A Statistical Distance Approach to Evaluate CE Data Disclosure Utility

Daniel K. Yang and Daniell Toth
Office of Survey Methods Research
U.S. Bureau of Labor Statistics

Content does not represent BLS policy

Disclosure Limitation

- An essential mission of BLS: Disseminating public data on labor market activity, working conditions, and price changes.

- Balance: confidentiality vs. satisfactory data utility.

Consumer Expenditure Survey

- Consumer Expenditure (CE) Survey Collects information on the buying habits of U.S. consumers.
- Two components: Quarterly Interview Survey (focus of this study) and the Diary Survey
- Provides data on expenditures, income, and consumer unit (families and single consumers) characteristics.

BLS

Importance of CE

- The only Federal survey to provide information on the complete range of consumers' expenditures, incomes, and characteristics.
- Economic policymakers: examining the impact of policy changes on economic groups,
- Businesses and academic researchers: studying consumers' spending habits and trends,
- Bureau of Economic Analysis uses CE data to analyze rental equivalence, maintenance and repairs, and travel
- For Consumer Price Index: market basket of goods and services and their relative importance.

BLS

CE SDL Process

- Consumer Expenditure Survey microdata release requires statistical disclosure limitation (SDL).
- Objective: Conceal personally identifiable information (PII) to preserve the confidentiality and anonymity of survey participants.
- Production Process: "top-coding" and numerical impact.

BLS

Top-coding

Let y_α be the α percentile value for variable y.

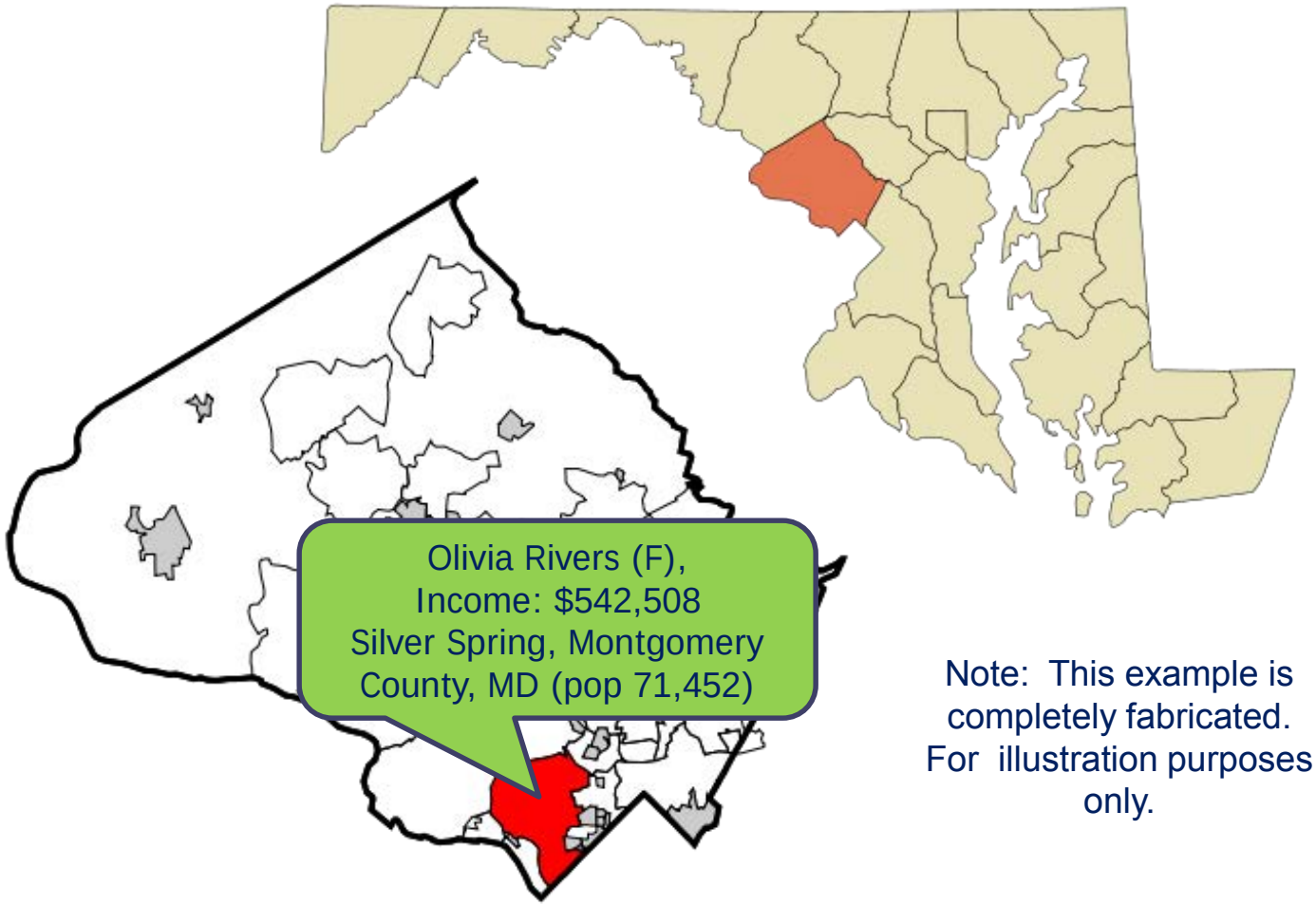
Then variable y is top-coded by replacing all values y_i in the set $A = \{y_i \mid y_i > y_\alpha\}$, by

$$y_i' = \frac{1}{|A|}\sum_{j \in A} y_j$$

This transformation preserves the mean of variable y.

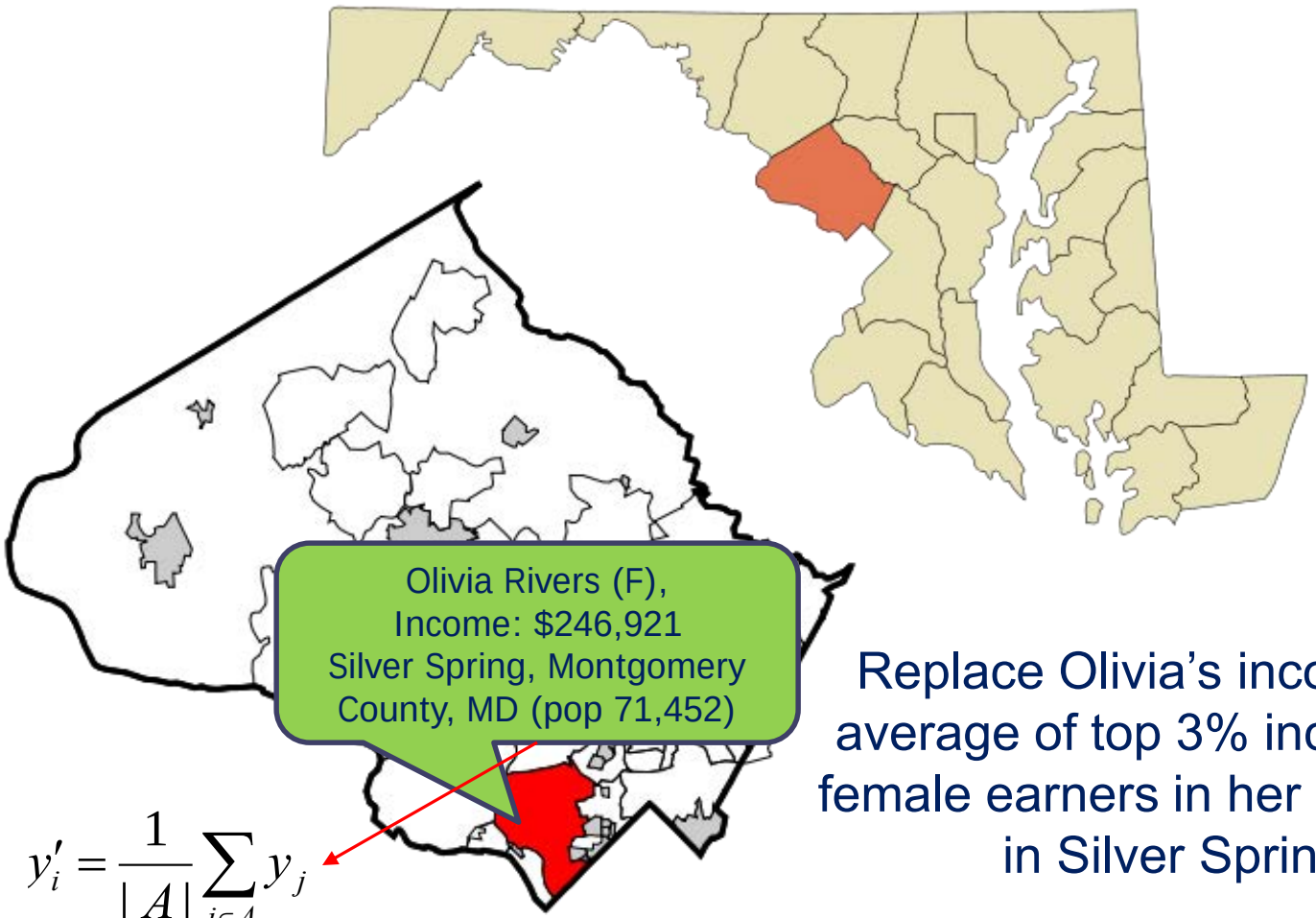
e.g. the reported annual income of a high income household might be replaced by the average of the annual income for all high-income household's in the region.

Olivia is an Outlier



Note: This example is completely fabricated. For illustration purposes only.

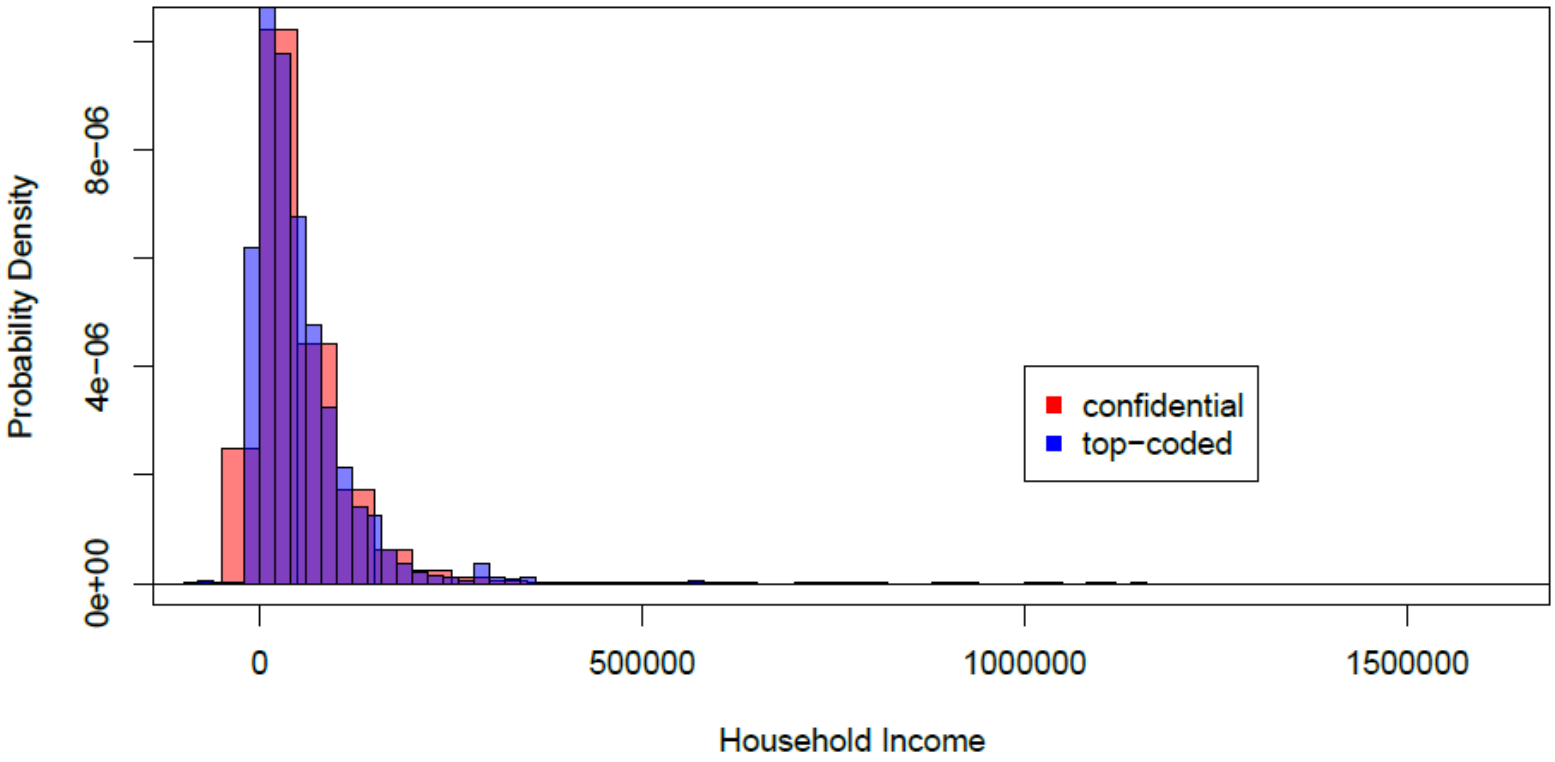
Top-Coding Example



$$y_i' = \frac{1}{|A|}\sum_{j \in A} y_j$$

Replace Olivia's income with average of top 3% incomes for female earners in her age group in Silver Spring.

Example: Overlap of Household Income



Data Utility Measures

- Analysis-specific:
  - Compare regression coefficients from confidential data vs. top-coded data for the same analysis.
  - Confidence Interval Overlap (IO) or Ellipsoid Overlap (EO)
- Global:
  - Compare propensity scores percentiles
  - Compare clusters in cluster analysis
  - Compare Empirical CDFs

BLS

Linear Models

- Topcoded: $\boldsymbol{y}^t = \boldsymbol{X}\boldsymbol{\beta}^t$, Confidential: $\boldsymbol{y}^c = \boldsymbol{X}\boldsymbol{\beta}^c$
- CE Data: 2008, 2009, 2010 and 2011 public released micro data and confidential data.
- Notation:
  - $\boldsymbol{y}^t$ – top-coded expenditure (property taxes, utilities, health care and domestic services)
  - $\boldsymbol{y}^c$ – confidential expenditure
  - $\boldsymbol{\beta}^t$, $\boldsymbol{\beta}^c$ – corresponding coefficients
  - $\boldsymbol{X}$ – covariates matrix (household income and other demographic variables for MLR) or household income for SLR

Demographic Variables

- 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
  - Gender

BLS

A Statistical Distance Measure

- Our contribution: a statistical distance based measure of bootstrap empirical distribution.
- Our goal:
  - To develop statistical distance measures that can be applied to expenditures model, e.g. multiple linear regression (MLR) or simple linear regression (SLR).
  - For program office to gauge how much changes had been made by top-coding.

Sufficient Statistics

- MLR: $\boldsymbol{y}_{n\times 1} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$, $T(y) = (\boldsymbol{y}^T\boldsymbol{y}, \boldsymbol{X}^T\boldsymbol{y})$ is the complete sufficient statistics.
- For SLR: $\hat{T}(y) = \left(\sum_{i=1}^{n} X_{1i}y_i,\ \sum_{i=1}^{n} y_i^2\right)$
- If assume $X_1 \sim N(\mu_1,\ \sigma_1^2)$, then in its complete sufficient statistics, $\sum_{i=1}^{n} X_{1i}^2$ seems to be a "spread-like" component.
- Consider a combined form of sufficient statistics:
- $\hat{T}(X, y) = \left(\sum_{i=1}^{n} X_{1i}y_i,\ \sum_{i=1}^{n} X_{1i}^2,\ \sum_{i=1}^{n} y_i^2\right)$

Bootstrap

- Sample is obtained from an unknown population and the sampling distribution of the statistics of interest is unknown.
- Use resampling of data to approximate the sampling distribution.
- Take m simple random samples of size n from $\{x_1, x_2, \ldots, x_n\}$ with replacement (SRSWR), for each resample, calculate $\hat{\theta}^* = T(\tilde{x}_R)$. The histogram of $\hat{\theta}^*$ values approximates the sampling distribution of $\hat{\theta}$.

Bootstrap of Sufficient Statistics

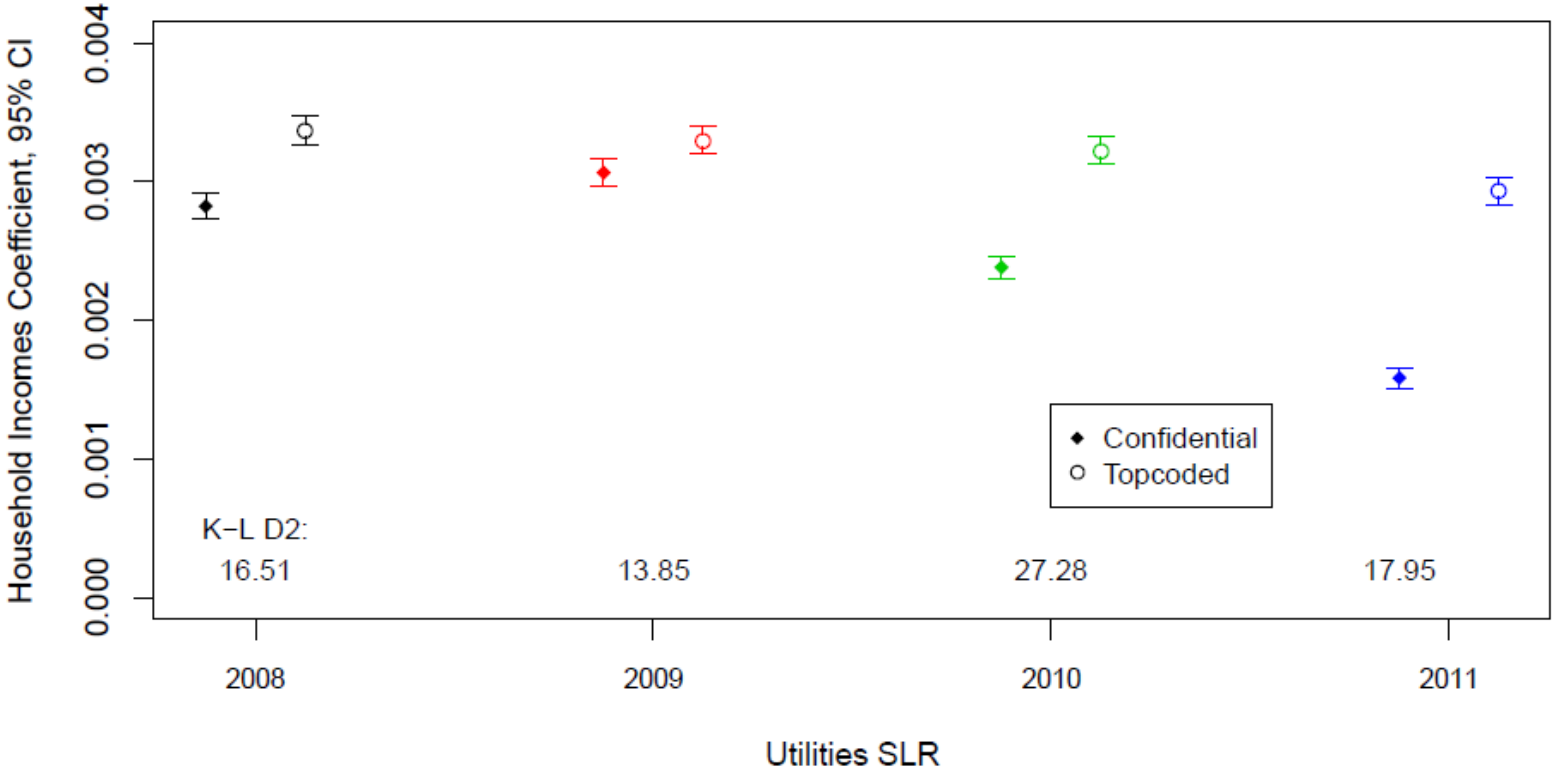
- For each data set from 2008 to 2011, we resampled 1,000 times of simple random sampling with replacement (SRSWR).
- At each SRSWR resample, a pair of confidential $\hat{T}(X, y)$ and top-coded $\hat{T}(\tilde{X}, \tilde{y})$ were computed.
- Bootstrap empirical distributions of confidential and top-coded sufficient statistics.

Modified Kullback–Leibler divergence

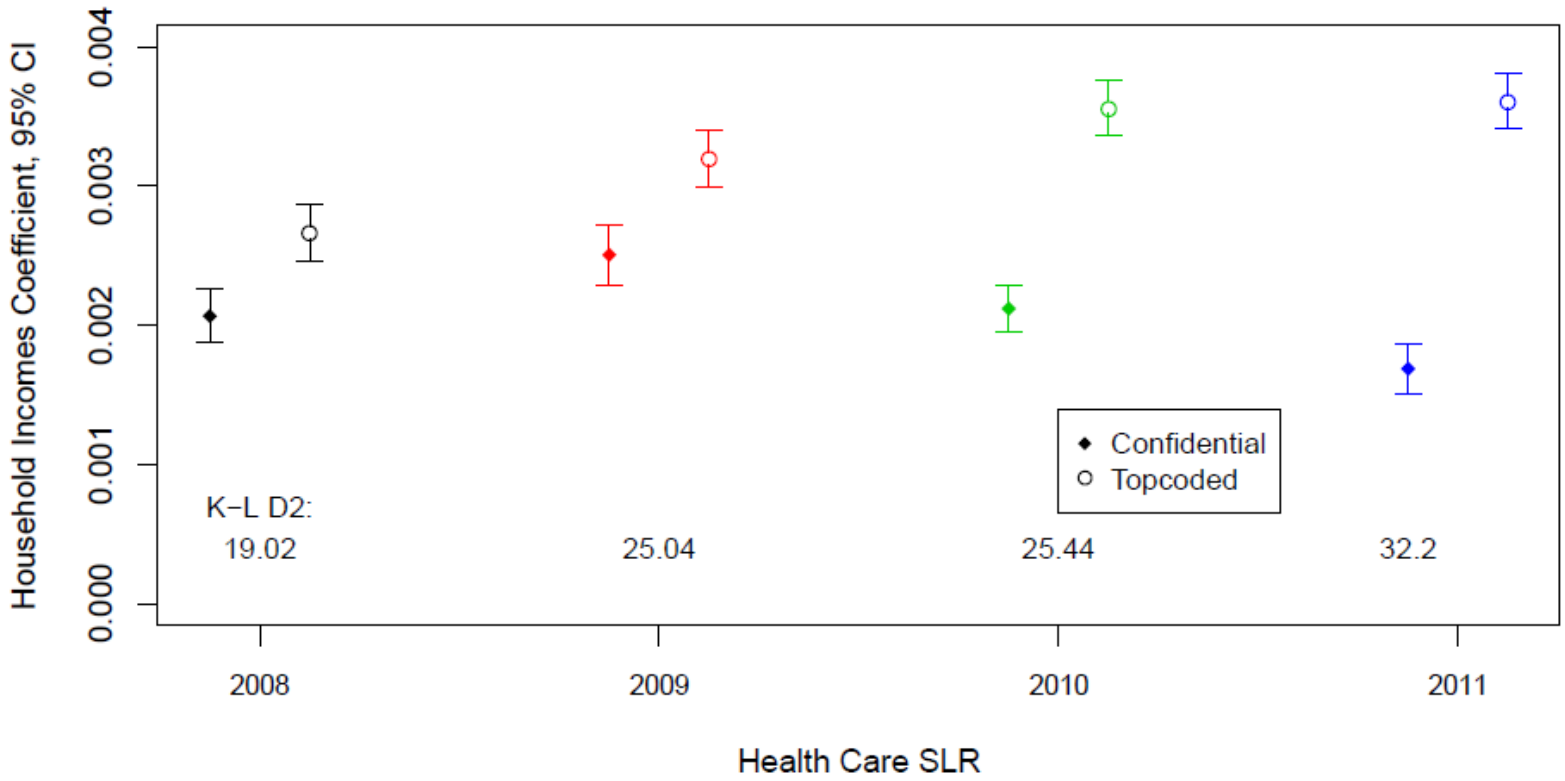
- Motivation: likelihood ratio between empirical distributions. Denote density functions:
- $P(i) = \widehat{\Phi}\left(\hat{T}(X, y)\right)$, $\mathrm{Q}(i) = \widehat{\widetilde{\Phi}}\left(\hat{T}(\tilde{X}, \tilde{y})\right)$
  - i – the ith resample/replicate
- $D_{KL2}(P||Q) = \sqrt{\sum_{i=1}^{n}\left(\ln\left(\frac{P(i)}{\mathrm{Q}(i)}\right)\right)^2 P(i)}$
- Estimate $\widehat{\Phi}$ and $\widehat{\widetilde{\Phi}}$ by Kernel-Smoothing (KS) numerical integration (base is equally divided).

SLR: Property Tax vs. Household Income

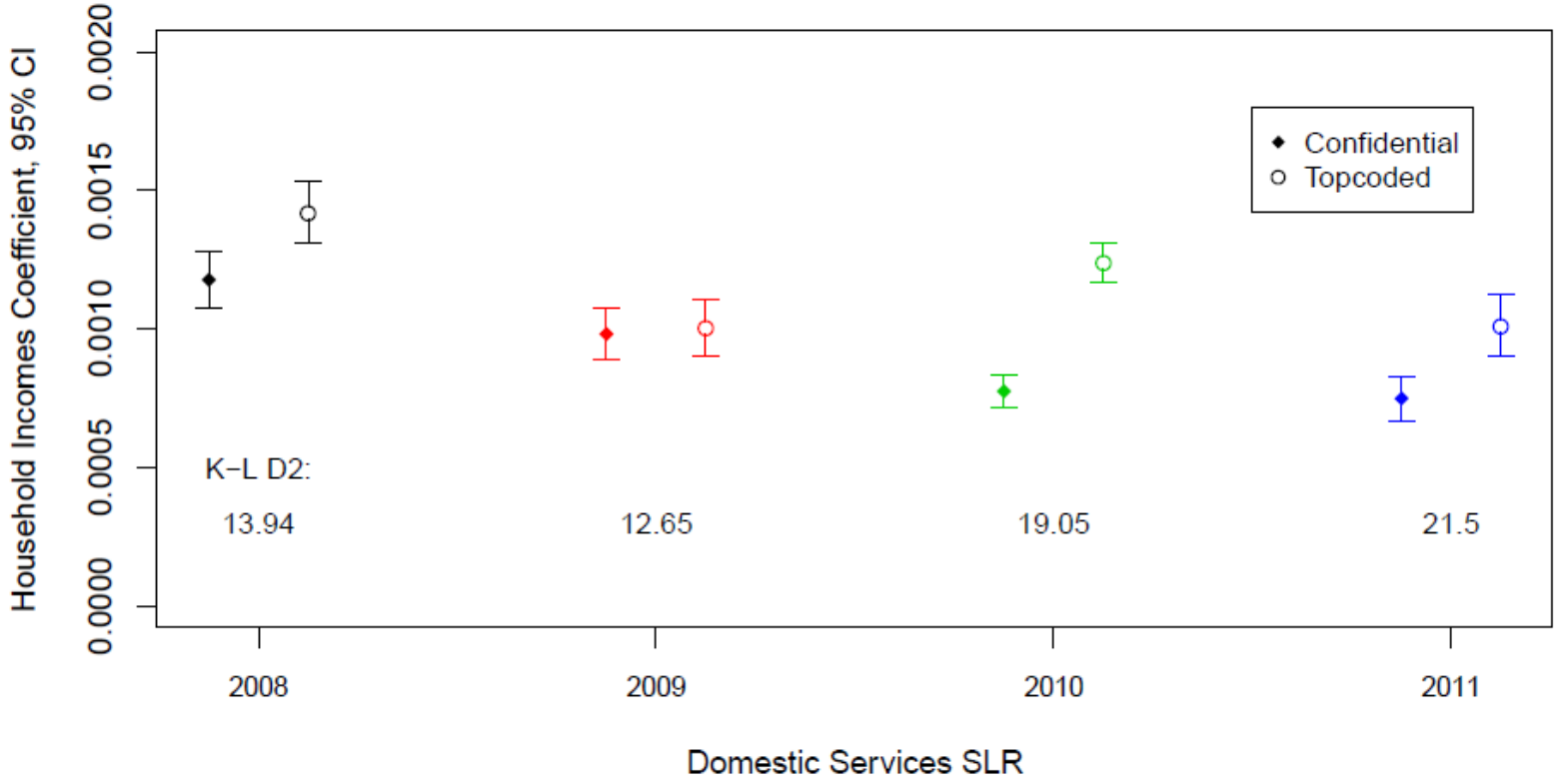
SLR: Utilities vs. Household Income



SLR: Health Care vs. Household Income



SLR: Domestic Services vs. Household Income



Limitations of Bootstrap Sufficient Statistics K-L D2

- None of them seem works perfectly.
- Sufficient statistics are under SLR model and unconditional.
- What about under a MLR model? Especially when one component must be conditioning on others?

Comparing Empirical Distribution of Bootstrap MLR $\boldsymbol{\beta}_{\textbf{HH Income}}$

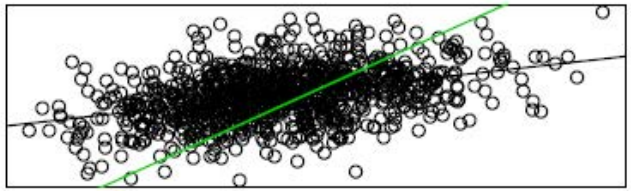
- MLR model: $y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$
- β_1: Household (HH) income
- $\beta_2, \cdots, \beta_p$: owing or not, region, household size, … , reference person's age, education, …
- For each data set from 2008 to 2011, we resampled 1,000 times of SRSWR
- At each SRSWR resample, a pair of confidential $\widehat{\beta_1}|_{\beta_2,\cdots,\beta_p}$ and top-coded $\widehat{\widetilde{\beta_1}}|_{\beta_2,\cdots,\beta_p}$ were computed.

Bootstrap of MLR $\beta_{\text{HH Income}}$

- Compare bootstrap empirical distributions between confidential (c) and top-coded ($t - c$) coefficients of HH income:
  - Plot $\beta_{t-c}\ vs.\ \beta_c$
  - Estimate K-L D2
- Consider: $\beta_{t-c} = b_0 + b_1\beta_c + \varepsilon$
- In a perfect world: $b_0 = 0$, $b_1 = 1$ and $\varepsilon = 0$
- But in reality: $b_0 \neq 0$, $b_1 \neq 1$ and $\varepsilon \neq 0$

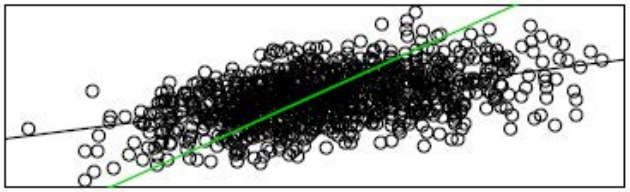
Bootstrap MLR $\beta_{\text{HH Income}}$: Property Tax

Property Taxes 2008



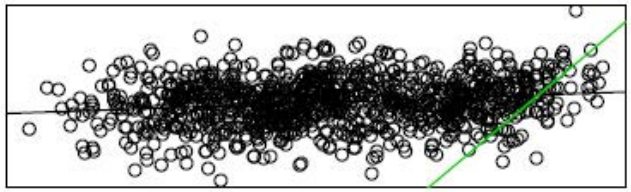
$\beta_{t-c} = 0.0018 + 0.2484\beta_c$

Property Taxes 2009



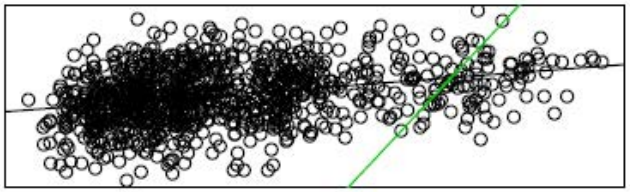
$\beta_{t-c} = 0.0017 + 0.2841\beta_c$

Property Taxes 2010



$\beta_{t-c} = 0.0027 + 0.0414\beta_c$

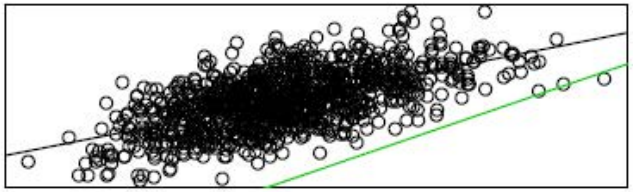
Property Taxes 2011



$\beta_{t-c} = 0.0022 + 0.0715\beta_c$

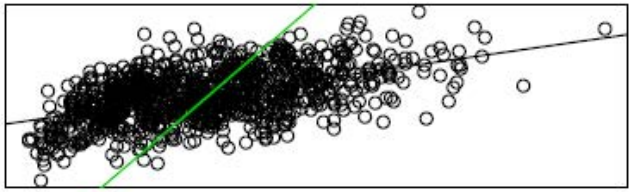
Bootstrap $\beta_{\text{HH Income}}$: Utilities

Utilies 2008



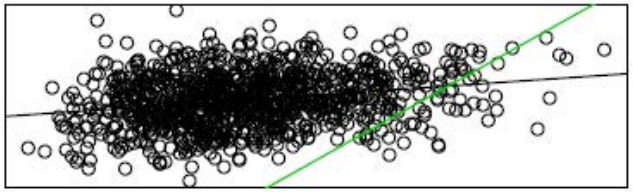
$\beta_{t-c} = 9e{-}04 + 0.5762\beta_c$

Utilies 2009



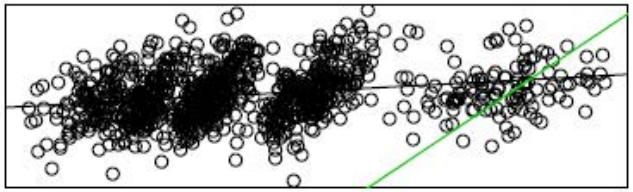
$\beta_{t-c} = 0.0015 + 0.1674\beta_c$

Utilies 2010



$\beta_{t-c} = 0.0017 + 0.1222\beta_c$

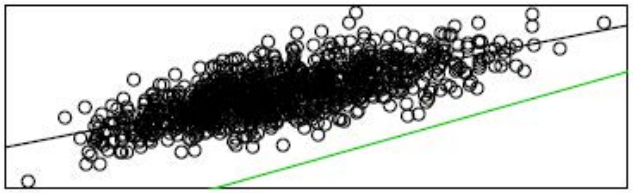
Utilies 2011



$\beta_{t-c} = 0.0016 + 0.0794\beta_c$

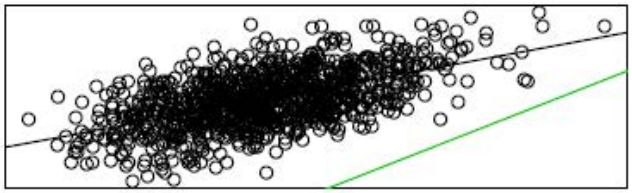
Bootstrap $\beta_{\text{HH Income}}$: Health Care

Health Care 2008



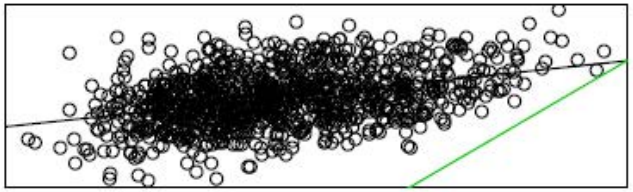
$\beta_{t-c} = 8e{-}04 + 0.6948\beta_c$

Health Care 2009



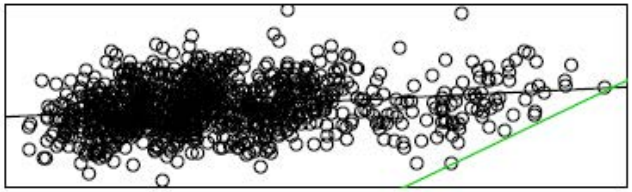
$\beta_{t-c} = 0.0014 + 0.4706\beta_c$

Health Care 2010



$\beta_{t-c} = 0.0024 + 0.1823\beta_c$

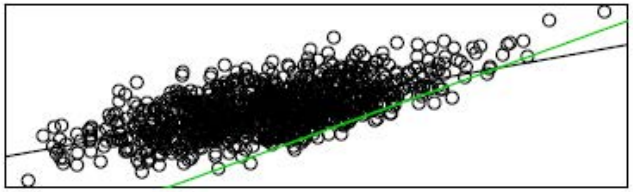
Health Care 2011



$\beta_{t-c} = 0.0025 + 0.1\beta_c$

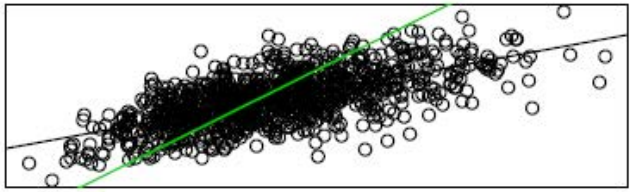
Bootstrap $\beta_{\text{HH Income}}$: Domestic Services

Domestic Services 2008



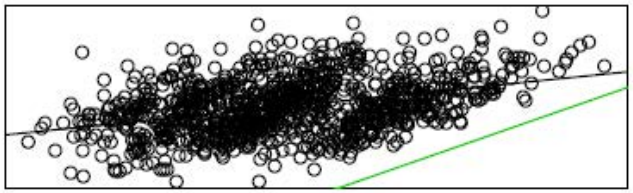
$\beta_{t-c} = 7e{-}04 + 0.497\beta_c$

Domestic Services 2009



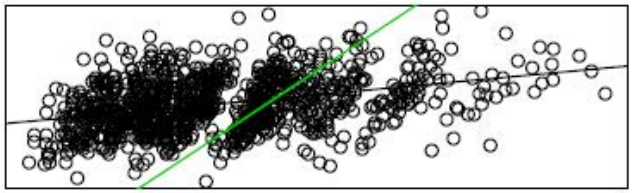
$\beta_{t-c} = 5e{-}04 + 0.3688\beta_c$

Domestic Services 2010



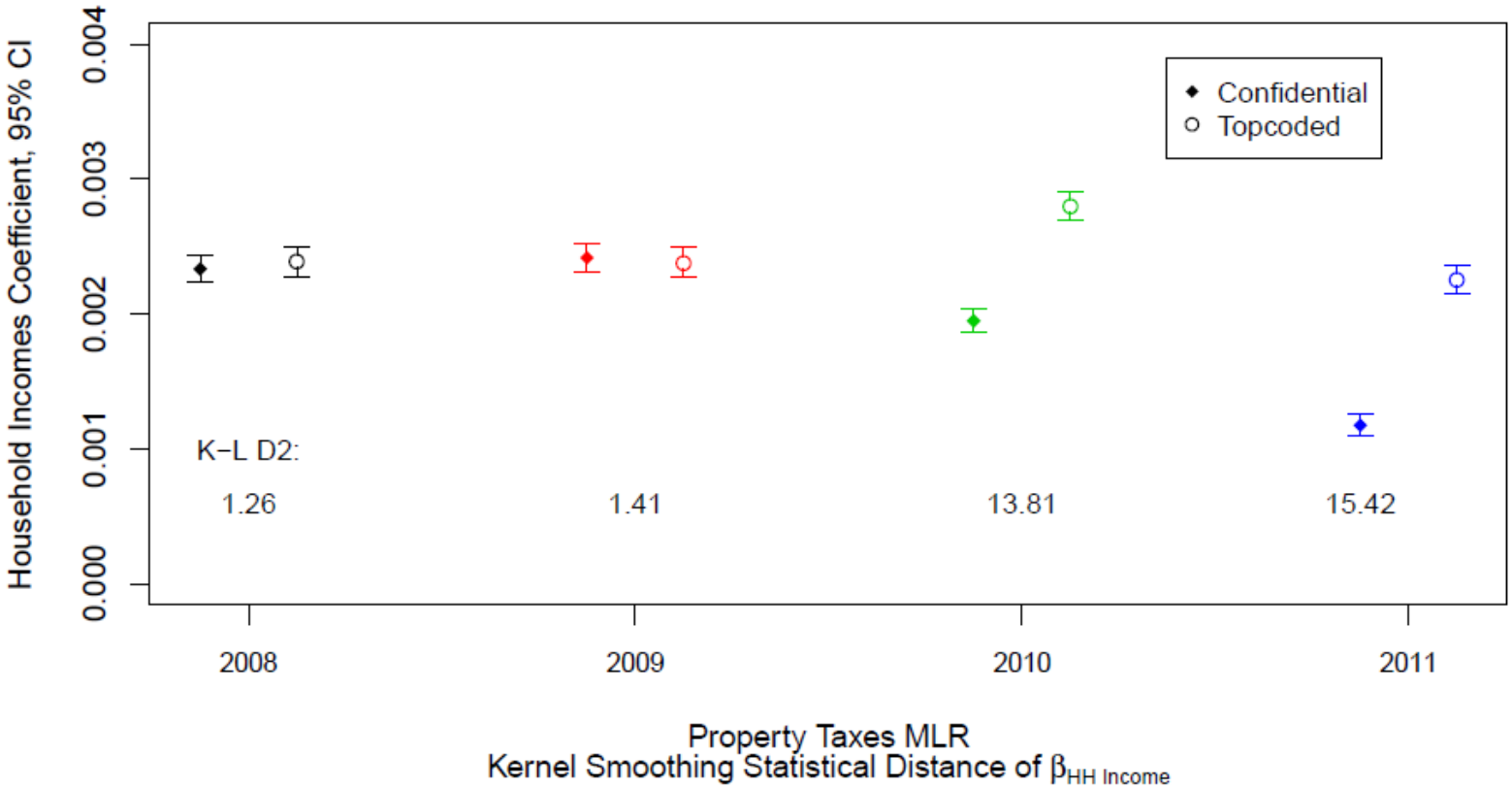
$\beta_{t-c} = 9e{-}04 + 0.2911\beta_c$

Domestic Services 2011

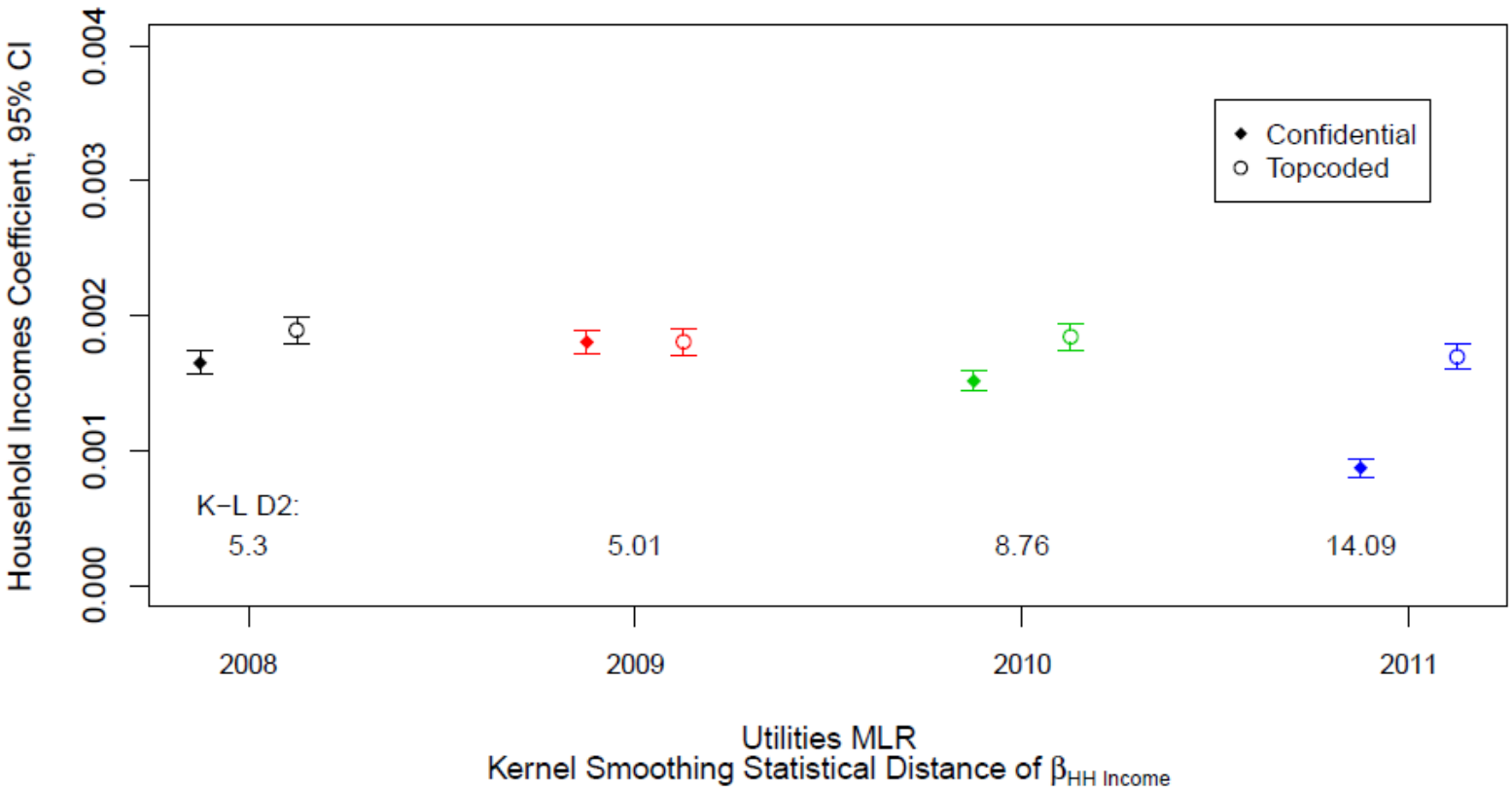


$\beta_{t-c} = 7e{-}04 + 0.1399\beta_c$

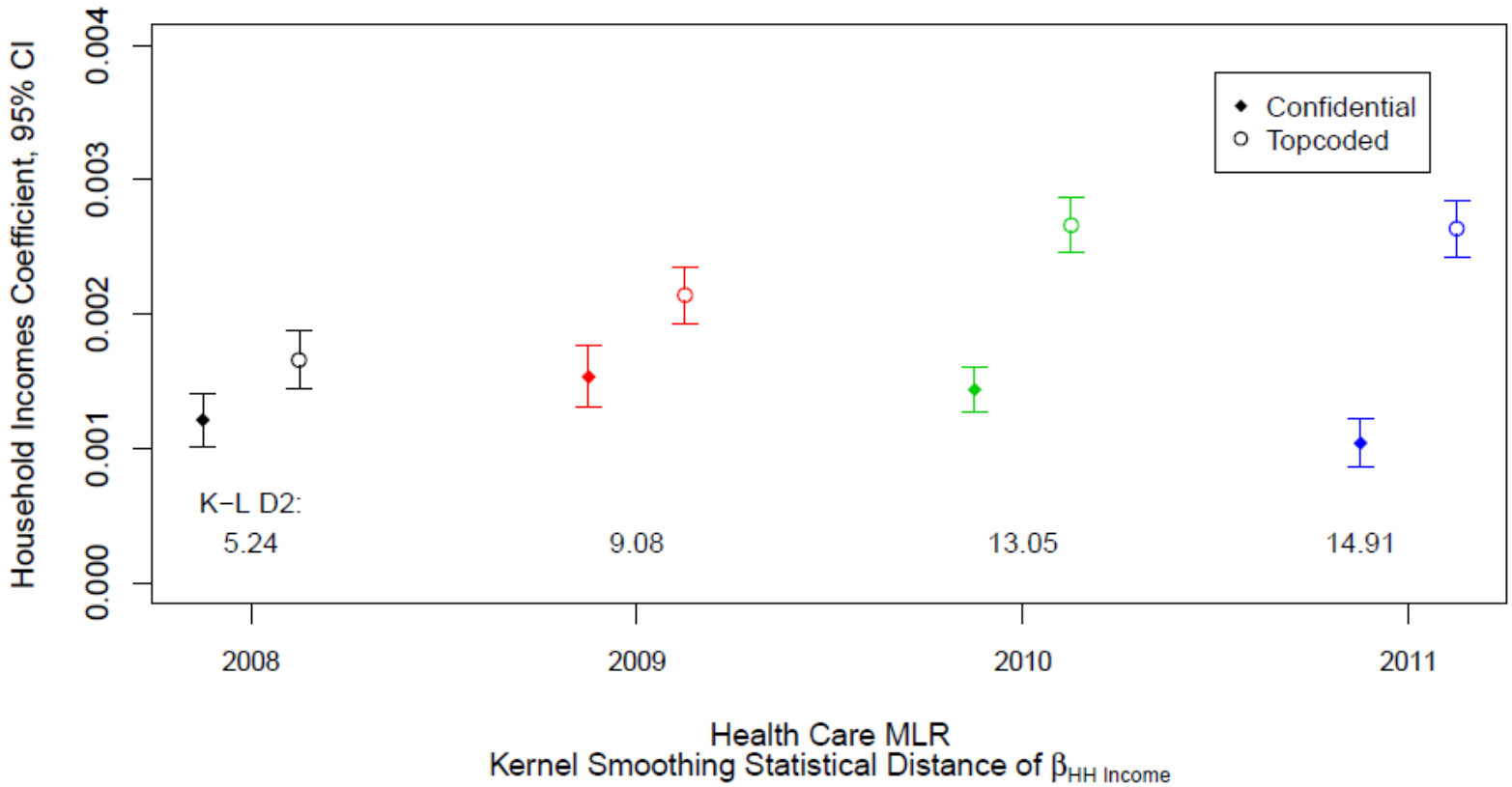
MLR: Property Tax vs. Household Income



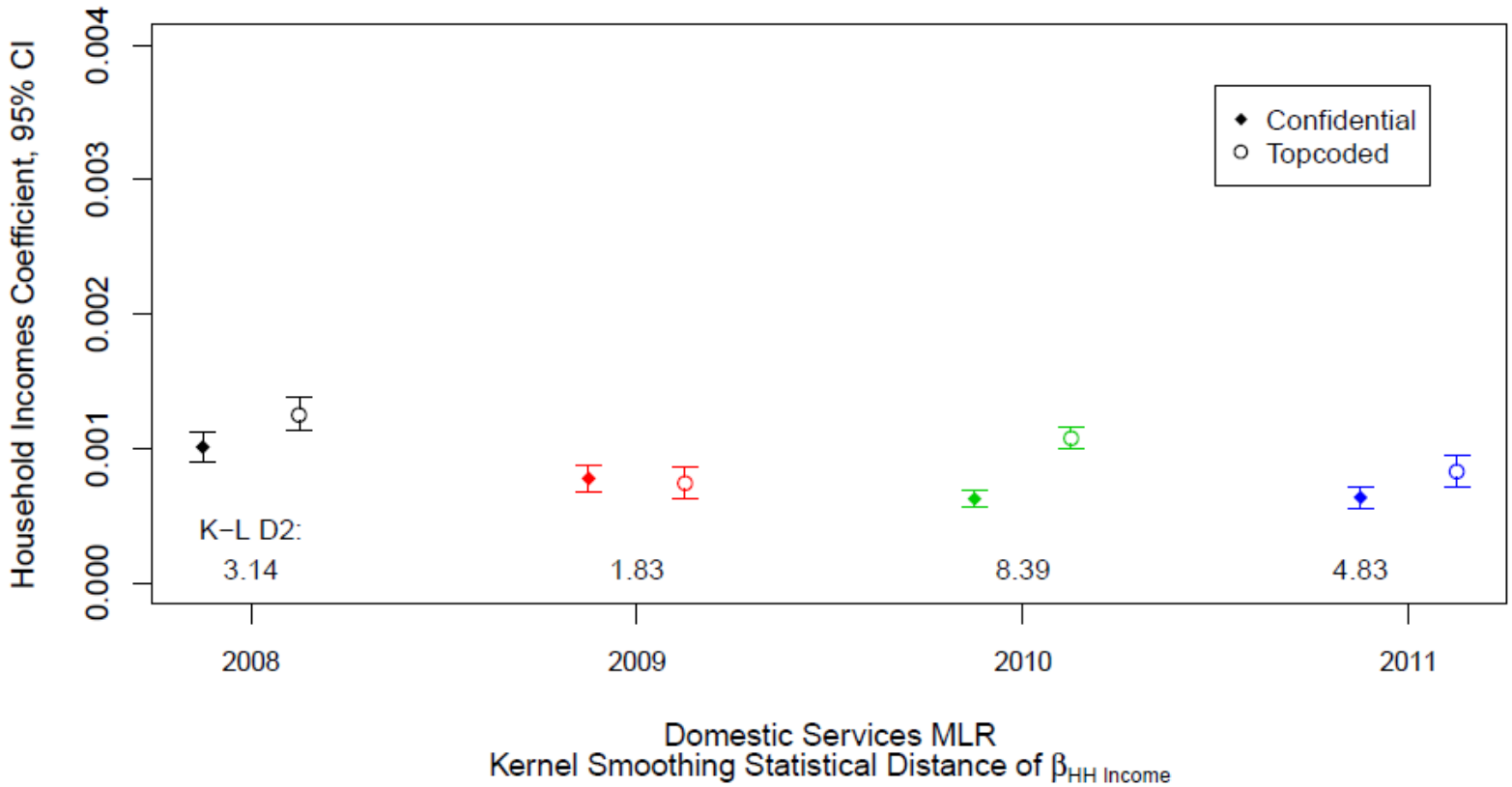
MLR: Utility vs. Household Income



MLR: Health Care vs. Household Income



MLR: Domestic Services vs. Household Income



Summary

- Sufficient statistics based distributional distance does not seem to work perfectly.
- Bootstrap empirical distribution of coefficients approach seems promising to reflect the numerical impact of top-coding when compared to original data (statistics distance vs. deviation of parameter of interest).
- Future works:
  - Could the modified K-L D2 be scaled into a sort of standardized indicator?
  - Bootstrap Procedure implementation for program office?
  - How does the modified K-L D2 affect the shape of $\beta_{t-c}\ vs.\ \beta_c$?

THANK YOU!

Contact Information

Daniel K. Yang
Research Mathematical Statistician
Office of Survey Methods Research (OSMR)
www.bls.gov/osmr/home.htm
yang.daniel@bls.gov

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

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR