Indicator for Top-Coding Effects on a Household Survey Income Elasticity of Demand Estimates

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

FCSM Government Advances in Statistical Programming Workshop

October 24, 2018

Overview

- Introduction: Consumer Expenditure Surveys (CE) and Statistical disclosure limitation (SDL)
- Data utility and economics model
- Impact on Income Elasticity of Demand and Proportion Ratio Indicator:
  - Log linear regression model for expenditures
  - Propensity model for consumption
- Conclusion

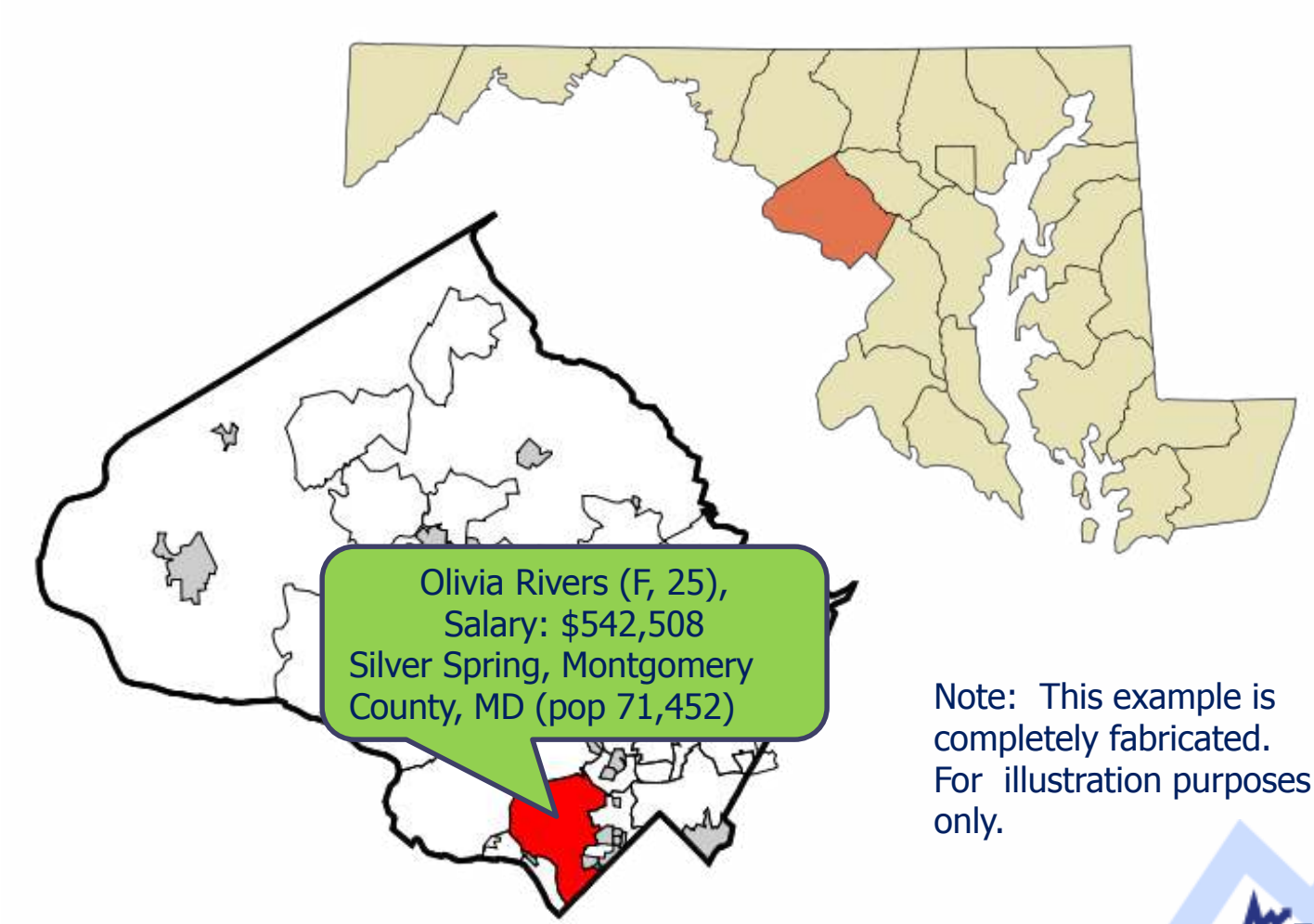
Consumer Expenditure Survey

- Consumer Expenditure Survey (CE) Collects information on the buying habits of U.S. consumers.
- CE's goal: Provides data on expenditures, income, and consumer unit (families and single consumers) characteristics.
- Balance: confidentiality vs. satisfactory data utility.

CE SDL Process

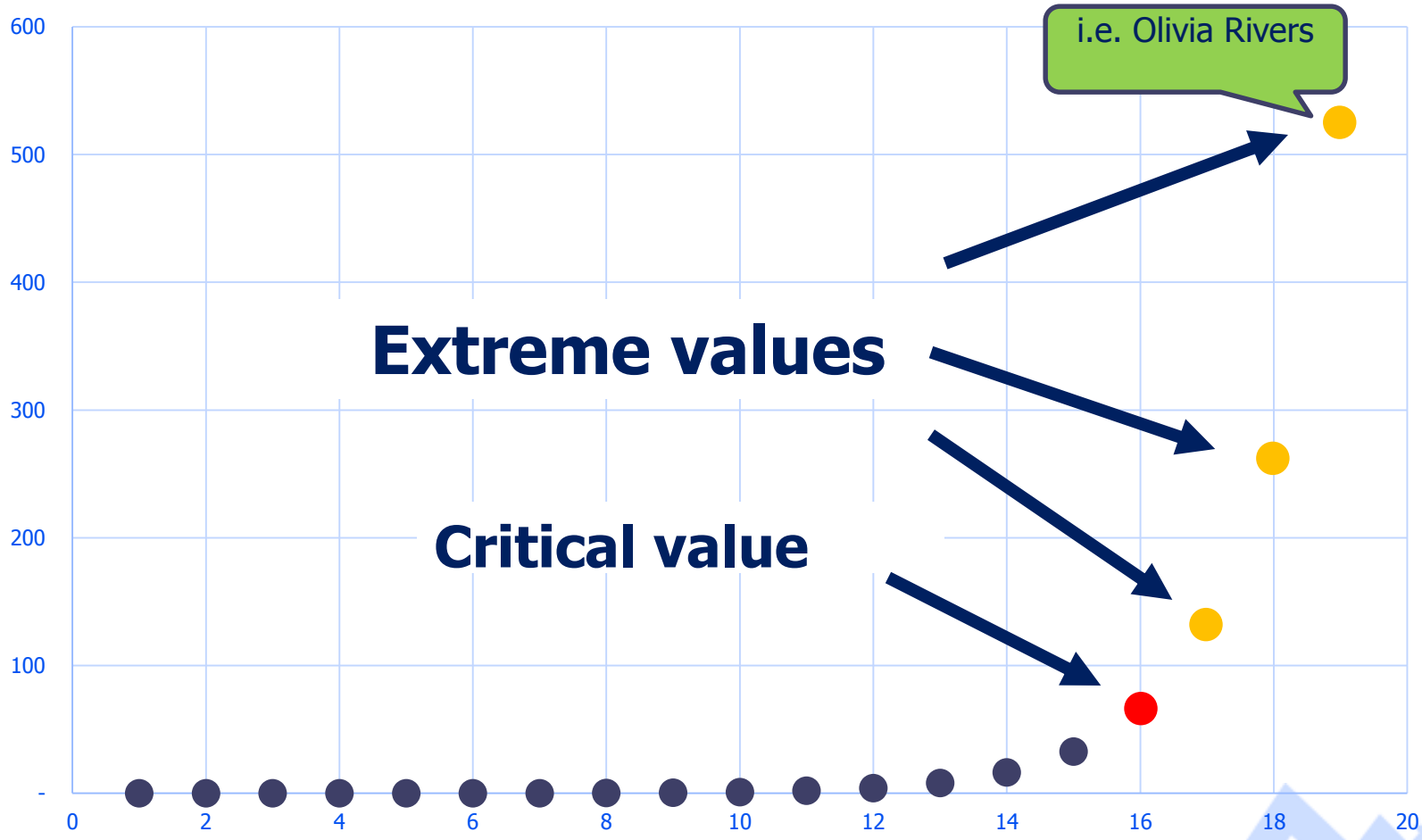
- CE microdata release requires statistical disclosure limitation (SDL).
- Objective: Conceal personally identifiable information (PII) to preserve the confidentiality and anonymity of survey participants.
- Production Process: “top-coding” and numerical impact.

Top-coding



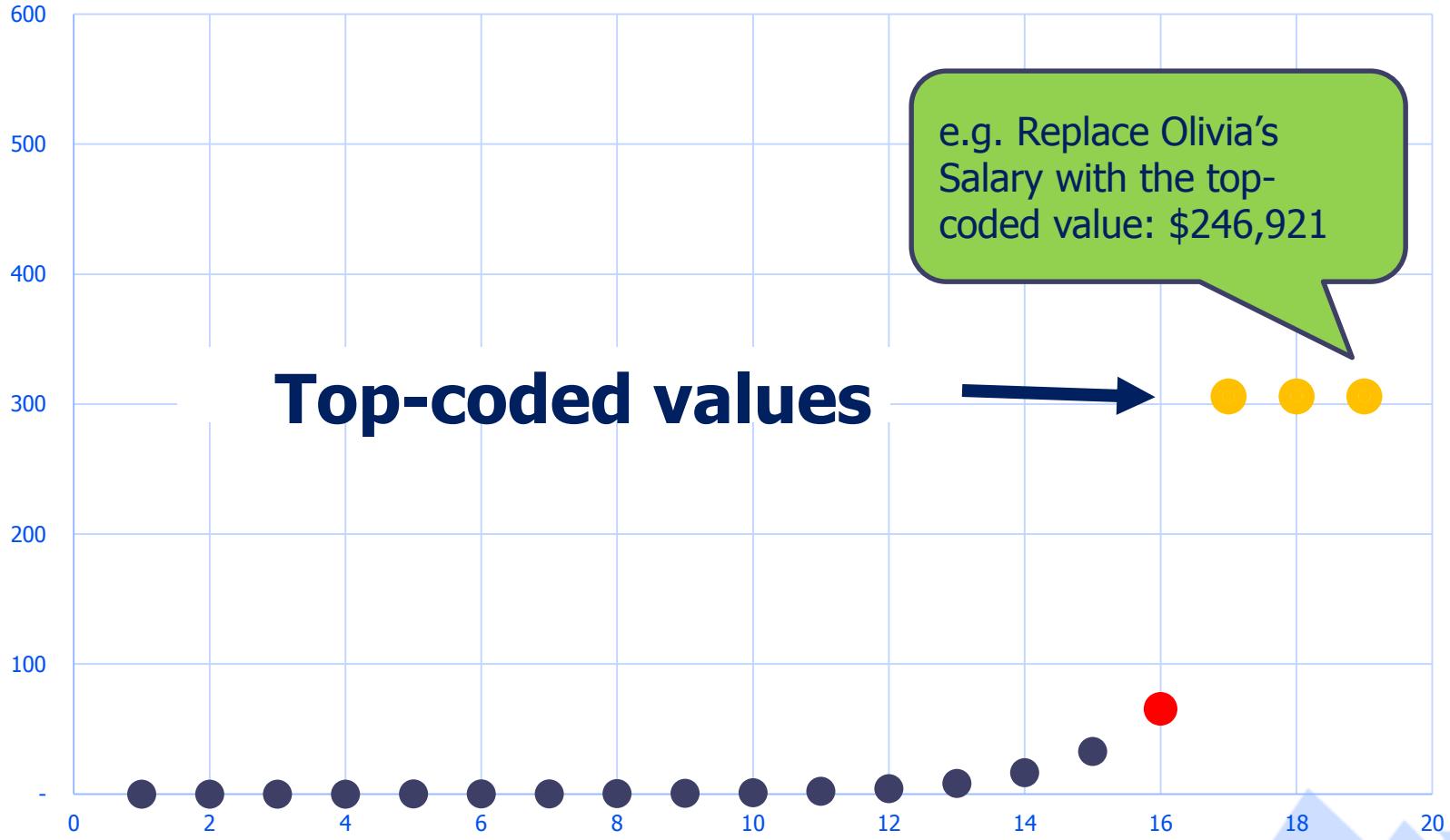
Note: This example is completely fabricated. For illustration purposes only.

Top-coding Illustration



Source: *Balancing Respondent Confidentiality and Data User Needs*, Aaron E. Cobet, BLS, 2014 CE Microdata User's Workshop.

Top-coding Illustration (cont.)



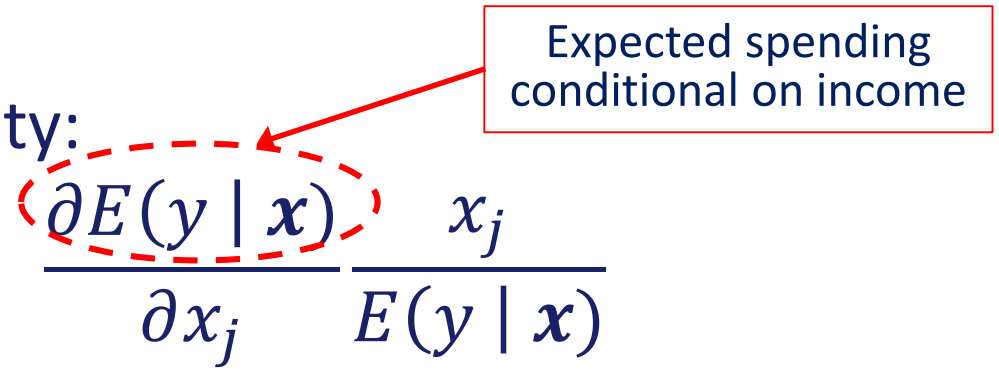
Source: *Balancing Respondent Confidentiality and Data User Needs*, Aaron E. Cobet, BLS, 2014 CE Microdata User's Workshop.

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 score percentiles.
  - Compare clusters in cluster analysis.
  - Compare Empirical Cumulative Density Functions (CDF's).

Economics Model: Income Elasticity of Demand

- Income Elasticity:



- Here, y – Expenditure, $\boldsymbol{x}$ – covariates, x_j - household income.
- Used in papers including: Altonji and Villanueva 2003, Bonisa and Silvestrinia 2012, Harris and Blisard 2002, Iacoviello 2011, Keen 1986, Kumhof and Laxton 2013, Tsekeris 2012, Weagley and Huh 2004.

Cragg's Double-Hurdle Model

- The unconditional expectation is

$$E(y|\boldsymbol{x}) = P(y > 0|\boldsymbol{x})E(y|\boldsymbol{x}, y > 0) + P(y = 0|\boldsymbol{x})E(y|\boldsymbol{x}, y = 0)$$
$$= P(y > 0|\boldsymbol{x})E(y|\boldsymbol{x}, y > 0) + 0$$
$$= P(y > 0|\boldsymbol{x})E(y|\boldsymbol{x}, y > 0).$$

Probability of spending conditional on income: logistic regression model

Expected amount of spending conditional on income: log linear regression model

$$E(y|\boldsymbol{x}) = P(y > 0|\boldsymbol{x})\, E(y|\boldsymbol{x}, y > 0)$$

- Assume a Logistic propensity model of consumption:

$$P(y > 0 \mid \boldsymbol{x}) = \Psi(\boldsymbol{x}\gamma) = \frac{e^{x\gamma}}{1 + e^{x\gamma}}$$

- If assume the outcome follows:

$$\log(y_i) \mid y_i > 0 = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \mid \boldsymbol{x}_i \sim N(0, \sigma^2)$$

$$\text{where } \boldsymbol{x}_i = \left[1, \log\left(x_j\right), \boldsymbol{x}_{i,k\neq j}\right]$$

- Unconditional expectation of $E(y \mid \boldsymbol{x})$ is

$$E(y \mid \boldsymbol{x}) = \Psi(\boldsymbol{x}\gamma)\exp(\boldsymbol{x}\boldsymbol{\beta} + \sigma^2/2)$$

Income Elasticity of Demand

$$\tau_{x_j} = \frac{\partial E(y \mid \boldsymbol{x})}{\partial x_j} \frac{x_j}{E(y \mid \boldsymbol{x})} = \gamma_j [1 - \Psi(\boldsymbol{x}\gamma)] x_j + \beta_j$$

coefficient from logistic model

coefficient of income from log linear model

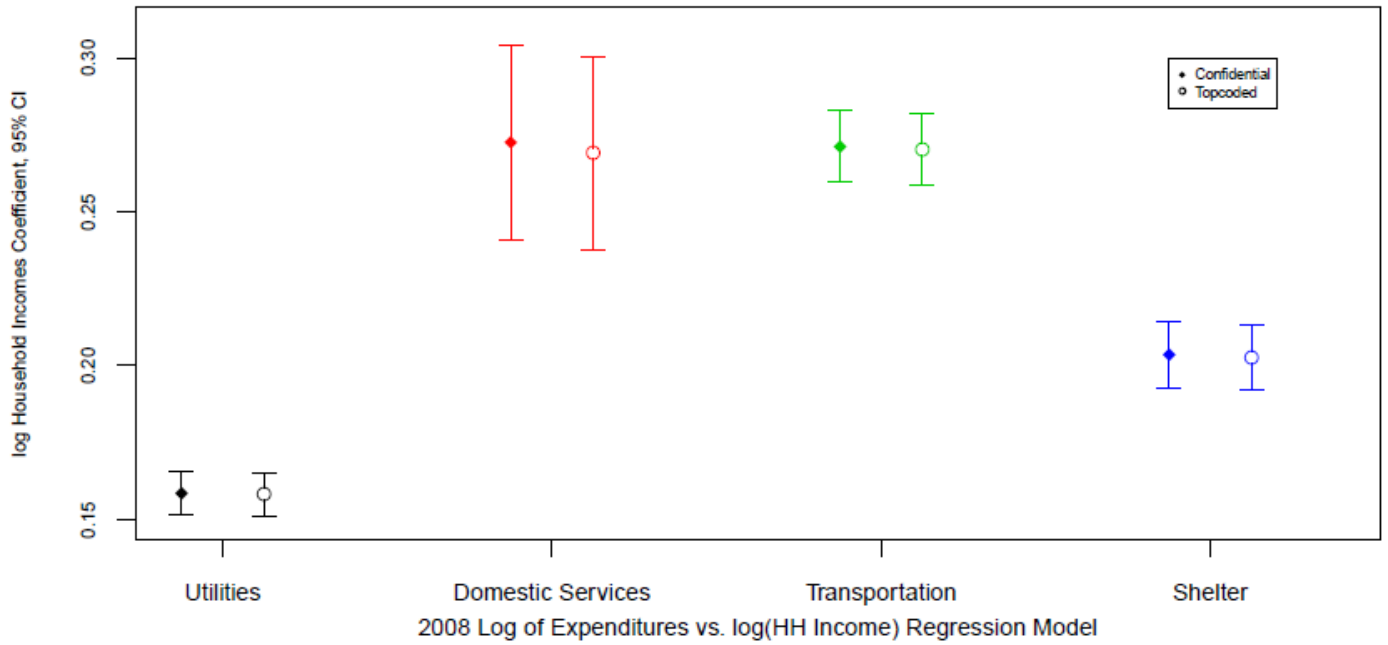
Expenditure Data and Demographics

- CE Data: 2008 public released micro data and confidential data.
- Expenditure outcomes: Utilities, Domestic Services, Transportation, Shelter, Medical Supplies, Major Appliances, Other Vehicle, and New Cars and Trucks
- Covariates (adopted from Omori 2010): household (HH) income, family type (ref.: married couple), geographical region (ref.: Northeast), numbers of children (age 0-5, 6-12 and 12-18), reference person's: education attainment (ref.: Less than HS), Occupation (ref.: Other), ethnicity (ref.: White), age.

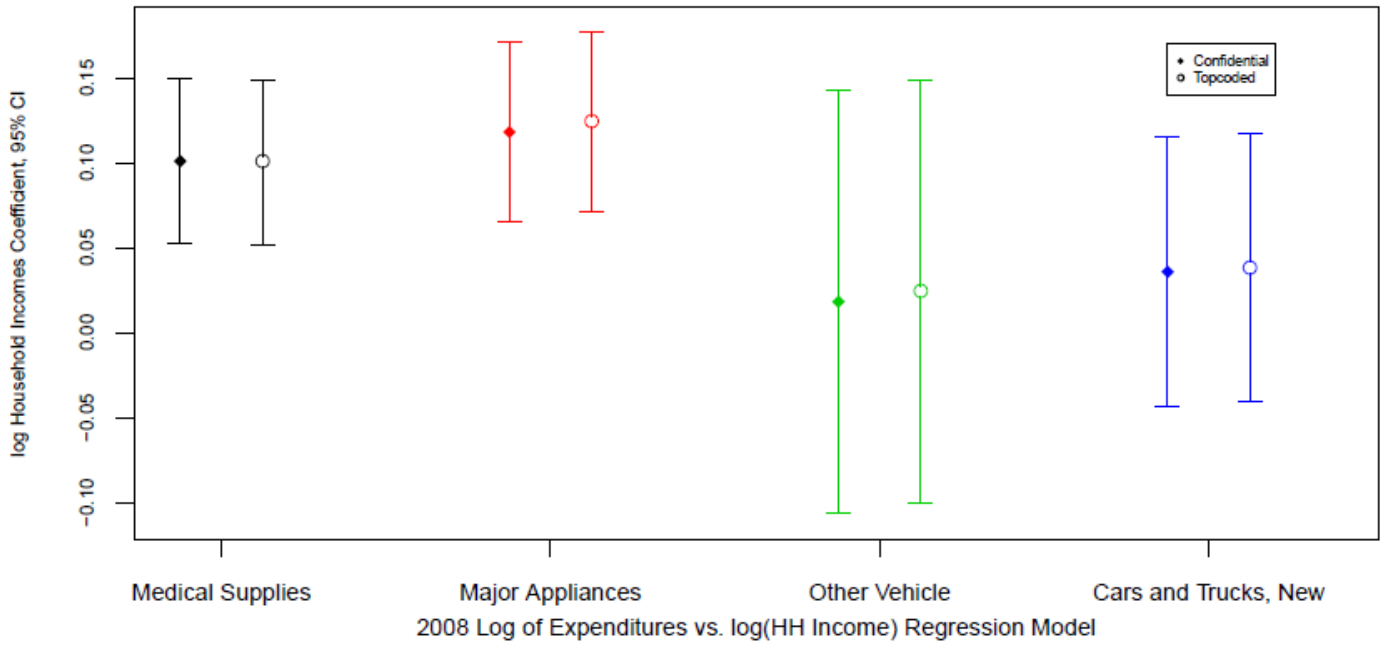
ref.: reference level, HS: high school

Log Linear Part of the Model

– $\beta_{ln(Household\ Income)}$ and 95% CI (1)

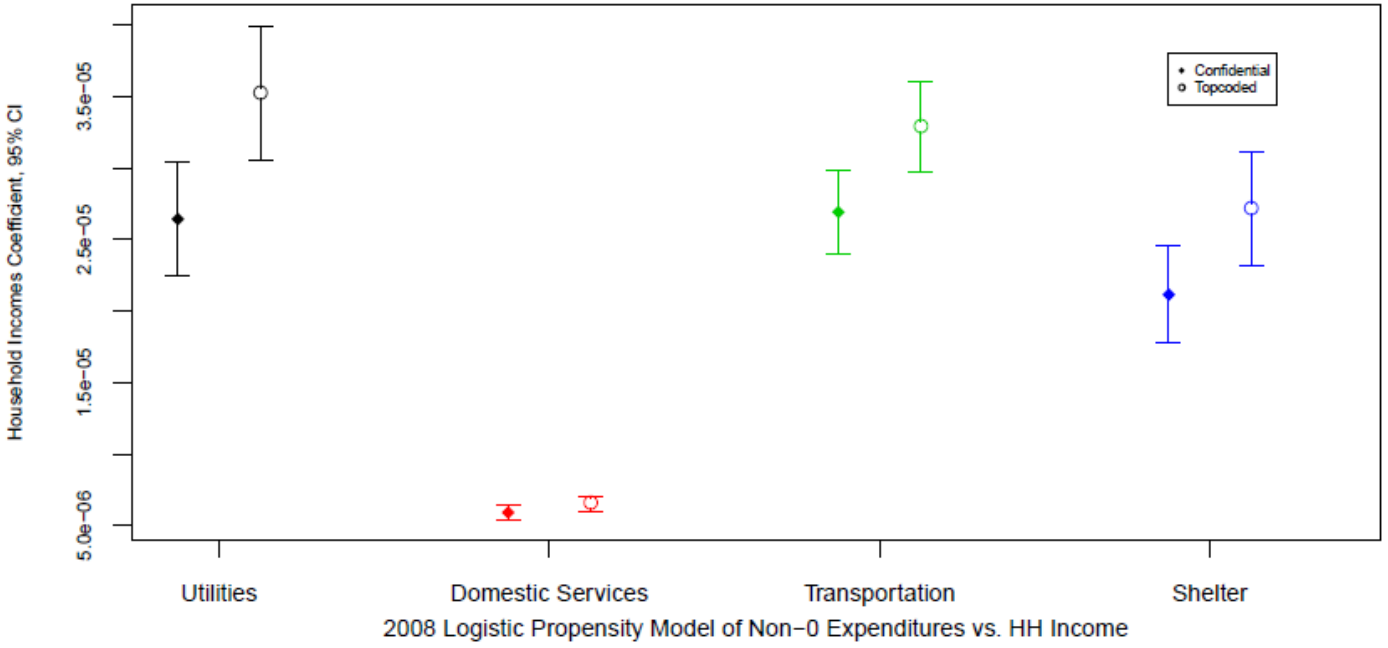


Log Linear Part of the Model
– $\beta_{ln(Household\ Income)}$ and 95% CI (2)



Logistic P.S. Part of the Model

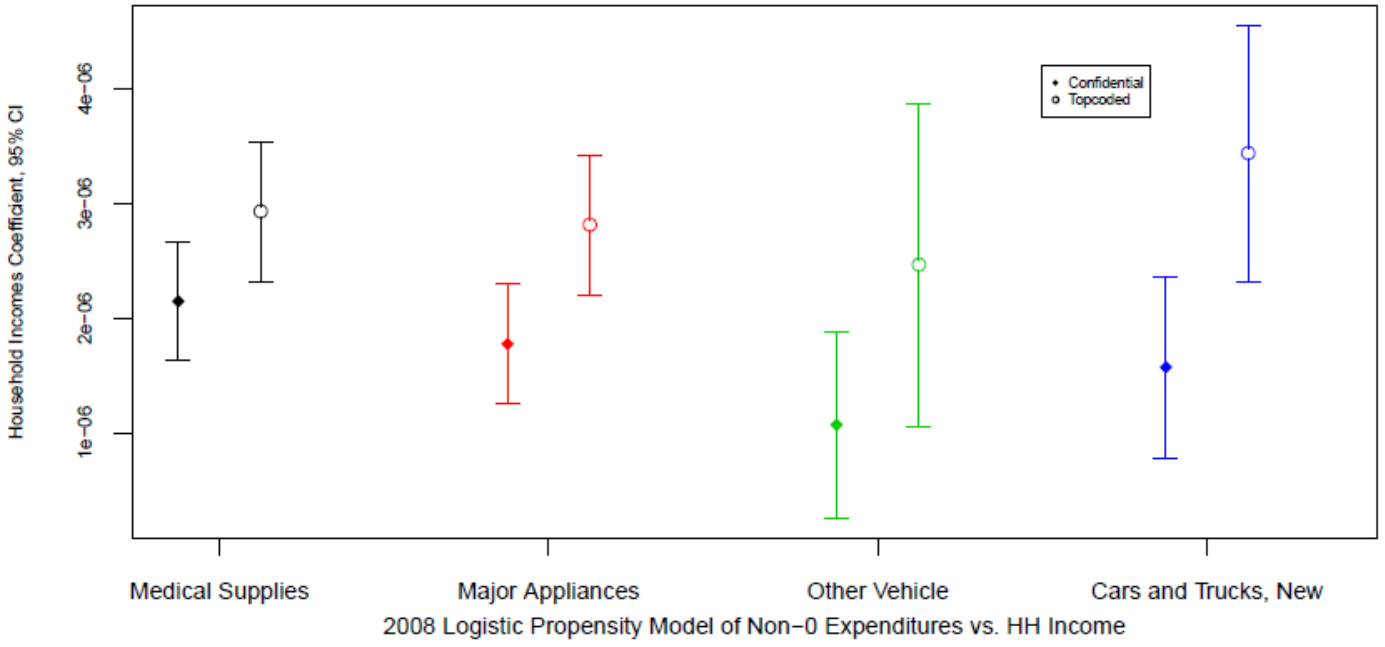
$\gamma_{Household\ Income}$ and 95% CI (1)



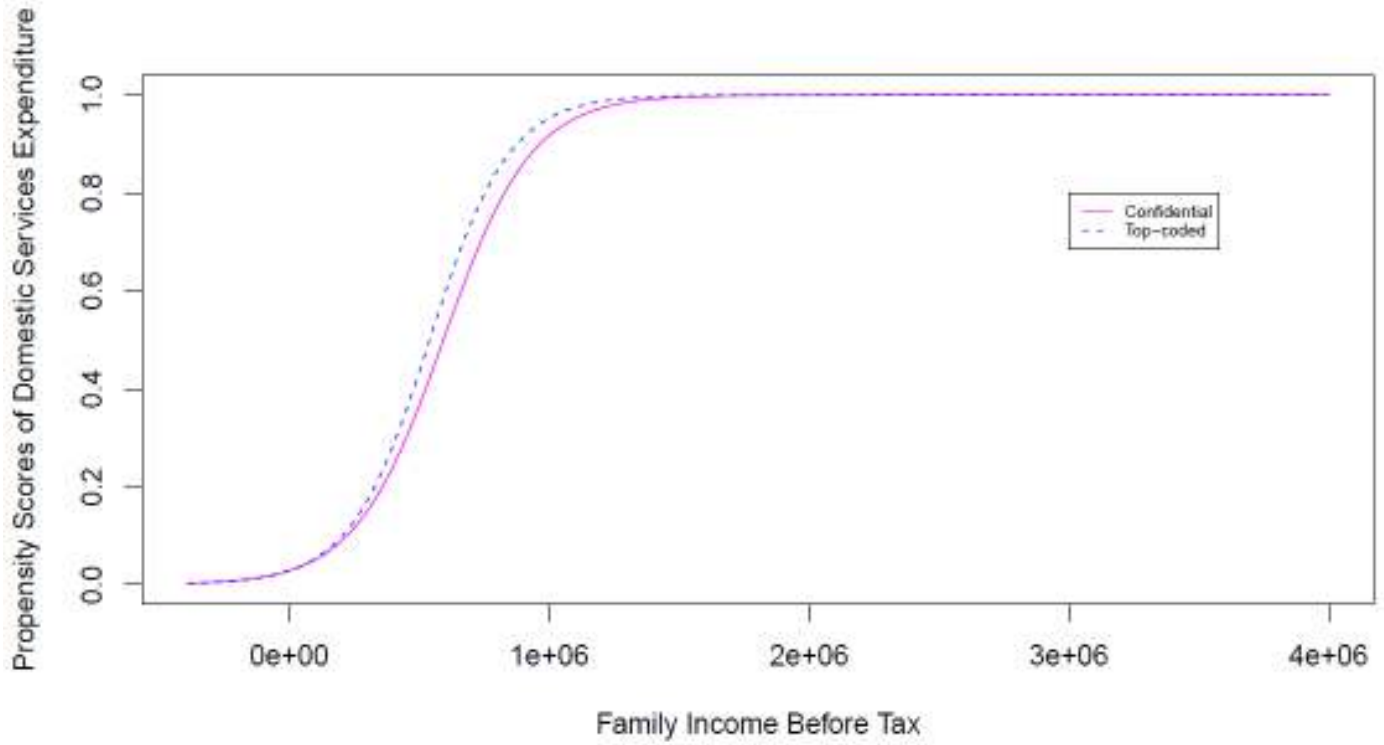
P.S.: propensity scores

Logistic P.S. Part of the Model

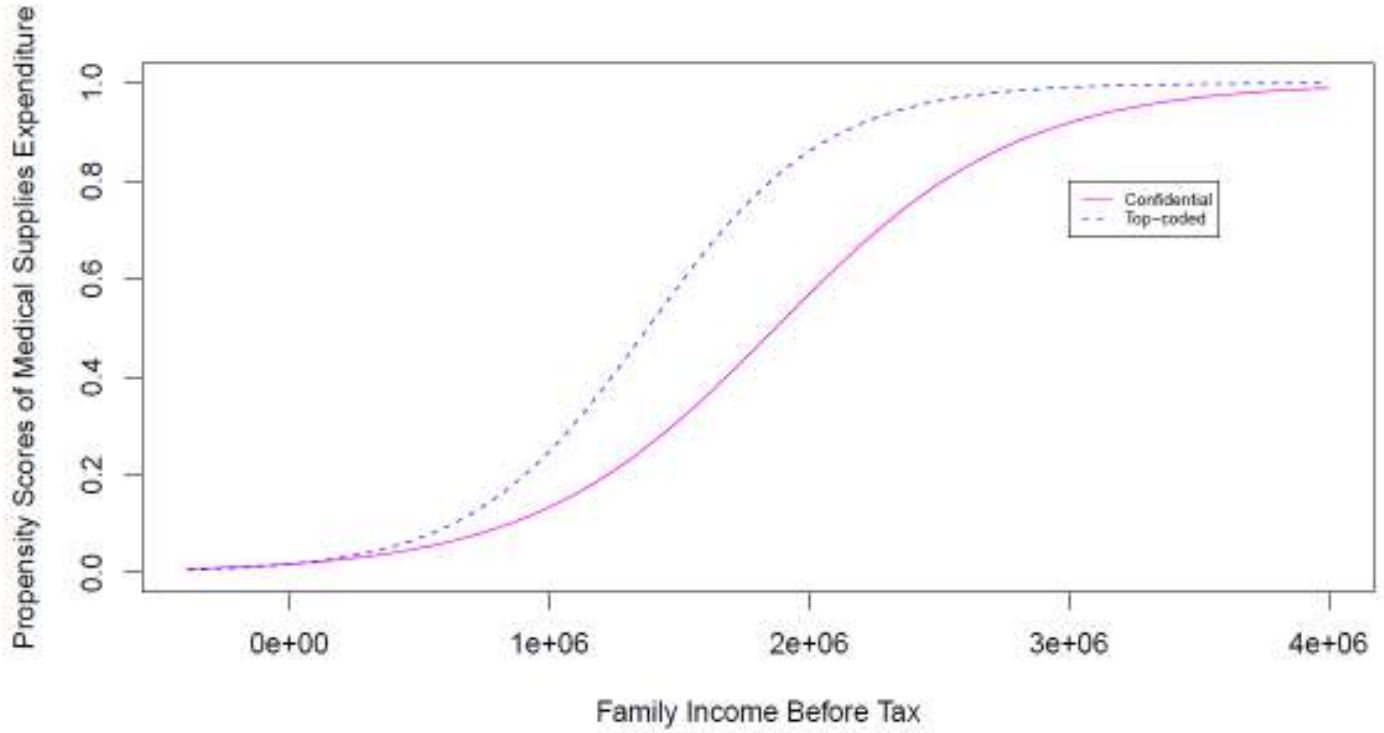
$\gamma_{Household\ Income}$ and 95% CI (2)



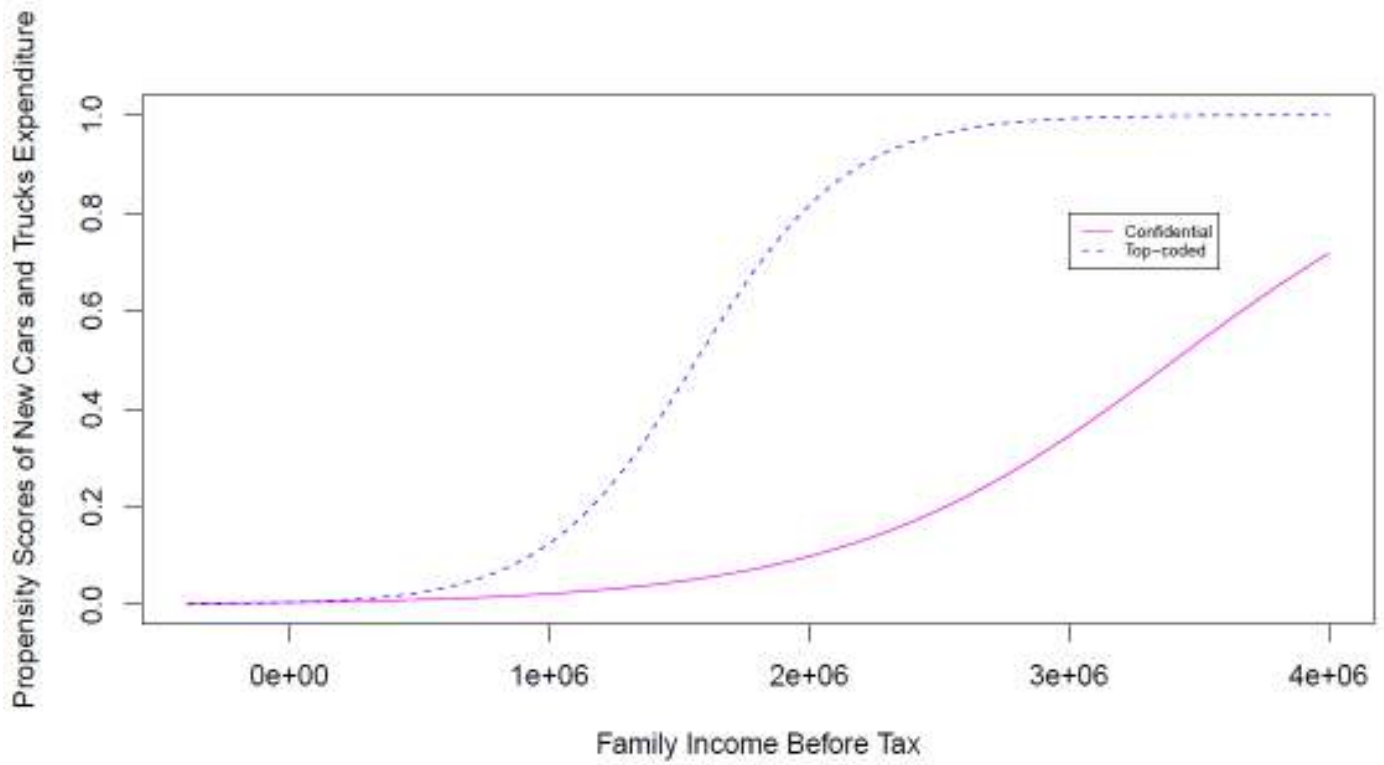
Propensity Scores: Domestic Services



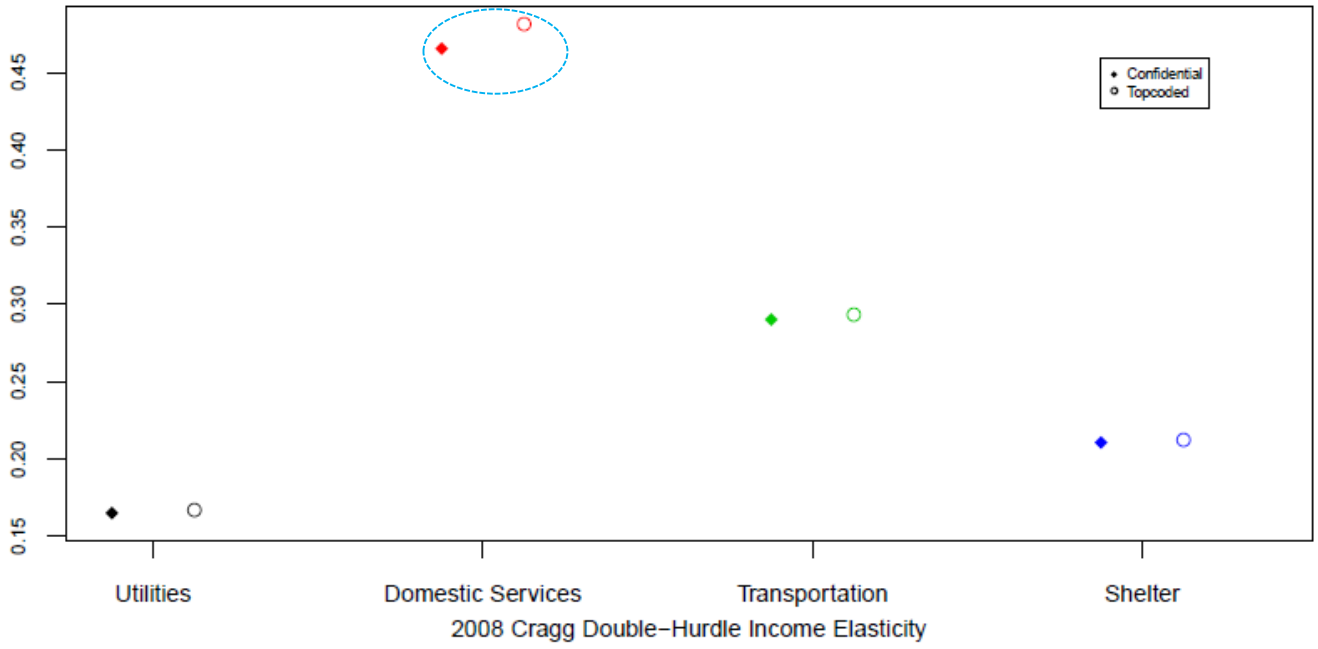
Propensity Scores: Medical Supplies



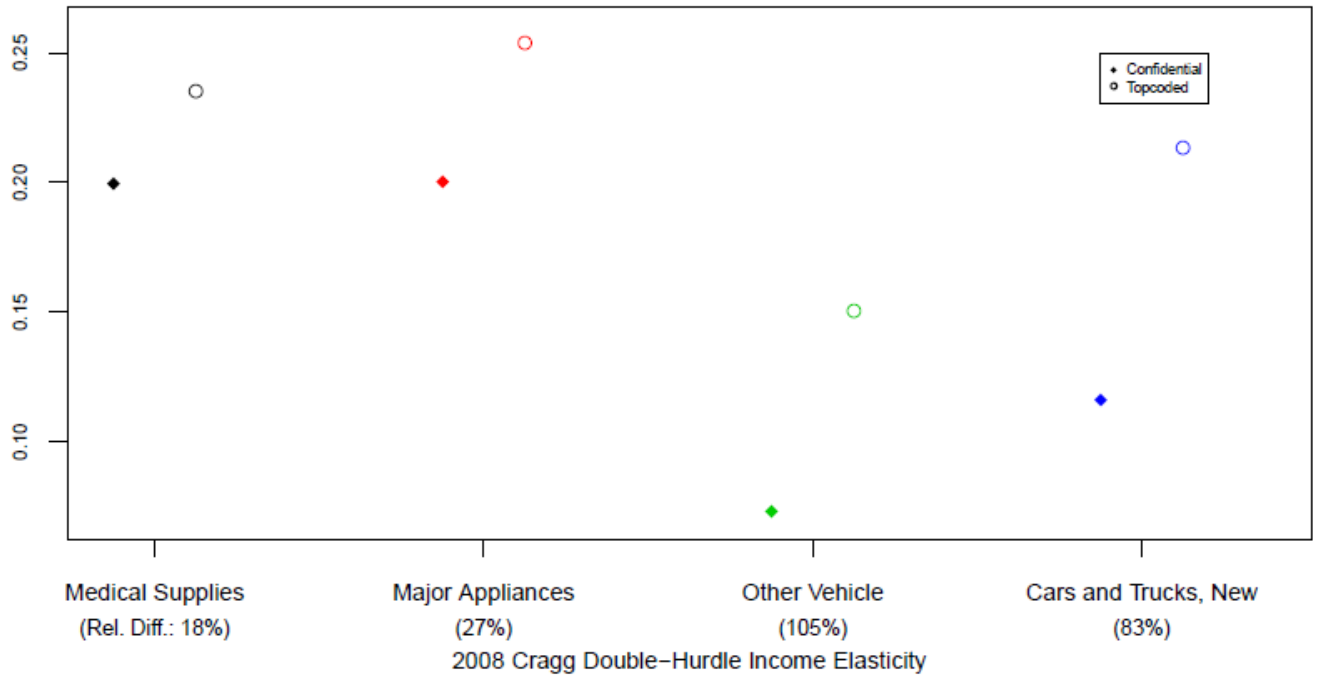
Propensity Scores Curve: New Cars and Trucks



Income Elasticity (1)



Income Elasticity (2)



Income Elasticity of Demand

Expenditures	Ratio of Top-coded over Confidential
Utilities	1.012
Domestic Services	1.034
Transportation	1.009
Shelter	1.006
Medical Supplies	1.176
Major Appliances	1.268
Other Vehicle (i.e. Motorcycle and Airplanes)	2.056
Cars and Trucks, New	1.922

Proportion Ratio Indicator:

- Proportion Ratio Indicator (PRI) of Top-coding:

- $$PRI = \frac{proportion\ \ of\ top{-}coded\ purchaser}{proportion\ of\ nontop{-}coded\ purchaser} - 1$$

Income Elasticity of Demand

Expenditures	Ratio of Top-coded over Confidential	Proportion Ratio Indicator (PRI) of Top-coding
Utilities	1.012	0.0026
Domestic Services	1.034	0.6840
Transportation	1.009	0.0021
Shelter	1.006	0.0037
Medical Supplies	1.176	0.2420
Major Appliances	1.268	0.2313
Other Vehicle (i.e. Motorcycle and Airplanes)	2.056	0.3882
Cars and Trucks, New	1.922	0.4313

Summary

- No difference in log linear model between confidential and top-coded data.
- Differences from some of the propensity models. This translates into some differences in income elasticity of demand for some expenditures.
- Proportion Ratio Indicator (PRI) appears to reflect those differences.

Acknowledgements and Disclaimer

- We would like to acknowledge our fellow researchers and managers from the program office.

- The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of the Bureau of Labor Statistics.

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.