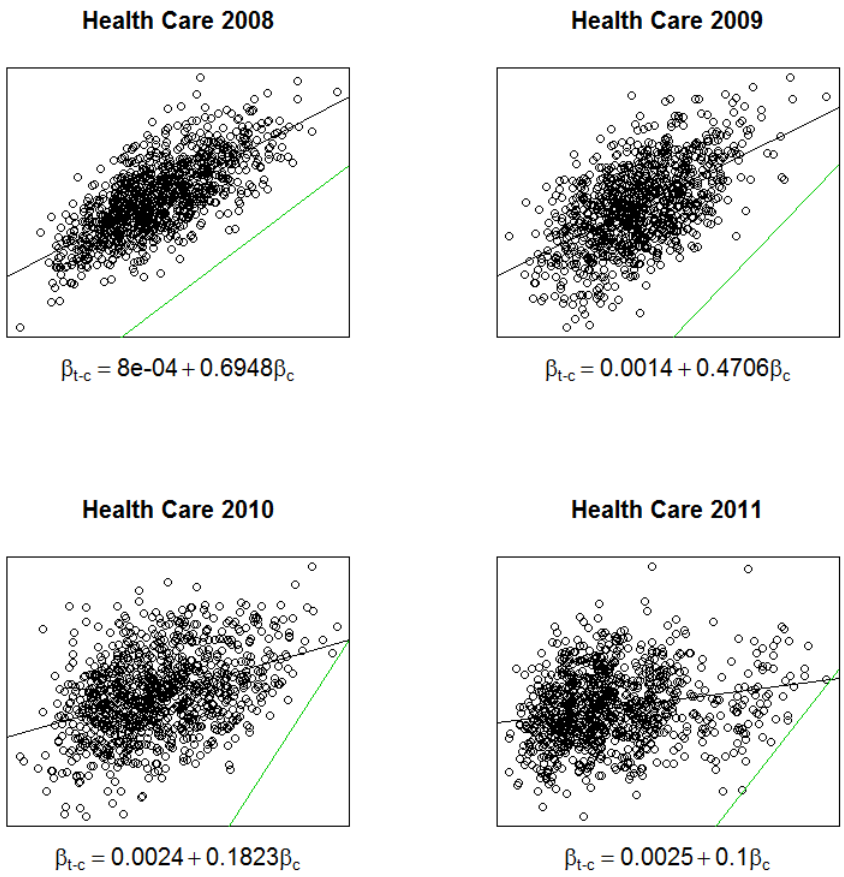

Figure 5: Health Care Model: Bootstrap MLR $\beta_{\text{HH Income}}$

different expenditures on a similar scale. This would help us establish a bootstrap procedure implementation for the program office. In addition, other economic indicators, such as the Gini Index and income elasticity (of consumption), are also of interest as seen in the economic and social science literature (Garner 1993, Landsburg 1999).

REFERENCES

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

Bureau of Labor Statistics. (2009). *BLS Handbook of Methods* (Chapter 16 Consumer Expenditures and Income). Available at 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.

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*. Available at 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.

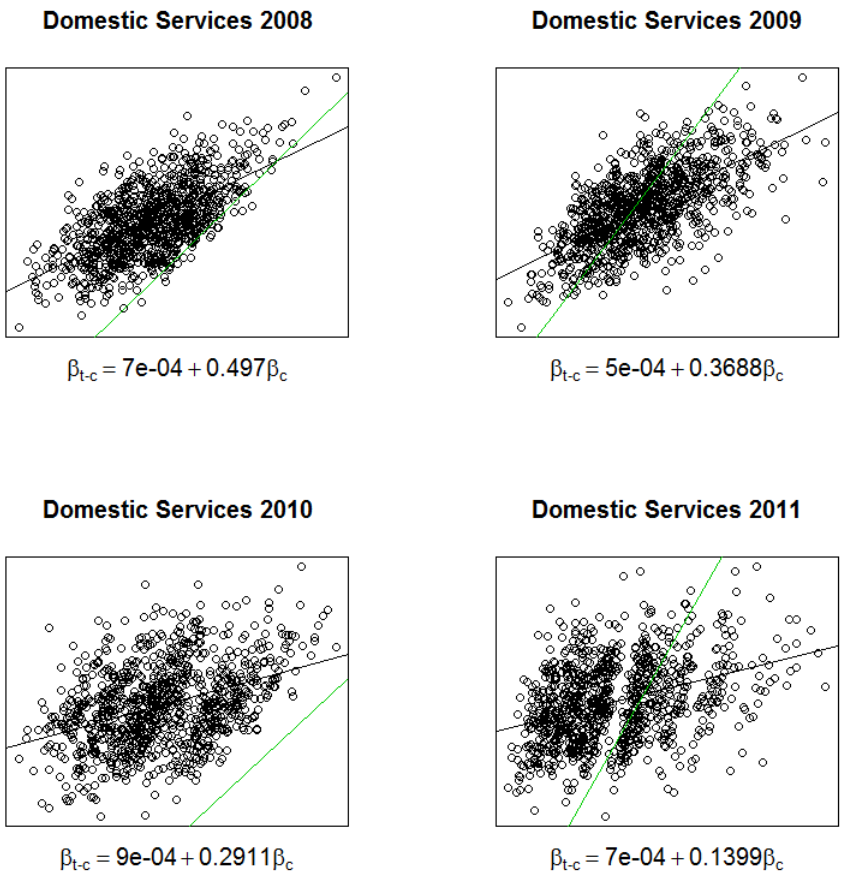

Figure 6: Domestic Service Model: Bootstrap MLR $\beta_{\text{HH Income}}$

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

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* (4th edition), South-Western College Publishing.

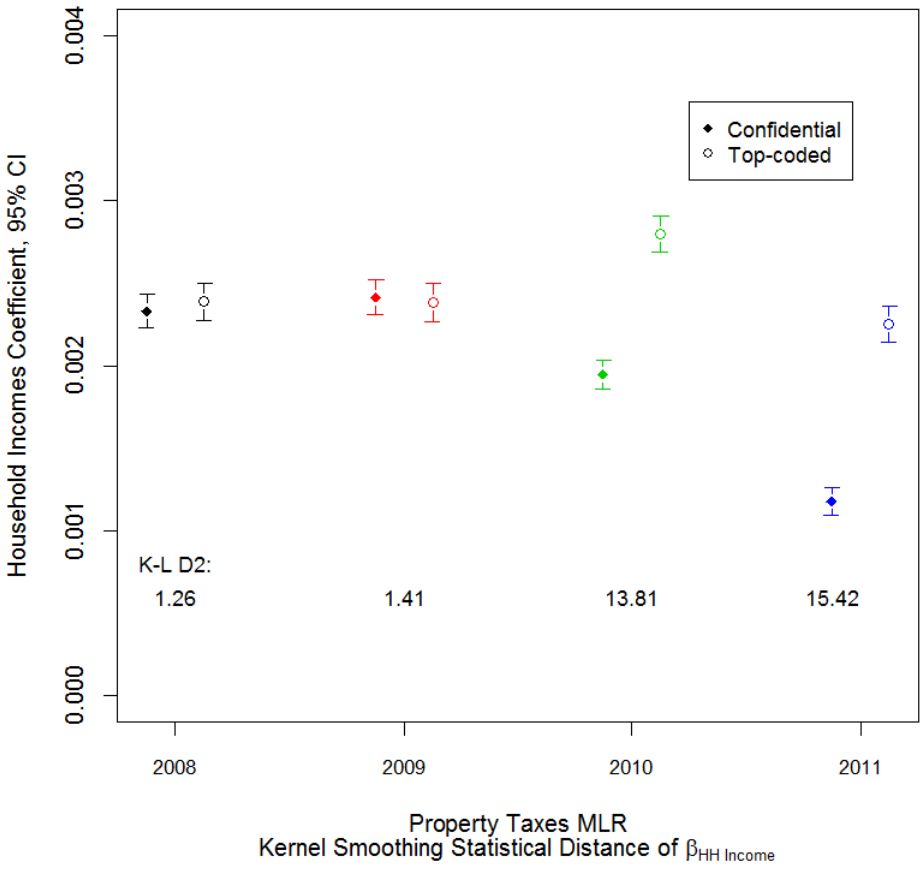

Figure 7: Property Taxes MLR Kernel Smoothing K-L D2 of $\beta_{\text{HH Income}}$

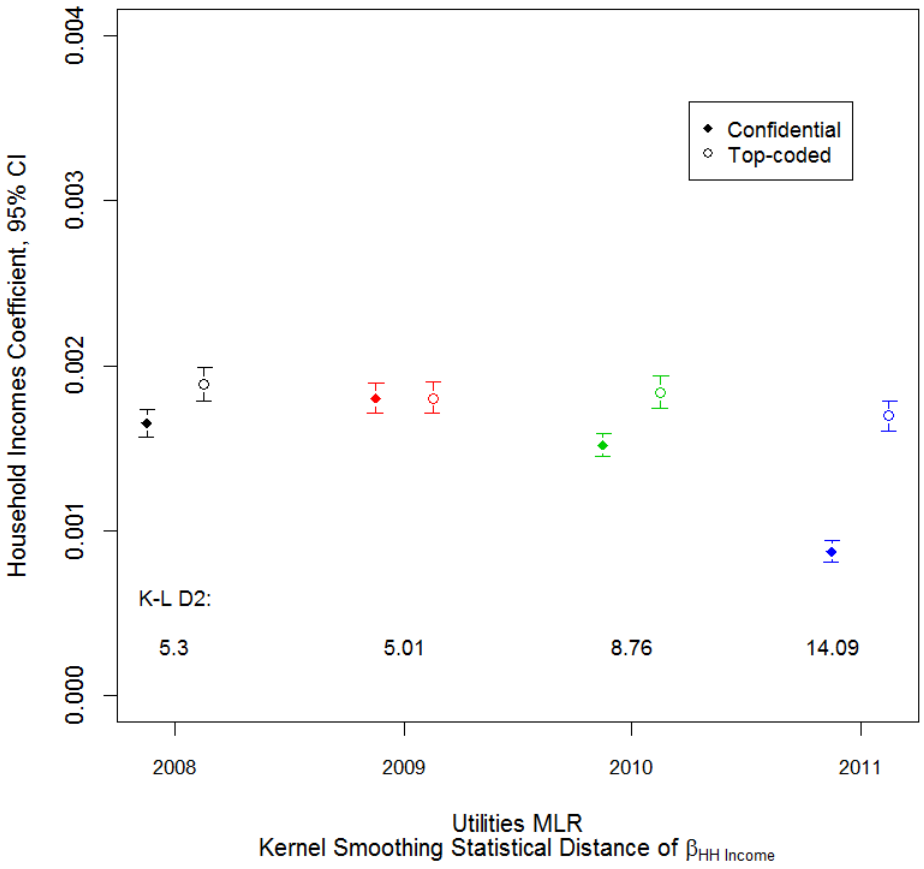

Figure 8: Utilities MLR Kernel Smoothing K-L D2 of $\beta_{\text{HH Income}}$

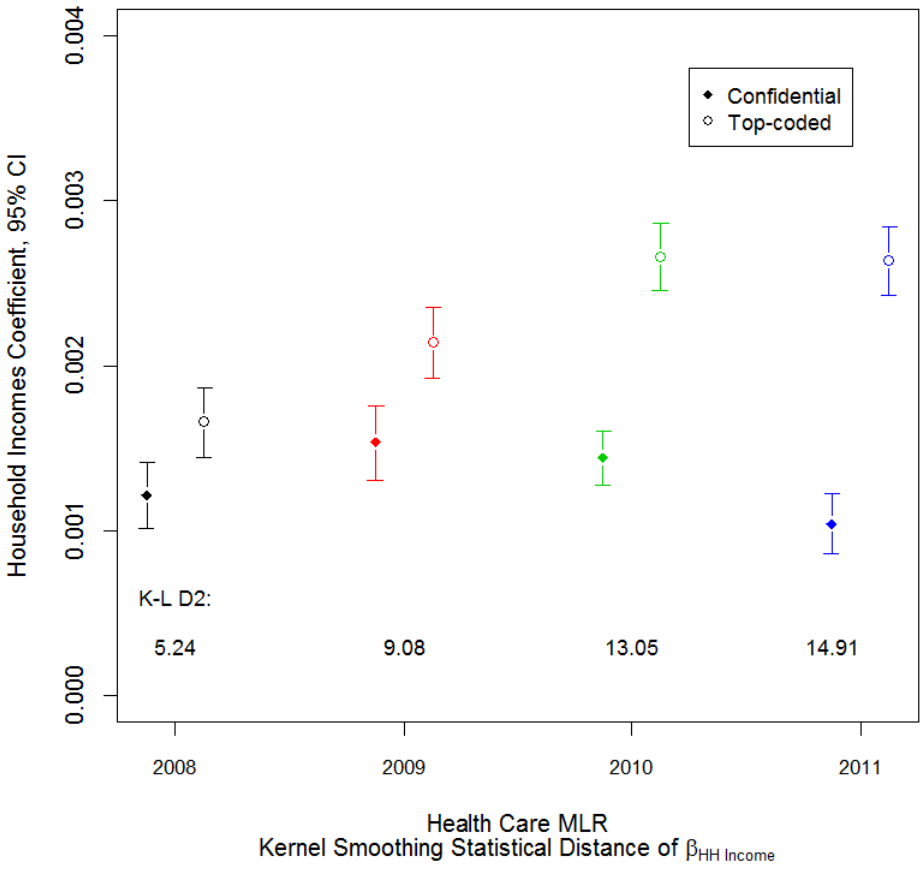

Figure 9: Health Care MLR Kernel Smoothing K-L D2 of $\beta_{\text{HH Income}}$

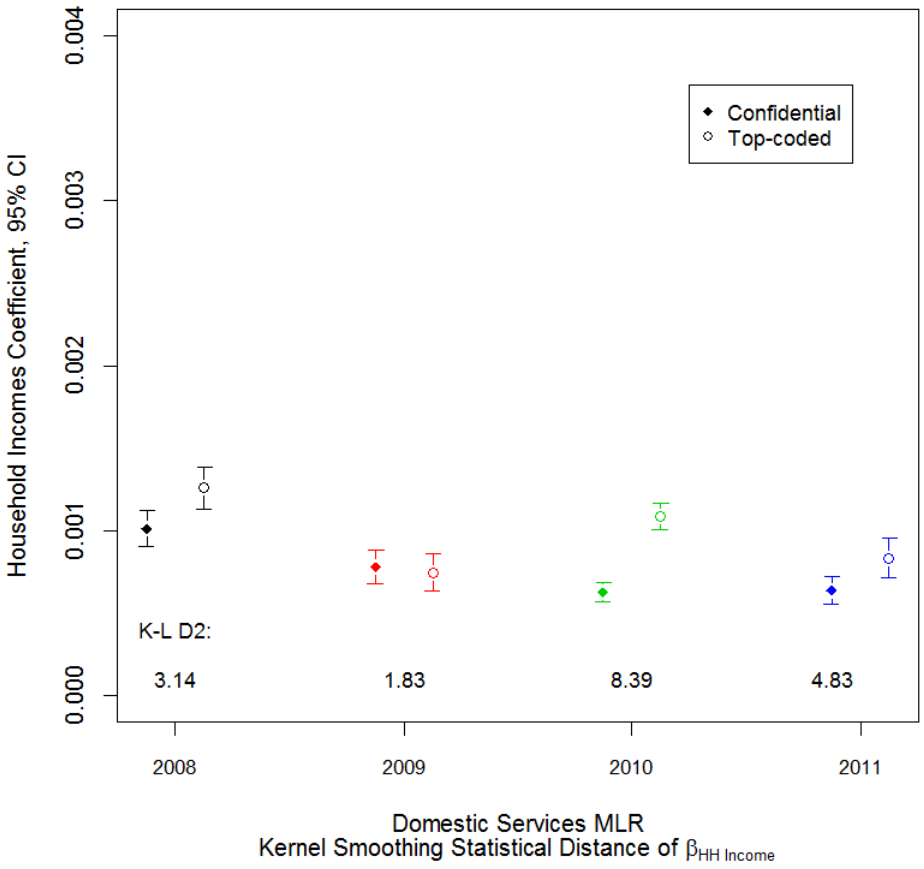

Figure 10: Domestic Service MLR Kernel Smoothing K-L D2 of $\beta_{\text{HH Income}}$