

# An Initial Experiment on Mixed Categorical and Numerical Data

Alan Karr  
Mi-Ja Woo  
January 31, 2006

## Setting

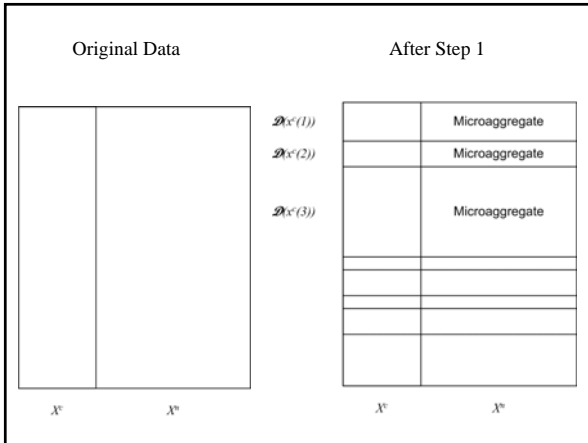
- Data:  $X = (X^c, X^n)$ 
  - $X^c$  = categorical variables
  - $X^n$  = numerical variables
- Problem: selection of microdata release
- Need
  - SDL procedures
  - Utility measure
  - Risk measure

## Outline

- SDL
  - Step 1: Microaggregation on  $X^n$ , conditional on  $X^c$
  - Step 2: Swapping on  $X^c$
  - Step 3: Linear transformation on  $X^n$ 
    - Conditional on  $X^c$
    - Unconditional
- Utility
  - Propensity scores
- Risk
  - ?????

## SDL Step 1

- $x^c$  = observed value of  $X^c$
- $\mathcal{D}(x^c) = \{X_j; X_j^c = x^c\}$
- Do microaggregation on each  $\mathcal{D}(x^c)$  separately
- Advantage
  - Preserves relationships between  $X^c$  and  $X^n$
- Disadvantage
  - $\mathcal{D}(x^c)$  may be too small for microaggregation to affect risk



### SDL Step 2

- Swap (only!)  $X^c$
- Choices
  - Attributes to swap
  - Swap rate

### SDL Step 3

- Use linear transformation on (post-Step 2)  $X^n$  to restore covariance
  - Essentially what Mi-Ja has been looking at
- Two possibilities:
  - Conditional on  $x^c$  (i.e., separately on each  $\mathcal{D}(x^c)$ )
    - Potential disadvantages
      - Swapping has occurred across the  $\mathcal{D}(x^c)$
      - Really want to restore global covariance of  $x^n$
  - Global

### Propensity Score Utility

- Need mixed model here