Evaluating the Consumer Expenditure Data Top-Coding Effects on Economics Models

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

Overview

- Consumer Expenditure Surveys (CE) and top-coding
- Income elasticity of Demand and Zero-Inflated Model
- Evaluations on log regression model for expenditures
- Evaluations on logistic model for propensity of consumption
- Effects on elasticity and conclusion

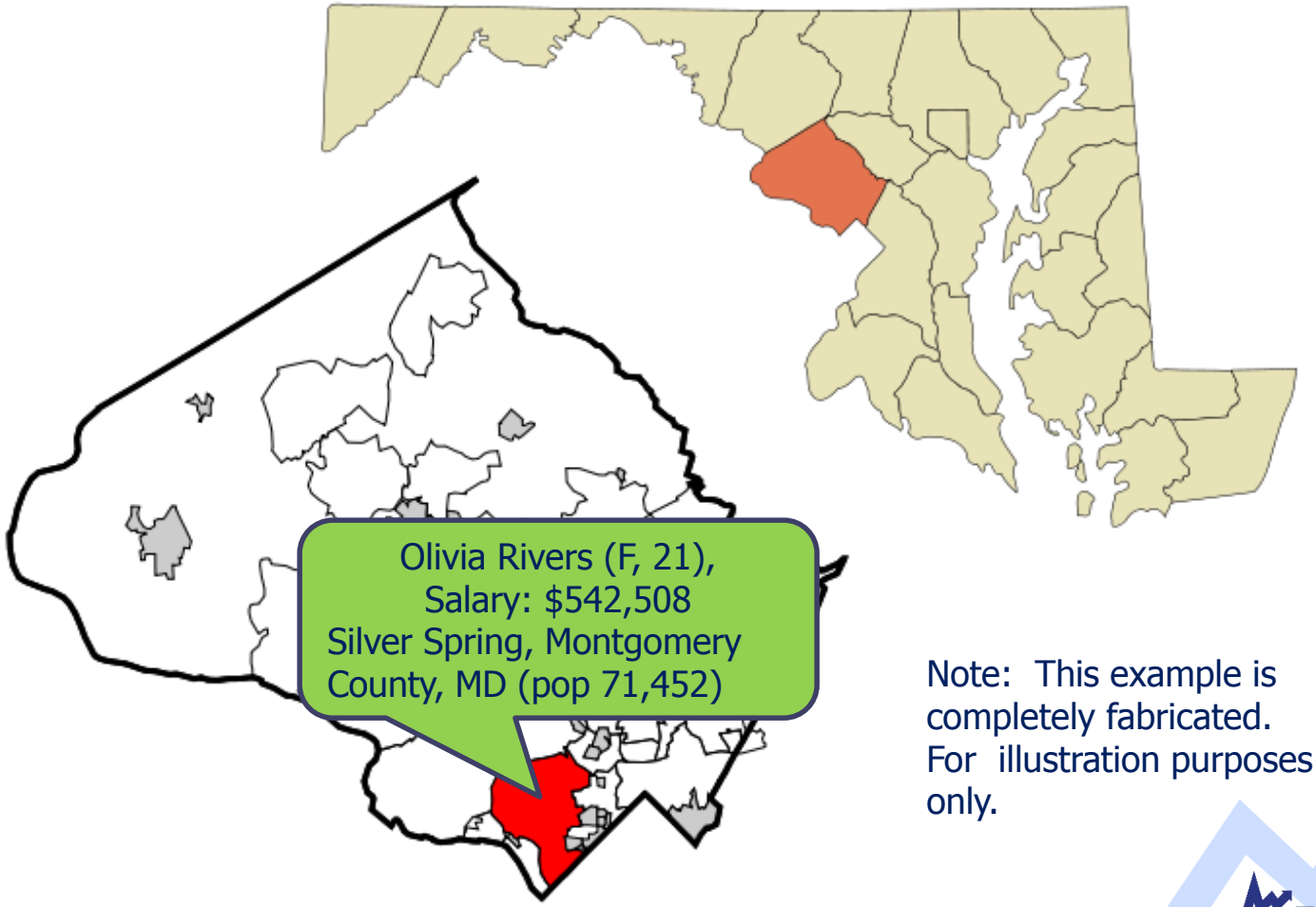
Consumer Expenditure Survey

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

CE SDL Process

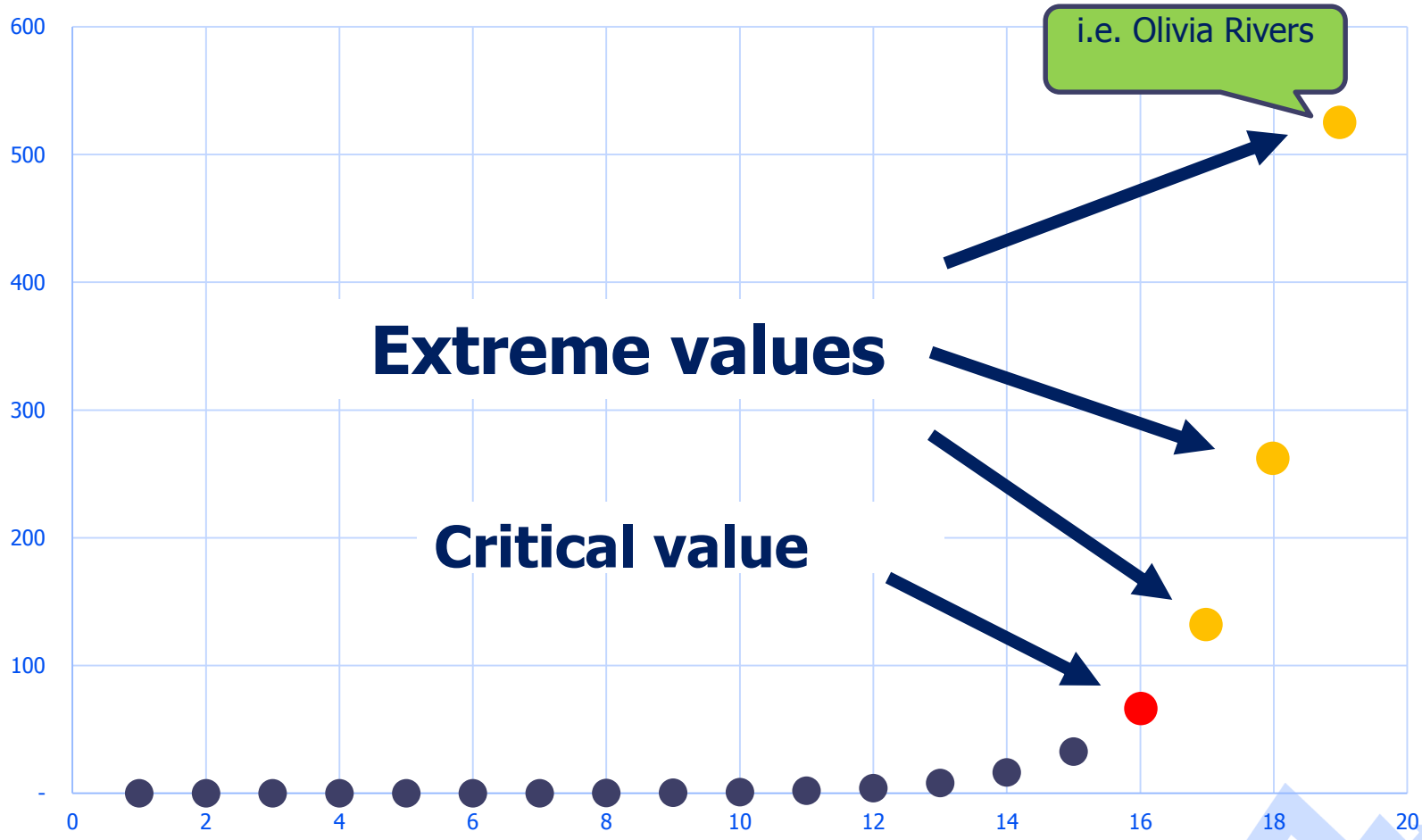
- CE microdata release requires statistical disclosure limitation (SDL).
- Objective: Conceal personally identifiable information to preserve the confidentiality and anonymity of survey participants.
- Production Process: “top-coding”
- Our goal is to assess its numerical impact.

Top-coding



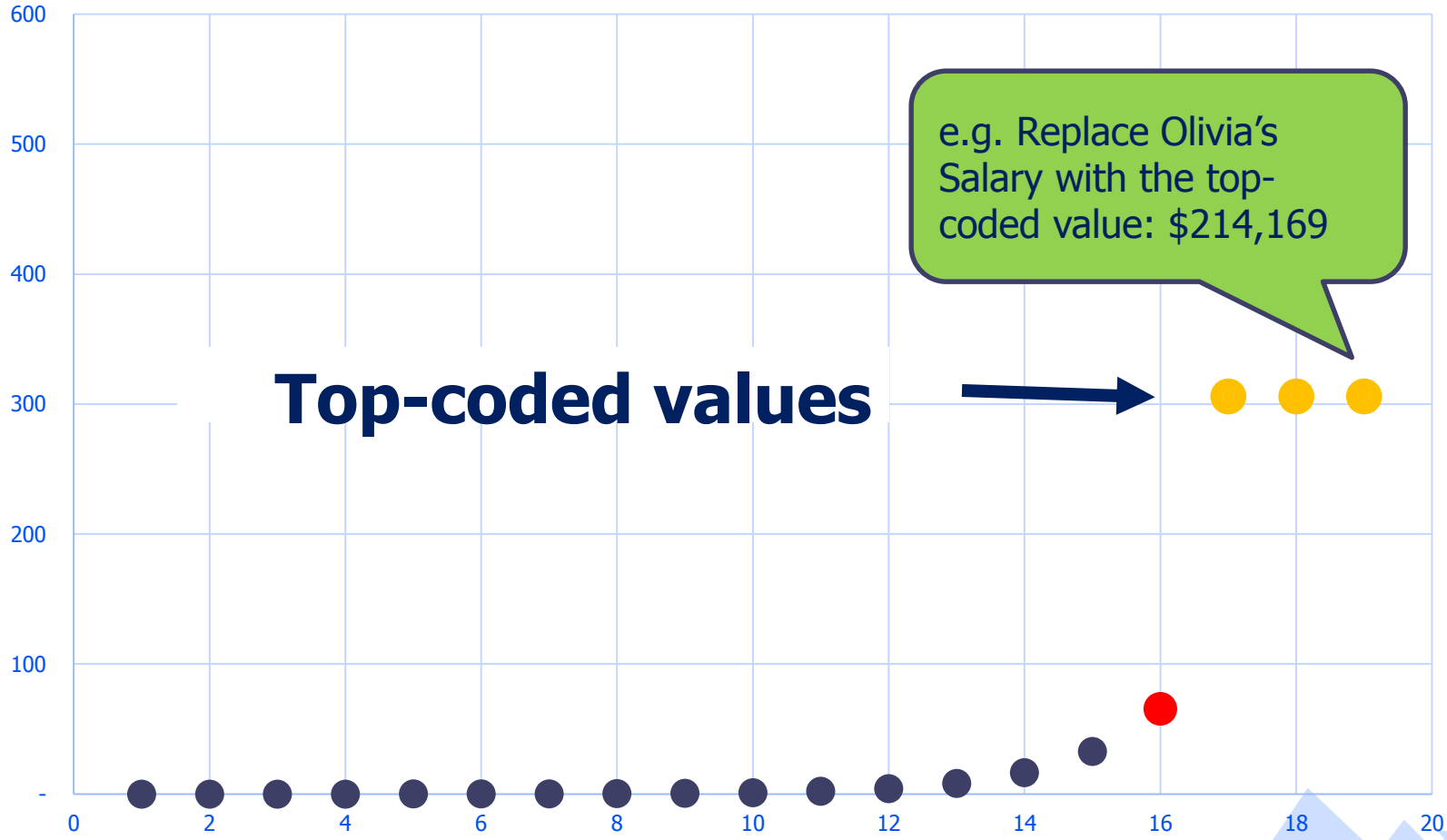
Note: This example is completely fabricated. For illustration purposes only.

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.)



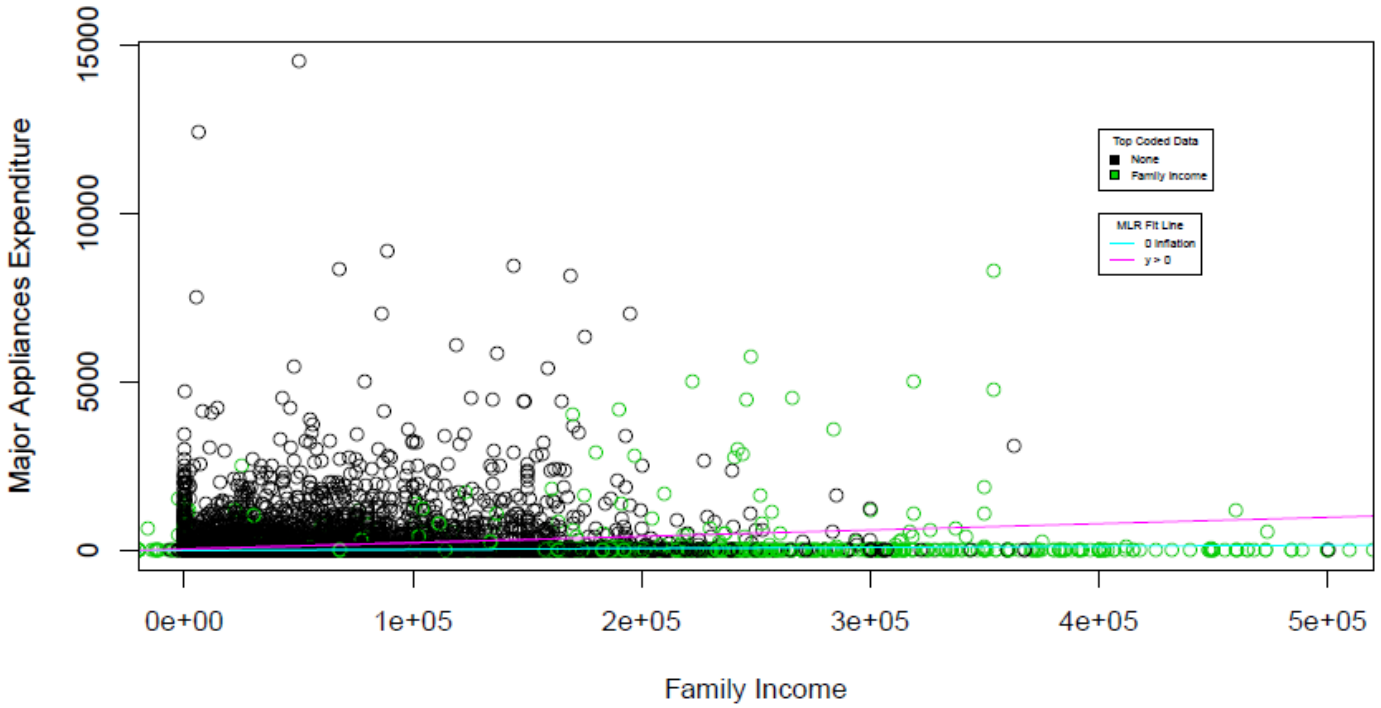
Source: *Balancing Respondent Confidentiality and Data User Needs*, Aaron E. Cobet, BLS, 2014 CE Microdata User's Workshop.

Notation

- Suppose y is an expenditure of the household
- $\boldsymbol{x}$ is a vector of covariates including income
- The expenditure will have 0-inflation because not all households made a purchase of a specific expenditure:

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

Expenditure vs. Household Income



Income Elasticity of Demand

- Here, y – Expenditure, $\boldsymbol{x}$ – covariates, x_j - household income.

- Income Elasticity of demand:

$$\frac{\partial E(y \mid \boldsymbol{x})}{\partial x_j} \frac{x_j}{E(y \mid \boldsymbol{x})}$$

- The income elasticity of demand can be interpreted as “the percent change in expenditure for a specific good given a 1-percent increase in income.”

References

- This measure is considered in several economics studies:

  - Joseph G Altonji and Ernesto Villanueva. The marginal propensity to spend on adult children. The BE Journal of Economic Analysis & Policy, 7(1):14, 2007
  - Riccardo De Bonis and Andrea Silvestrini. The effects of financial and real wealth on consumption: new evidence from oecd countries. Applied Financial Economics, 22(5): 409–425, 2012.
  - James Michael Harris, Noel Blisard, et al. Food-consumption patterns among elderly age groups. Journal of Food Distribution Research, 33(1):85–91, 2002.
  - Matteo M Iacoviello. Housing wealth and consumption. In International Encyclopedia of Housing and Home, pages 673–678. Elsevier Ltd., 2012.
  - Michael Kumhof and Douglas Laxton. Fiscal deficits and current account deficits. Journal of Economic Dynamics and Control, 37(10):2062–2082, 2013.
  - Theodore Tsekeris. Disaggregate analysis of gasoline consumption demand of greek households. Engineering Economics, 23(1):41–49, 2012.
  - Robert O Weagley and Eunjeong Huh. The impact of retirement on household leisure expenditures. Journal of consumer affairs, 38(2):262–281, 2004.

Zero-Inflated Model

- The unconditional expectation is

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

model with logistic regression

model with log linear regression

Model with Logistic Regression

- Assume a Logistic propensity model of consumption:

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

here, γ_j is the logistic coefficient of household income.

Model with Log Linear Regression

- Assume the outcome follow:

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

where income $x_j > 0$ is also logged and

β_j is the coefficient of $log(x)_j$.

- Then, the 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 τ_{x_j}

- Income Elasticity of Demand is

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

logistic coefficient

$1 - P(y > 0 \mid \boldsymbol{x})$

linear coefficient

Expenditure Data

- 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.

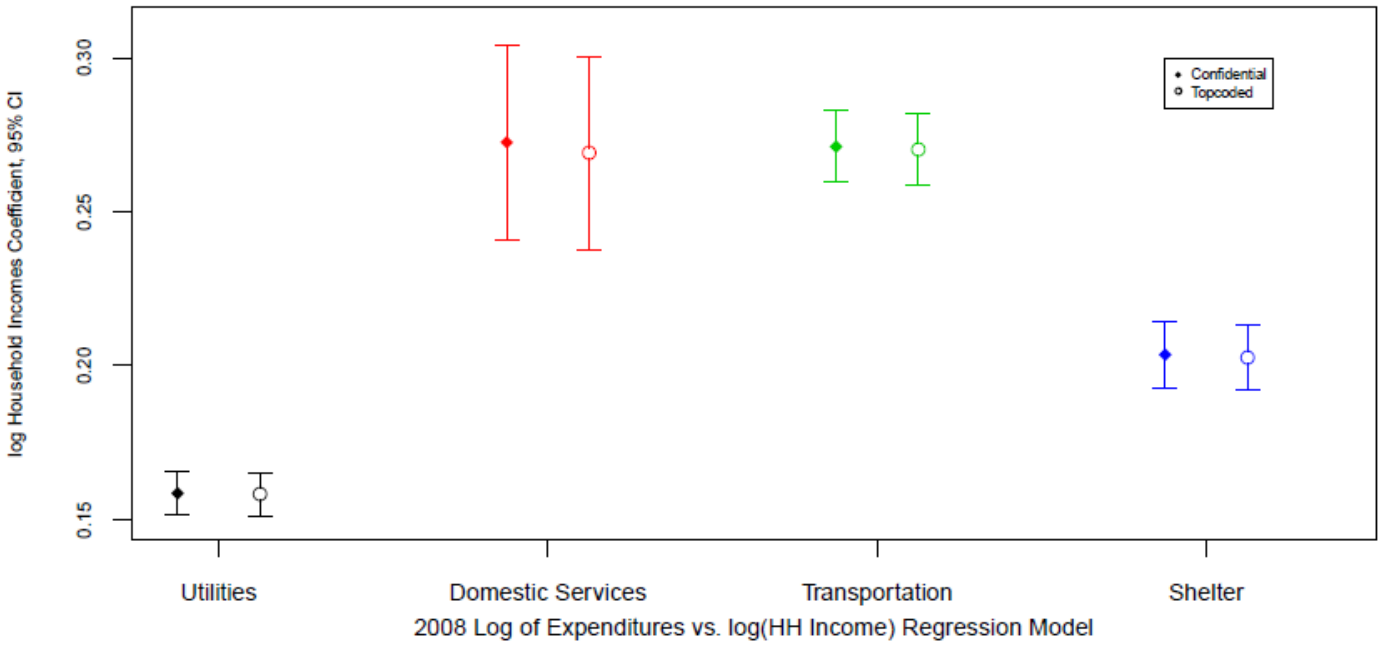
Covariates and Demographics

- 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 demographics: education attainment (ref.: Less than HS), Occupation (ref.: Other), ethnicity (ref.: White), age.

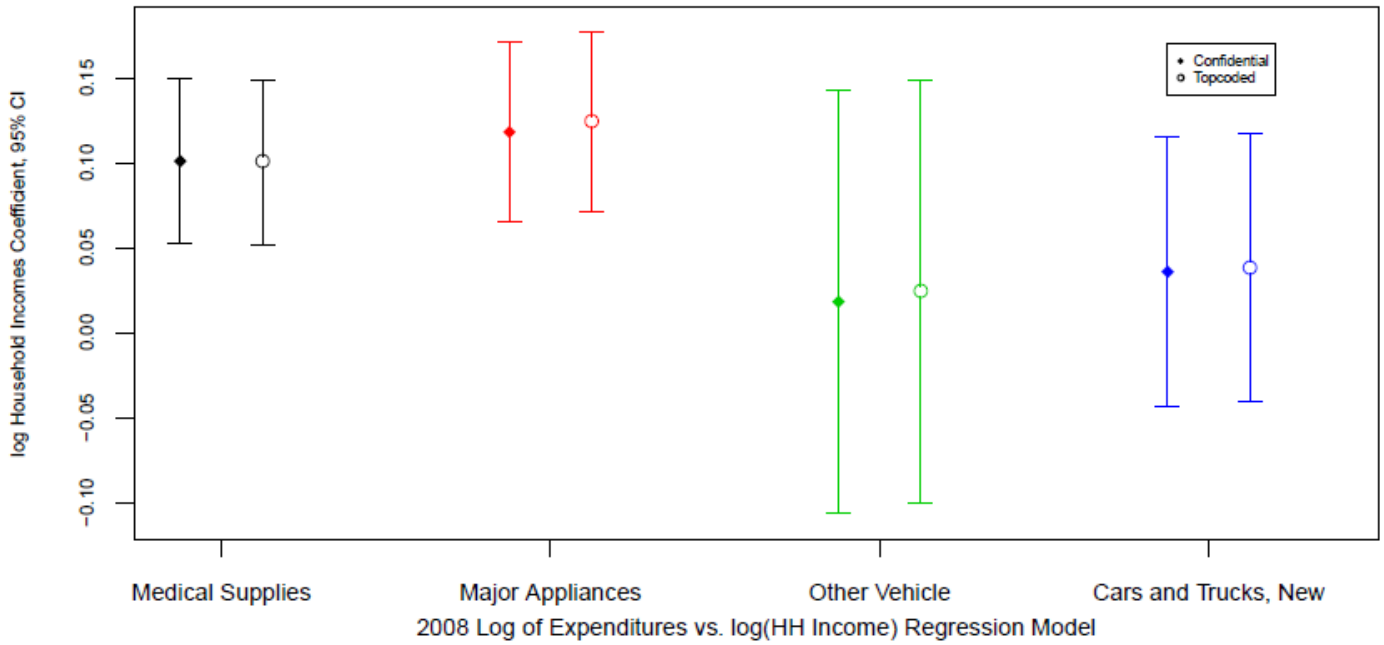
ref.: reference level, HS: high school

Log Linear Part of the Model: β_j and 95% CI (1)

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

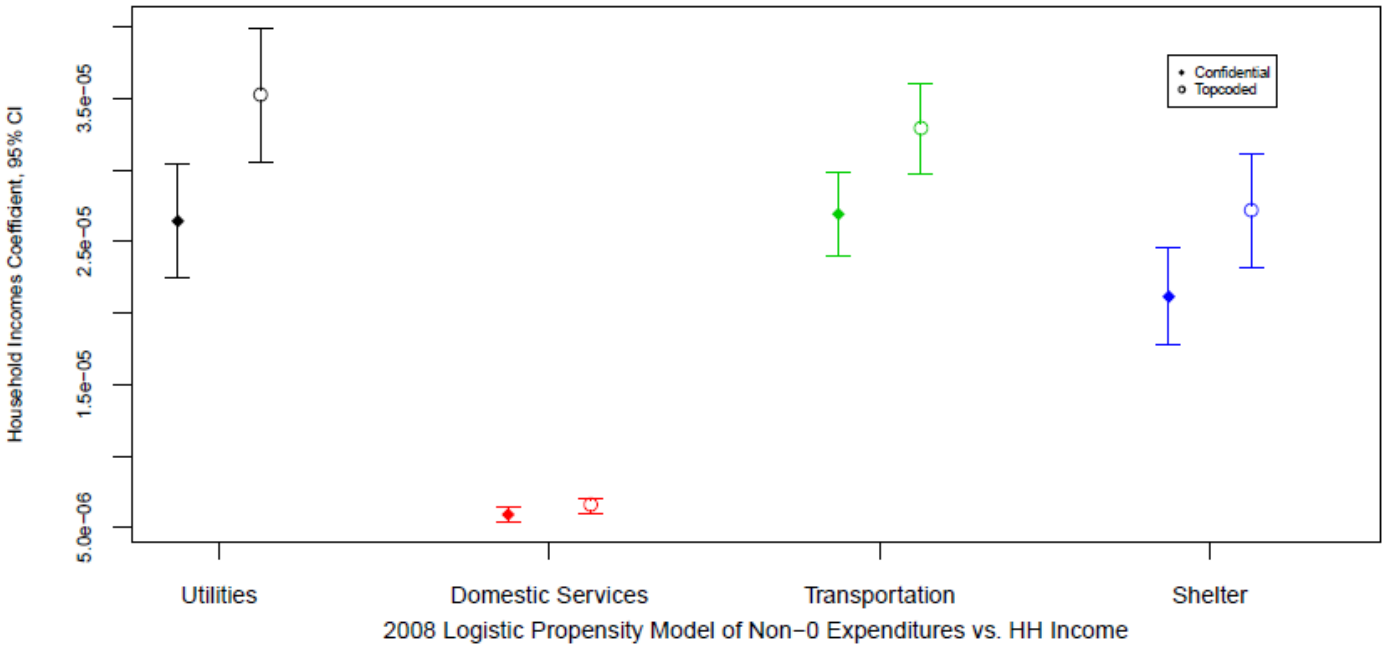


Log Linear Part of the Model: β_j and 95% CI (2)



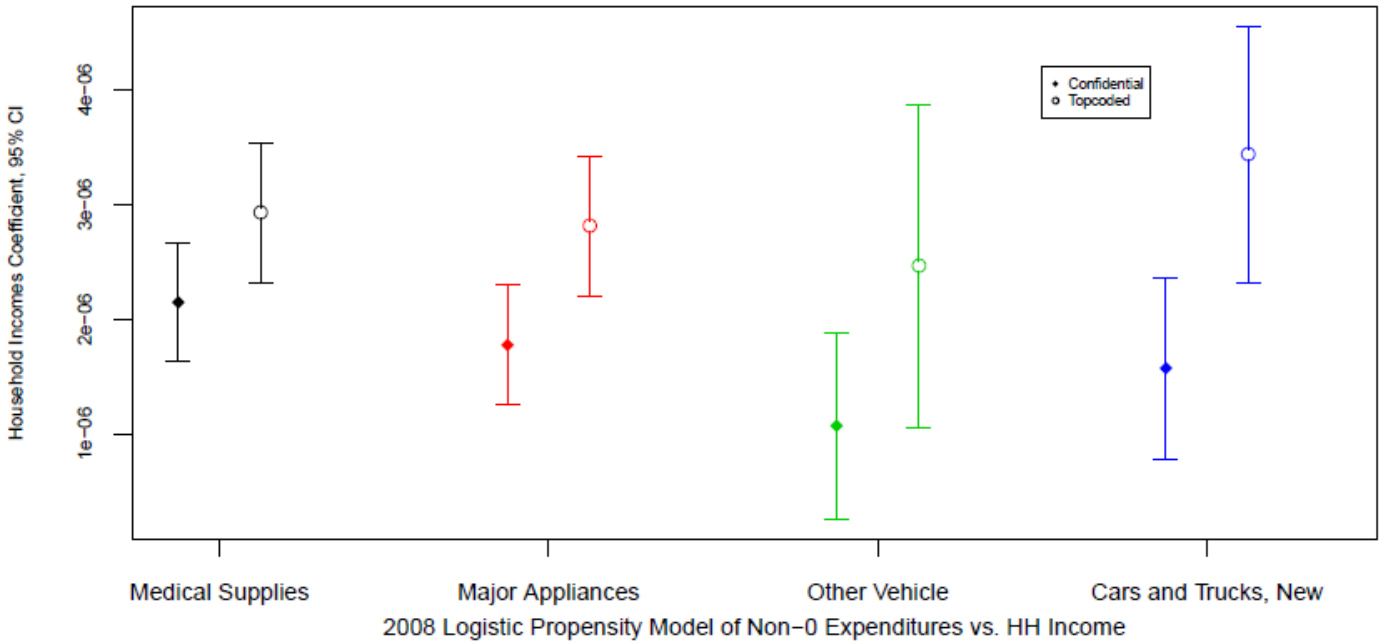
Logistic P.S. Part of the Model: γ_j and 95% CI (1)

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

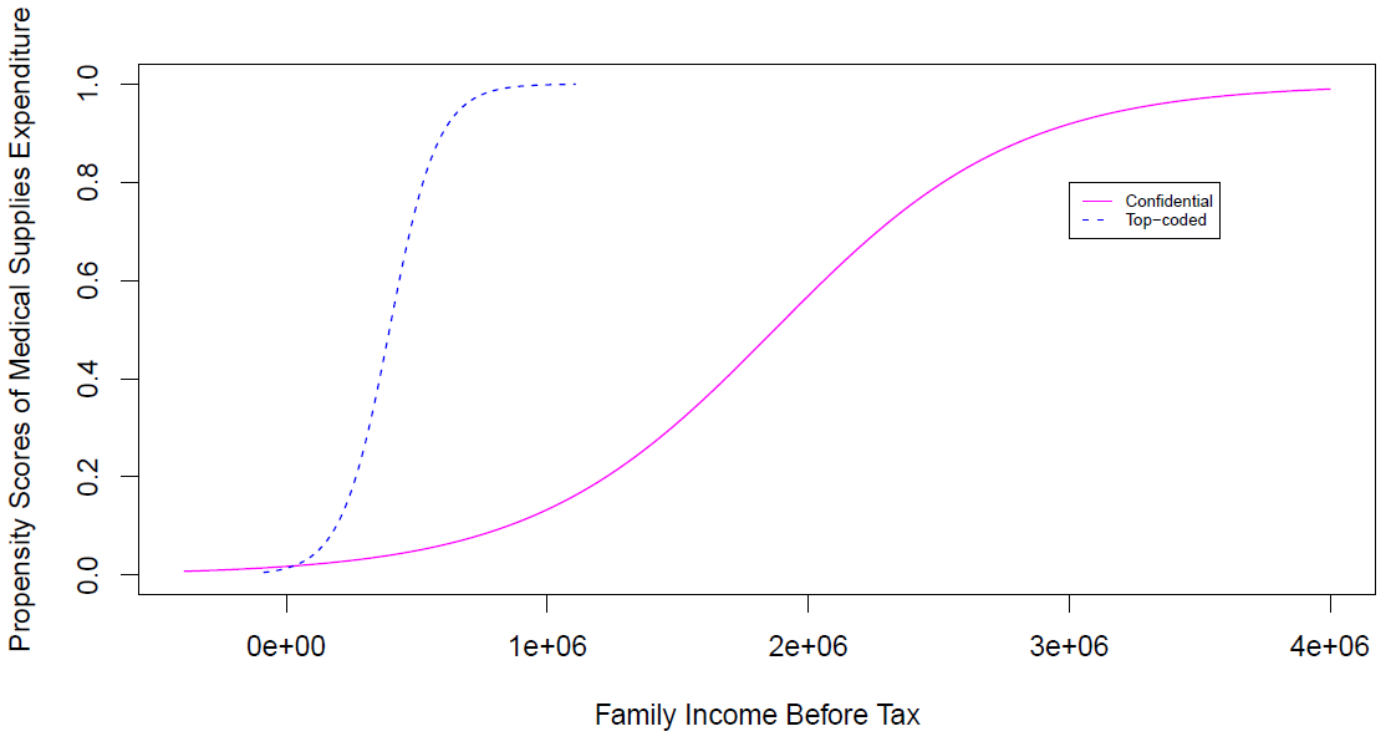


Logistic P.S. Part of the Model: γ_j and 95% CI (2)

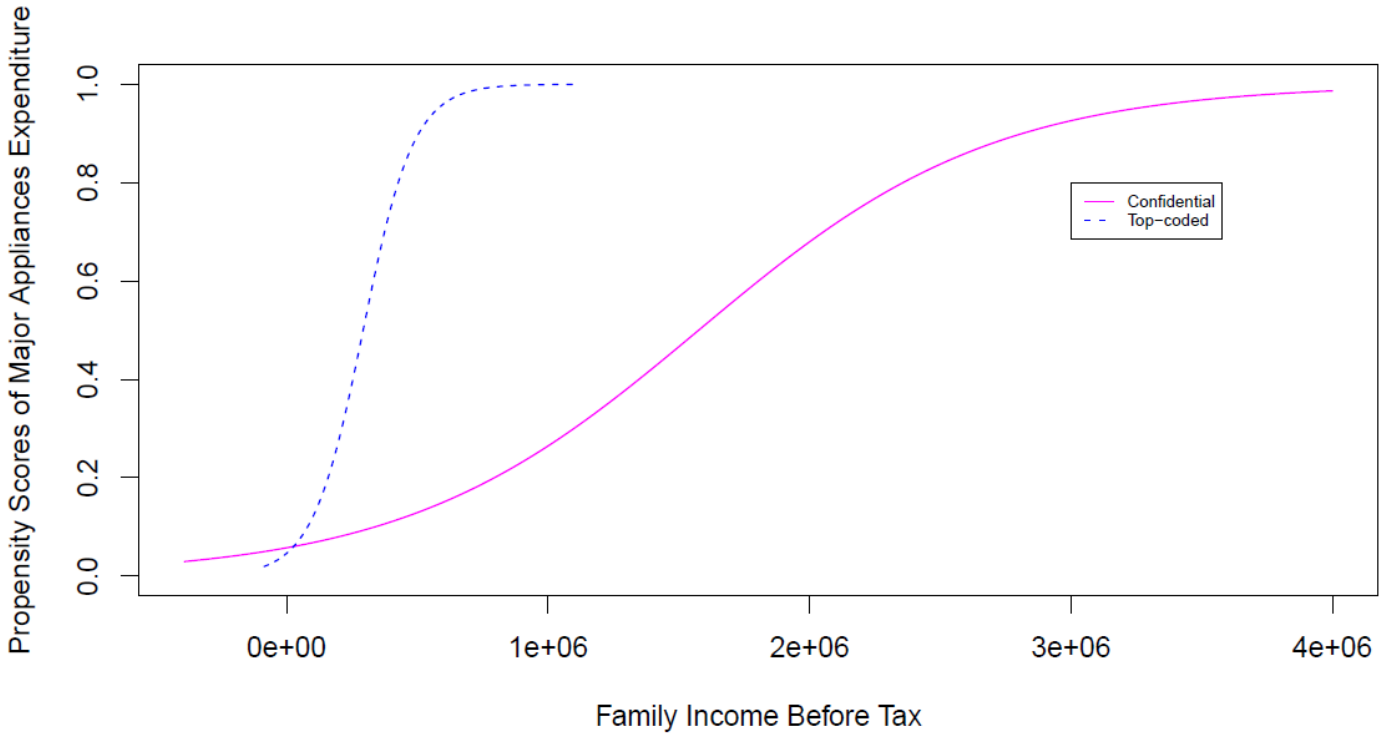
$$\tau_{x_j} = \frac{\partial E(y \mid \boldsymbol{x})}{\partial x_j} \frac{x_j}{E(y \mid \boldsymbol{x})} = \gamma_j [1 - \Psi(\boldsymbol{x}\gamma)] x_j + \beta_j$$



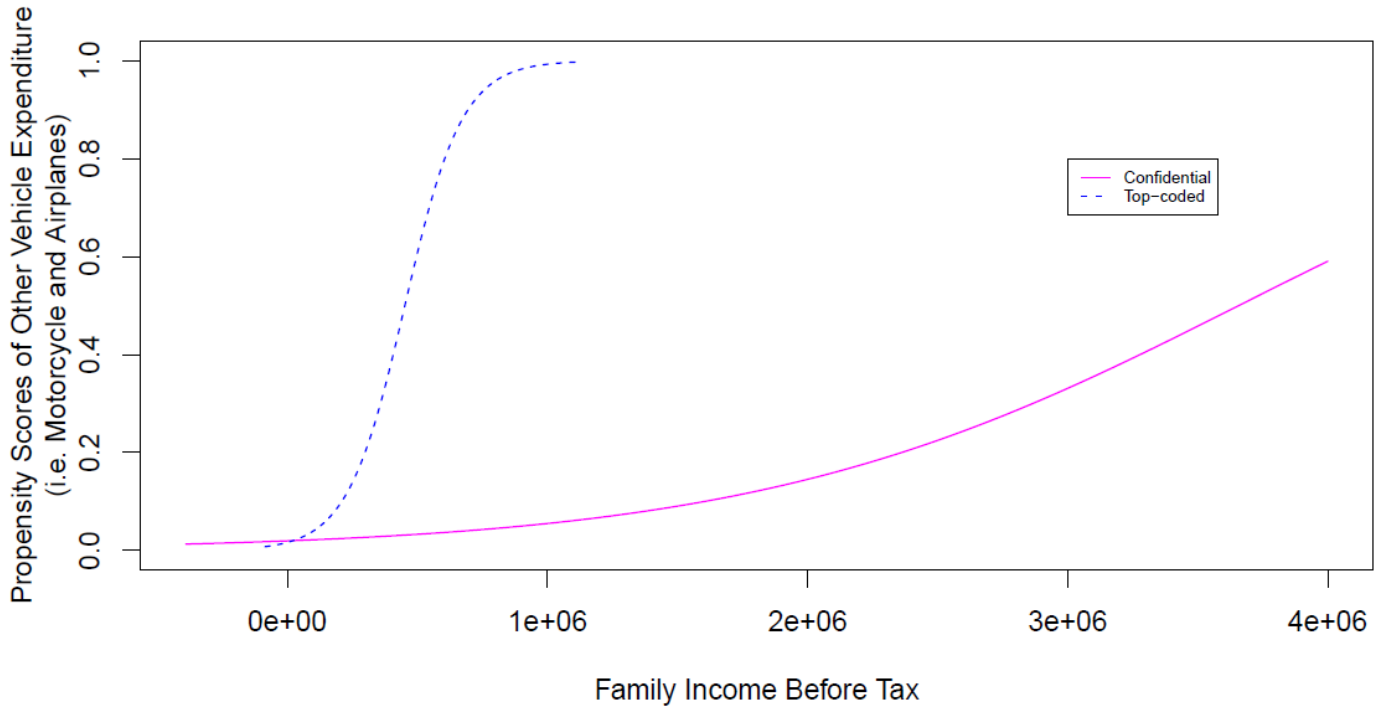
Propensity Scores Curve of Medical Supplies Expenditure



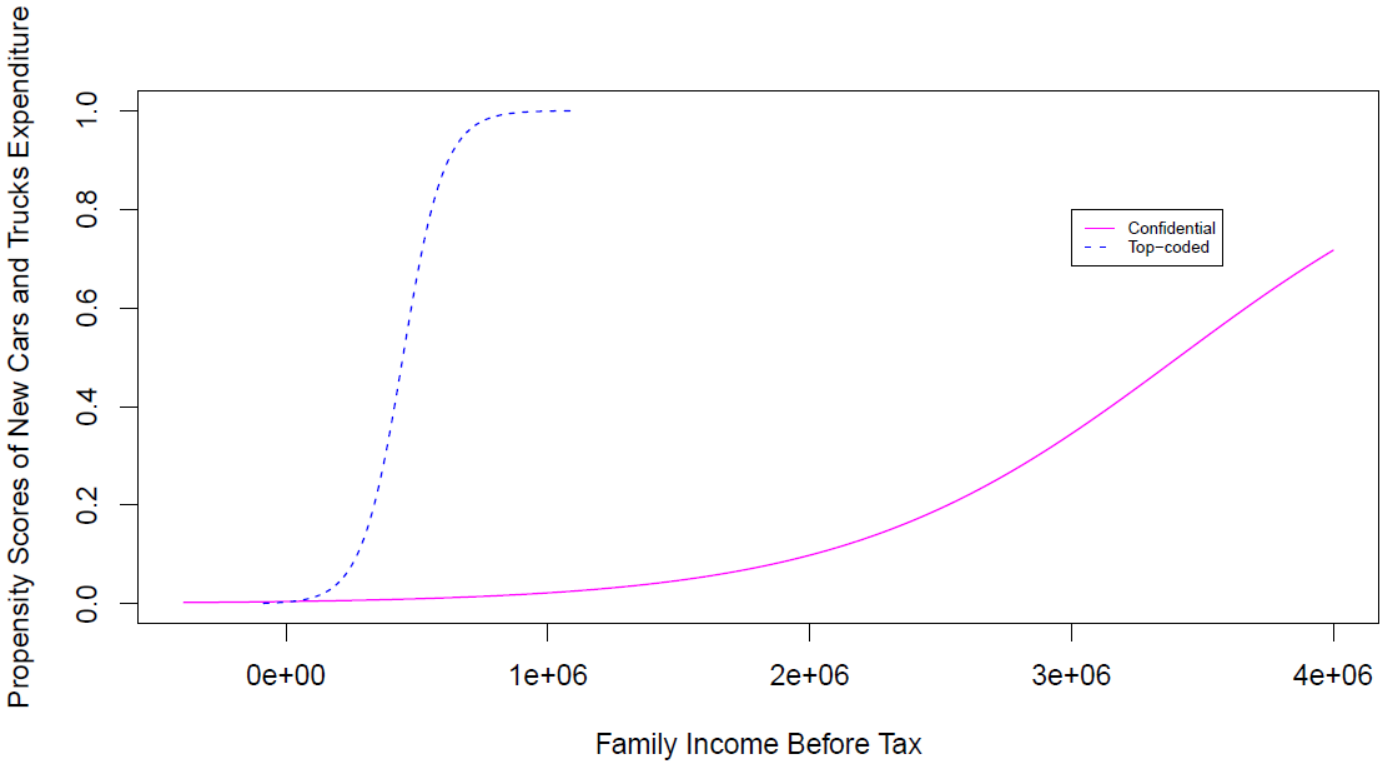
Propensity Scores Curve of Major Appliances Expenditure



Propensity Scores Curve of Other Vehicle Expenditure



Propensity Scores Curve of New Cars and Trucks Expenditure



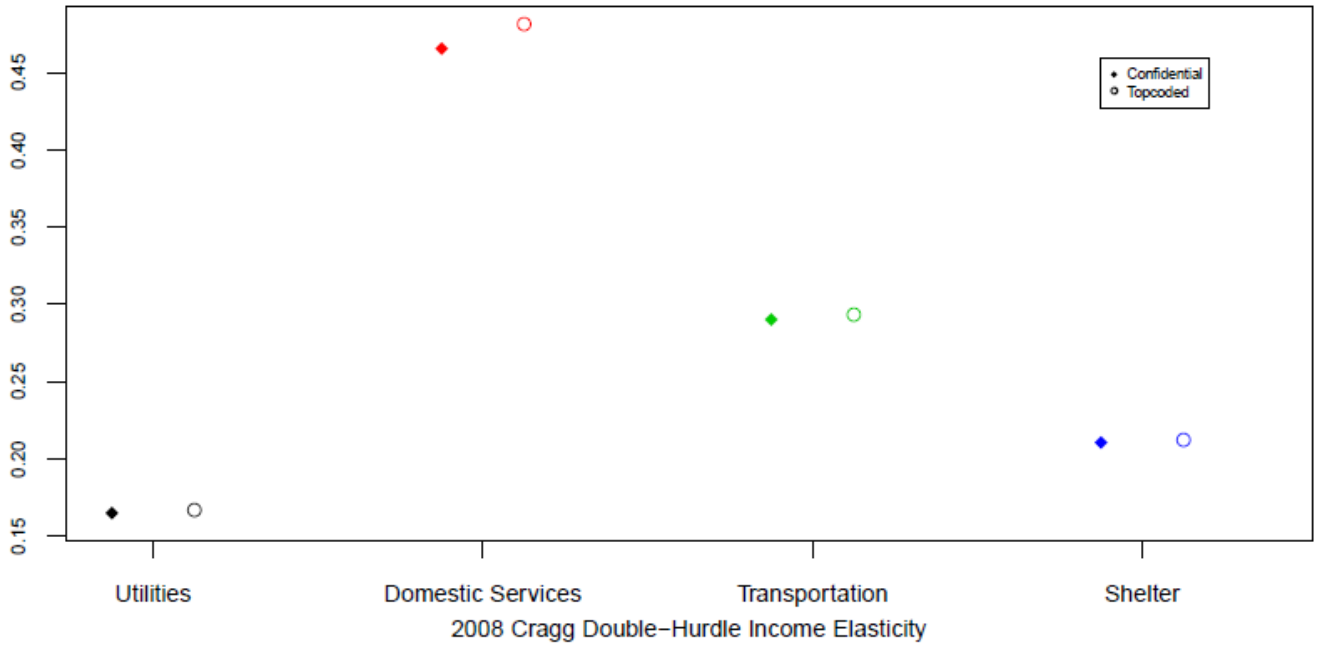
Recall Income Elasticity of Demand

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

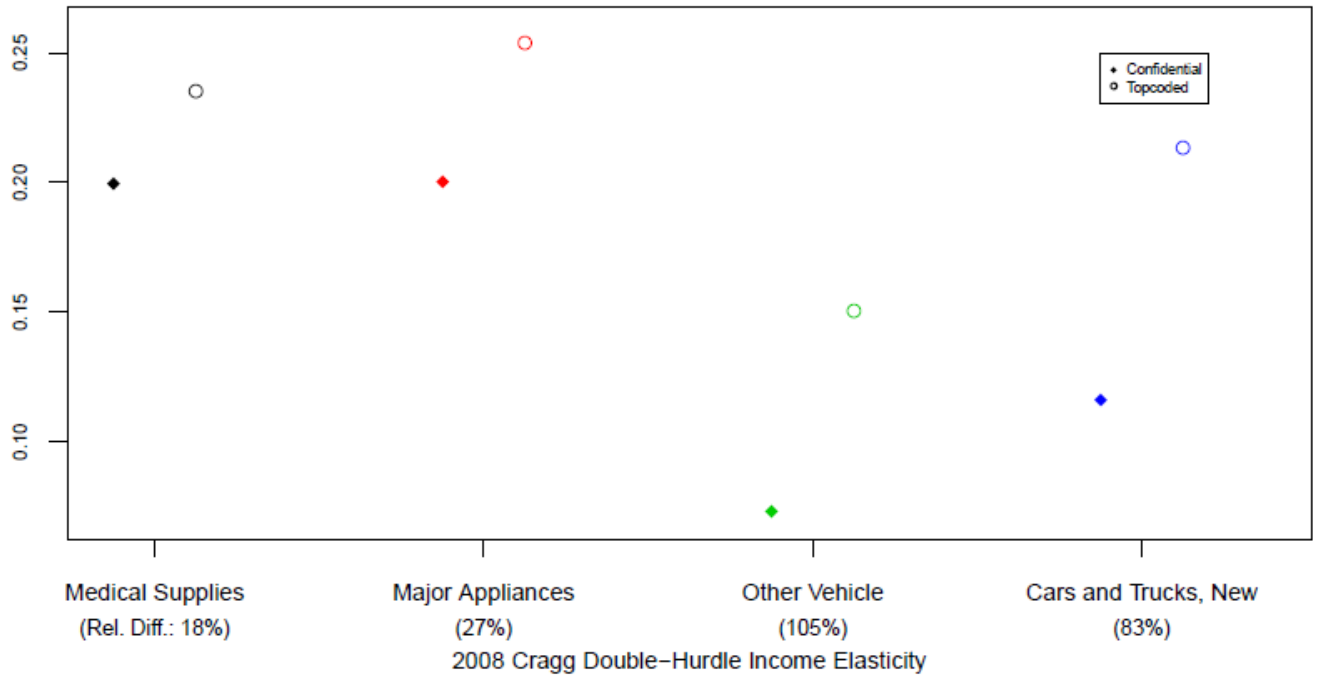
coefficient from logistic model

coefficient of income from log linear model

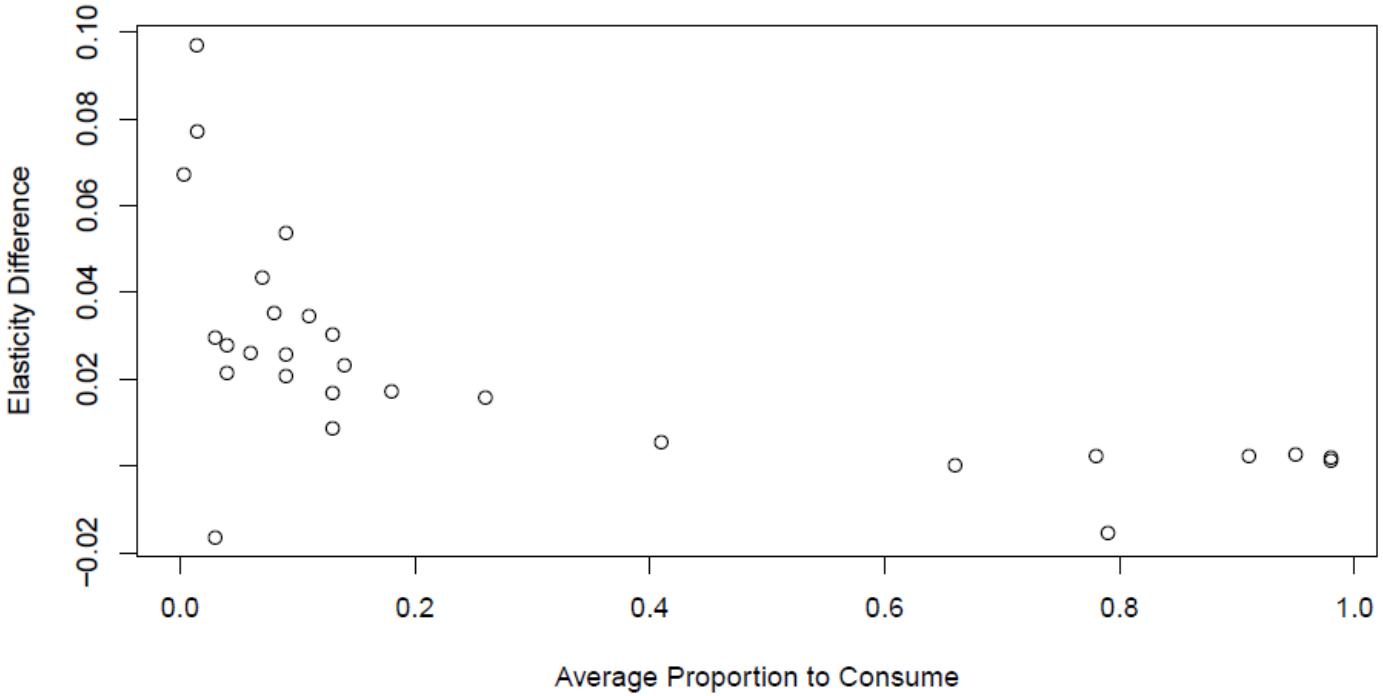
Income Elasticity of Demand (1)



Income Elasticity of Demand (2)



Estimated Elasticity Difference (for 28 Expenditures) vs. Average (Marginal) Proportion to Consume



Summary

- No difference in model for $E(y|\boldsymbol{x}, y > 0)$ between confidential and top-coded data.
- Differences in model for $P(y > 0)$ from some of the propensity models.
- Translates into some differences in income elasticity of demand for some expenditures.

Summary (cont.)

- On the other hand, even though certain expenditures are infrequent but they still are of interest to economists and industry.
- The program office may be able to come up with a warning to economists or researchers on the differences of economics measurements due to top-coding and acceptable threshold.

THANK YOU!

FCSM

Federal Committee on
STATISTICAL METHODOLOGY

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.