Top Coding in the 2010 1-Year ACS PUMS

Nicole Crimi
Advisor: Professor William Eddy

Carnegie Mellon University Department of Statistics

Introduction

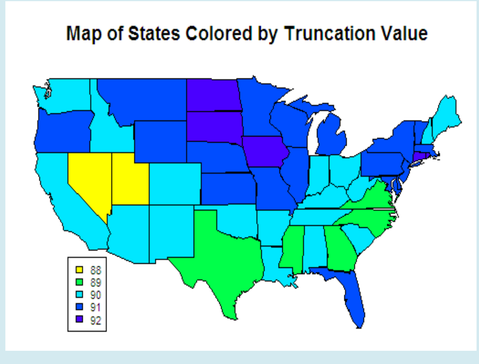
- ACS replaced Census long form
- Given to 3.5 million per year
- Wide range of topics

- Census Bureau gives general results
- PUMS are complete responses to survey
- About 1% of ACS sample in PUMS
- Used for research and allocation of government funds

- Individual privacy in PUMS a concern
- Many techniques to remove identifiers
- Identifiers include extreme age, wage, etc

Top Coding for Age

- Use mean-modified top coding
- Replace ages above chosen truncation point with mean of all ages above truncation point
- Effective for protection of privacy
- Damaging to integrity of individual data at ages above 85 years
- Data over age 85 must be looked at as group to be accurate



- Map of US states color coded by truncation point
- No one uniform national top code
- Top coding done for each individual state

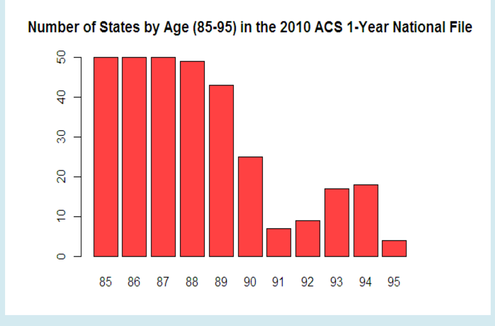
Issues with Top Coding in the National File

- No extreme effect on individual state data from top coding (if user understands the top coding)
- Much larger effect on national file

- Initially appears national file truncated at age 95
- Closer look shows national file is an aggregate of the state files
- Since ages top coded by state, some ages in national file don't contain all states

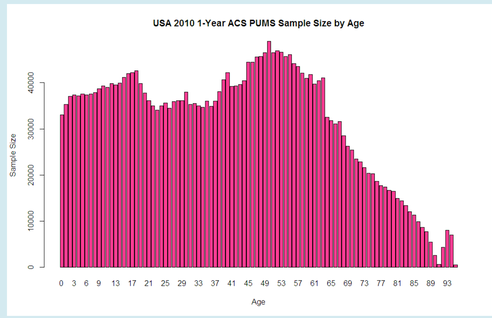
- Geographic effect when looking at ages above 85 years old

States Represented at Each Age



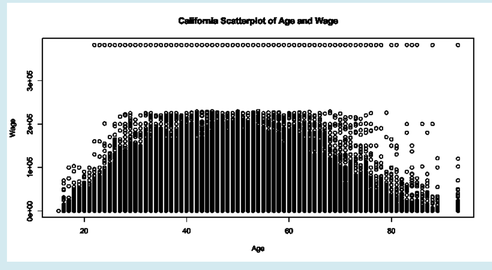
- Ages less than 85 (not pictured) in national file contain all 50 states (and District of Columbia)
- After 85, number decreases
- Only 4 states represented at age 95
- Introduces geographic bias into the data

Sample Size by Age



- US sample size by age pictured above
- Strange bump at tail caused by top coding.

Top Coding for Wage



- California distribution of wages by age
- Cutoff wage $230,000 and top coded placement value 382,000
- Done on state by state basis, altering national file
- Potential issue in individual records top-coded for both age and wage.

- Income variable another issue
- Income variable not top-coded itself, sum of several variables, many of which are top coded.