Measuring Survey Quality through Representativity Indicators using Sample and Population based Information

Chris Skinner, Natalie Shlomo,

Barry Schouten, Jelke Bethlehem

Li-Chun Zhang,

Representativity Indicators for Survey Quality

collaboration between:

national statistical institutes of Netherlands, Norway and Slovenia

& universities of Leuven & Southampton

SEVENTH FRAMEWORK PROGRAMME

Representativity Indicators

- quality indicators for survey non-response

- to supplement response rate
- to measure how well respondents represent population
- tools for use at different stages of survey process (data collection+)

Aim of paper

- to consider estimation of representativity indicators using either:
- **sample-based information** (microdata for both respondents and nonrespondents), or
- **population-based information** (microdata for respondents and aggregate data for population)

How to define representativity indicator?

two approaches

- both based on idea of **response propensity**

RISQ

Response propensity

Idea: R-indicator measures homogeneity of **response propensities:**

ρ_i = probability of response for unit i given values of auxiliary variables

R-indicator

$$R(\boldsymbol{\rho}) = 1 - 2S(\boldsymbol{\rho})$$

where $$S(\boldsymbol{\rho}) = \sqrt{\frac{1}{N-1}\sum_U (\rho_i - \bar{\rho}_U)^2}$$

$R(\boldsymbol{\rho}) = 1$ if ρ_i constant

$R(\boldsymbol{\rho}) = 0$ if $S(\boldsymbol{\rho}) = 0.5$

Schouten, Cobben & Bethlehem (2009, *Survey Methodology*)

Alternative Indicator

$$Q^2(\boldsymbol{\rho}) = [\sum_U \rho_i]^{-1} [\sum_U \rho_i (\phi_i - \bar{\phi})^2]$$

where $\phi_i = \rho_i^{-1}$

proposed by Särndal and Lundström (2008, JOS) in the context of selecting auxiliary variables for weighting adjustment

Estimated R-indicator

Sample-based – auxiliary variables recorded for whole sample

(1) estimate response propensities using e.g. logistic regression

(2) $$\hat{R}(\boldsymbol{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{s} d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)^2}$$

RISQ

Estimated R-indicator

Population-based – auxiliary variables only measured on respondents and in aggregated form for population

(1) Estimate response propensities using ordinary least squares

$$\tilde{\rho}_i^{OLS} = x_i'[N(S_{xx} + \overline{x}\overline{x}')]^{-1} \sum_r d_i x_i$$

if population covariance matrix S_{xx} known.

Estimate S_{xx} from respondents if only population mean vector $\overline{x}$ known

RISQ

Estimated R-indicator

Population based (continued)

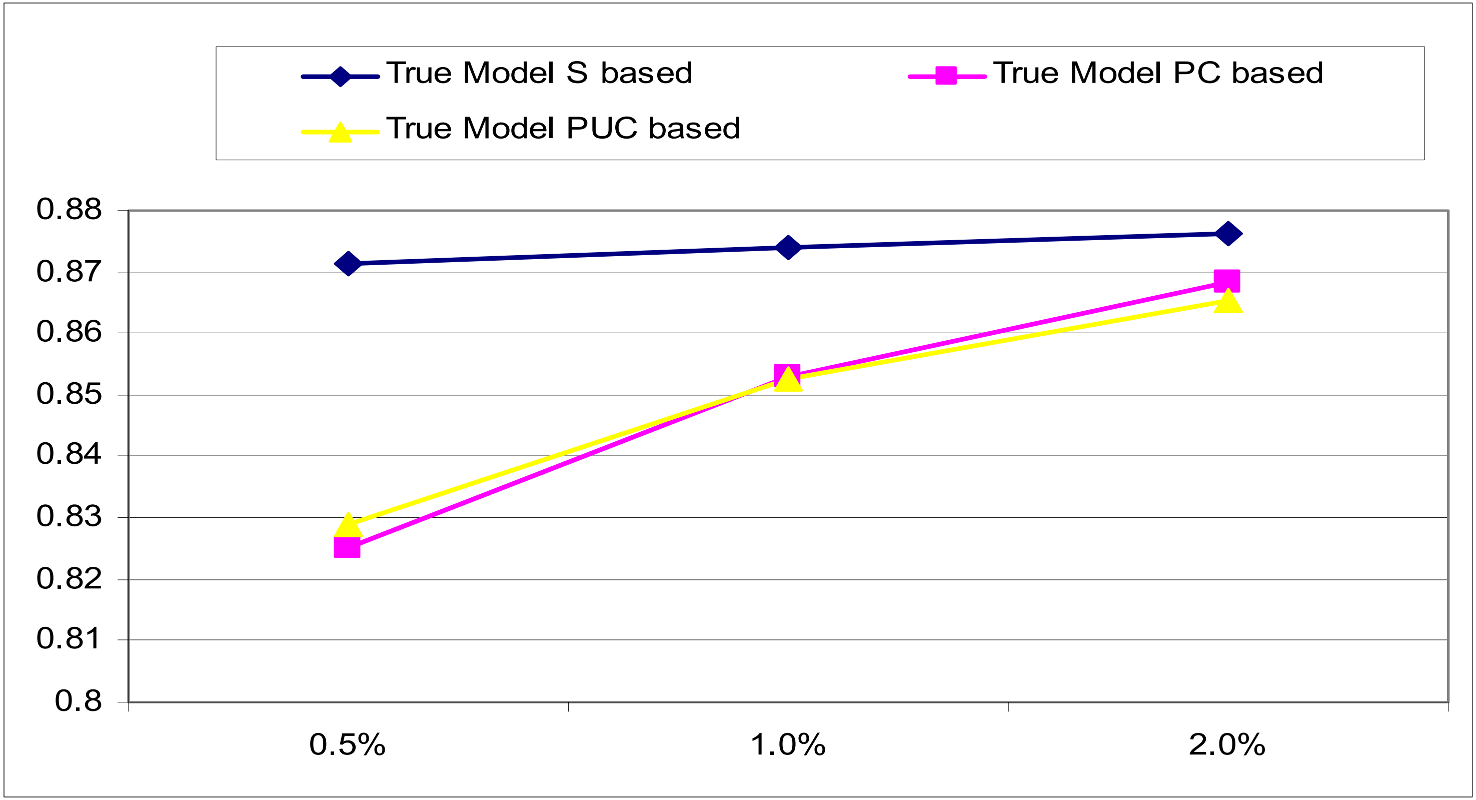
(2) $$\hat{R}_r(\boldsymbol{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_r d_i \hat{\rho}_i^{-1}(\hat{\rho}_i - \bar{\hat{\rho}}_r)^2}$$

RISQ

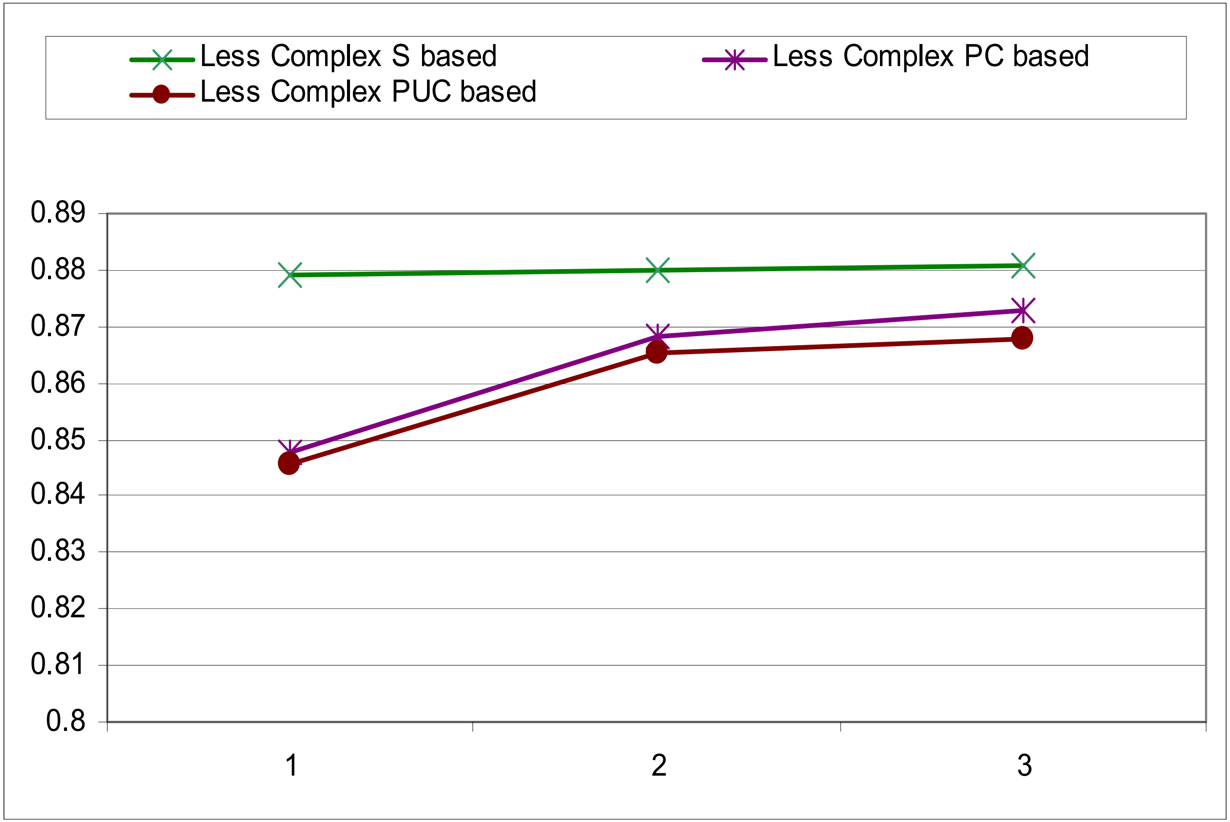
Simulation Study

- Samples from Israel census data on 753000 individuals
- ‘realistic’ sampling with fractions 1:50, 1:100, 1:200
- ‘realistic’ non-response based on type of locality, household size, children in household – overall response rate 82%

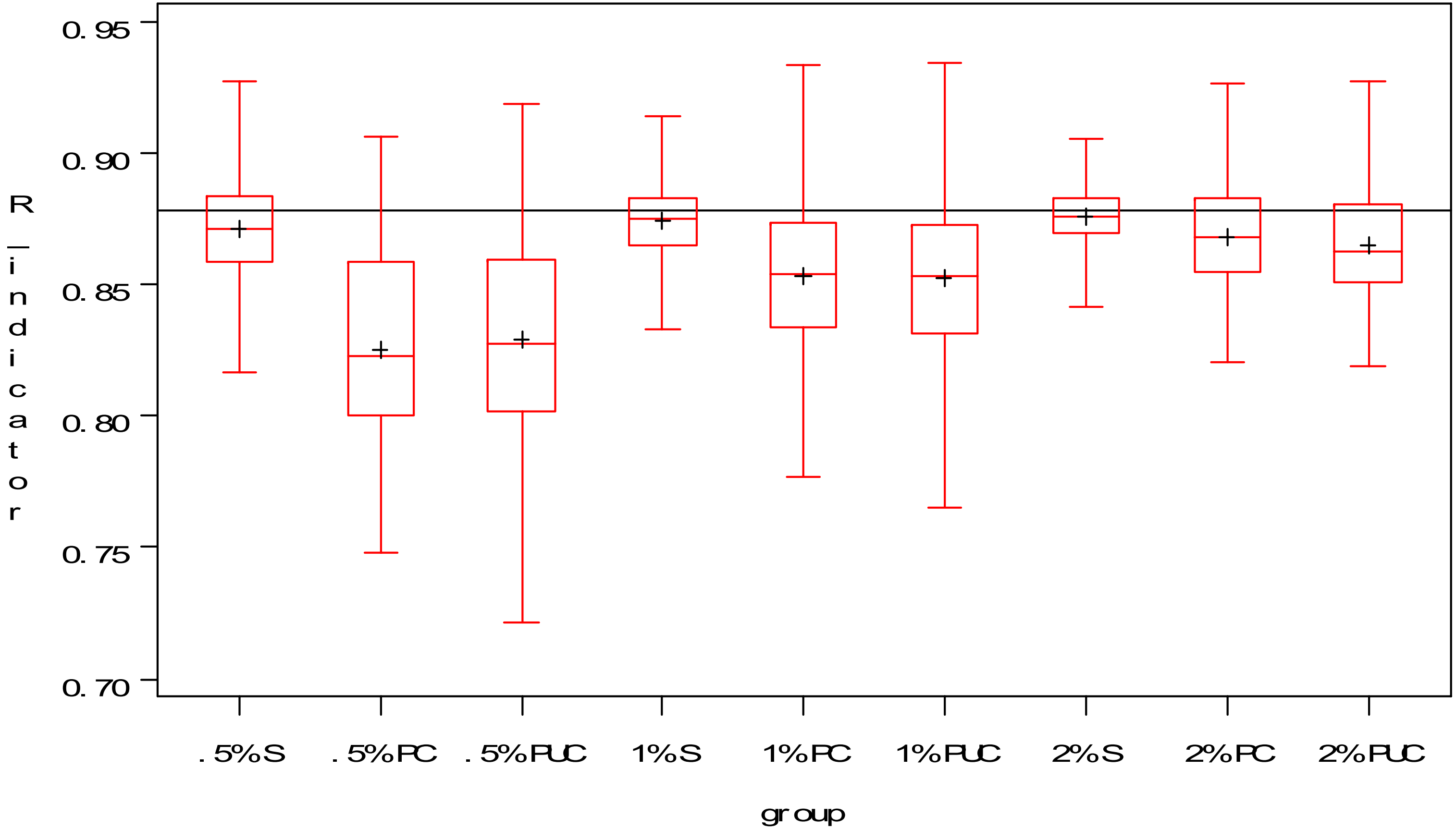
Simulation Means of $\hat{R}(\rho)$ for Sample (S), Population with Known Covariance matrix (PC) and Population with Unknown Covariance Matrix (PUC) of Auxiliary Variables – True Model



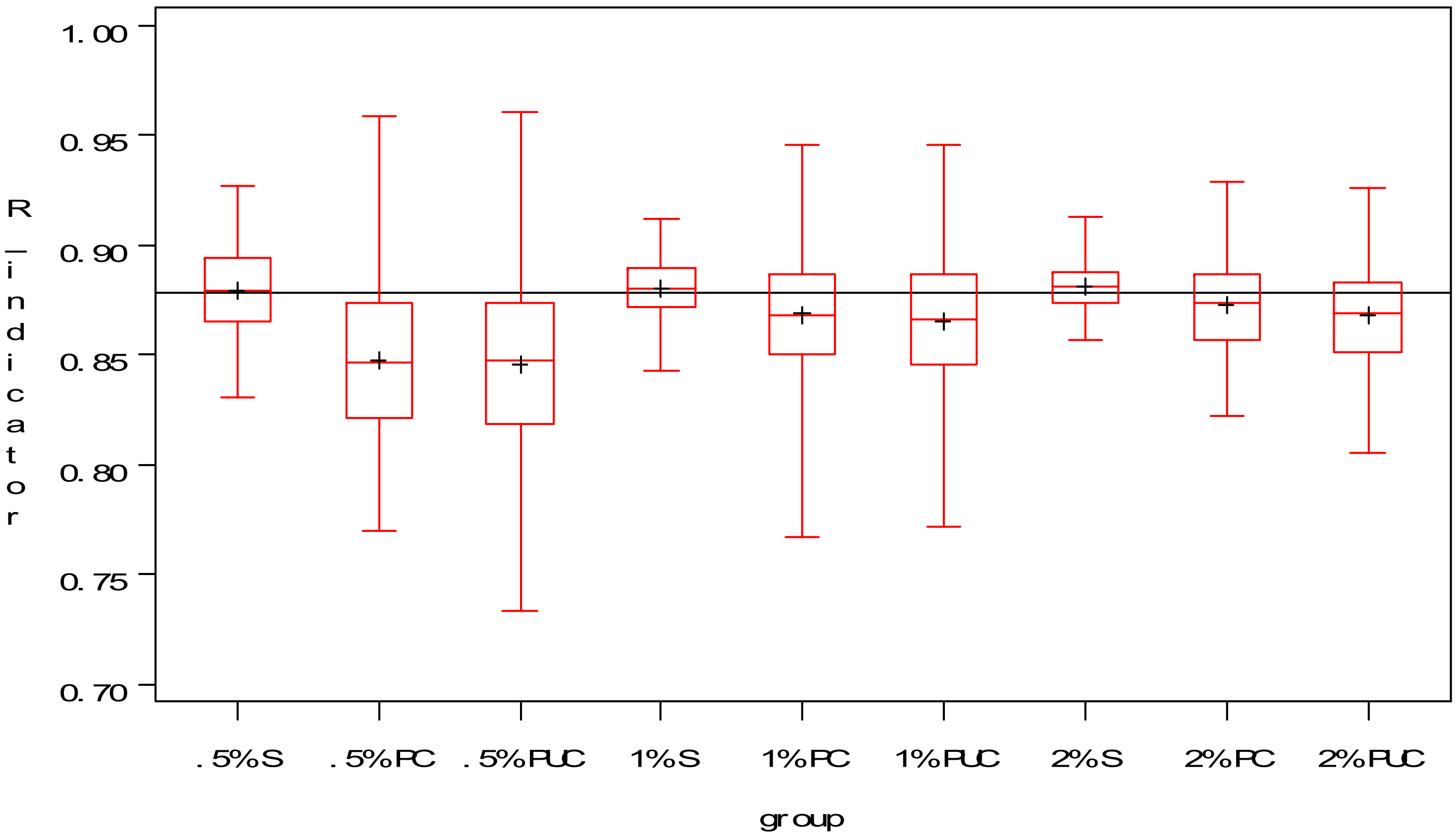
Simulation Means of $\hat{R}(\rho)$ for Sample (S), Population with Known Covariance matrix (PC) and Population with Unknown Covariance Matrix (PUC) of Auxiliary Variables - Less Complex Model



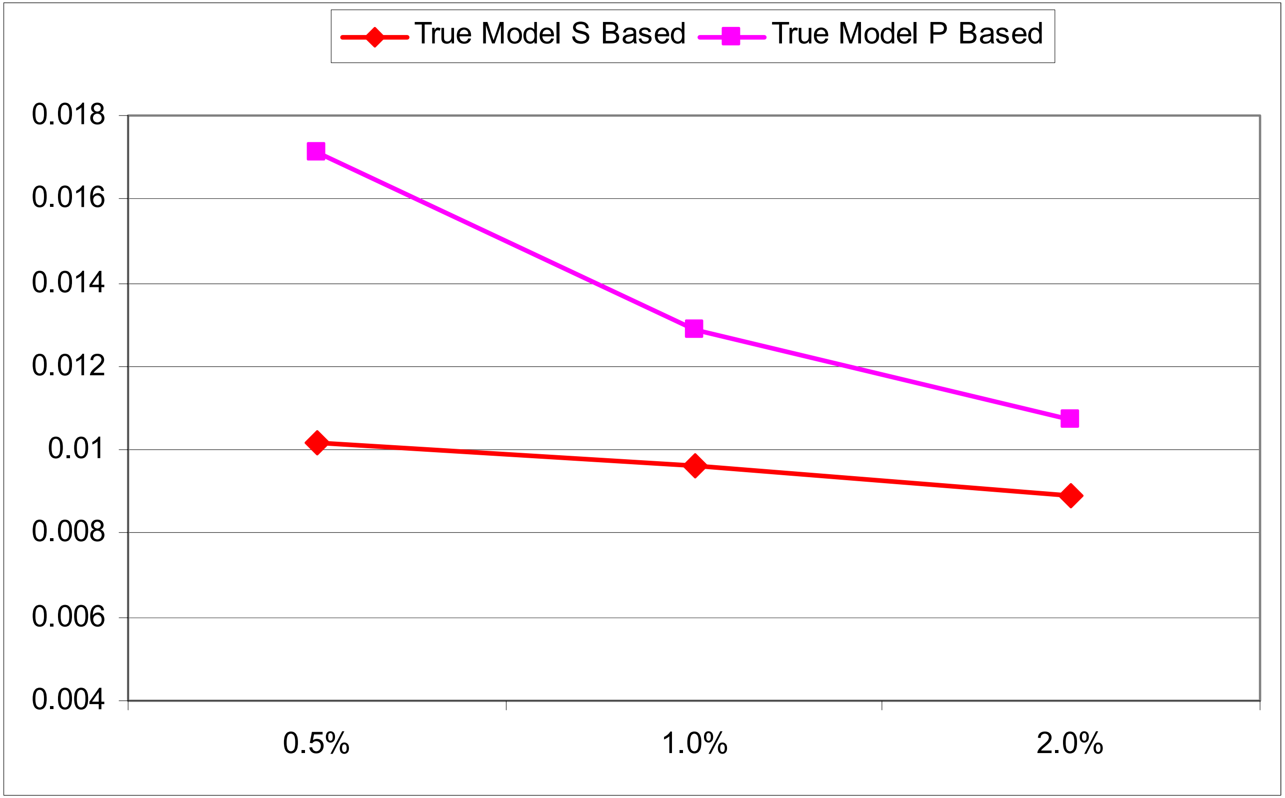
$\hat{R}(\rho)$ for Sample (S), Population with Known Covariance Matrix (PC) and Population with Unknown Covariance Matrix (PUC) of Auxiliary Variables – True Model



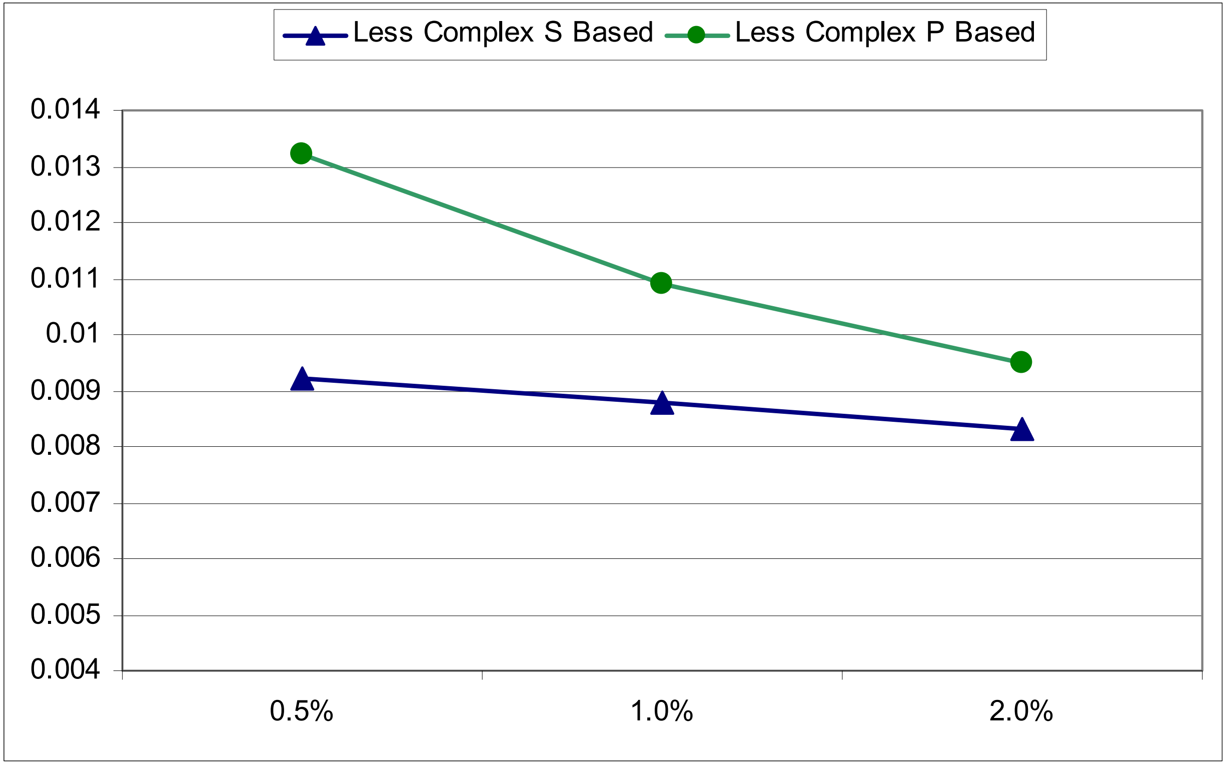
$\hat{R}(\rho)$ for Sample (S), Population with Known Covariance Matrix (PC) and Population with Unknown Covariance Matrix (PUC) of Auxiliary Variables – Less Complex Model



Simulation Means of q^2 for Sample (S) and Population (P) Based Auxiliary Variables – True Model

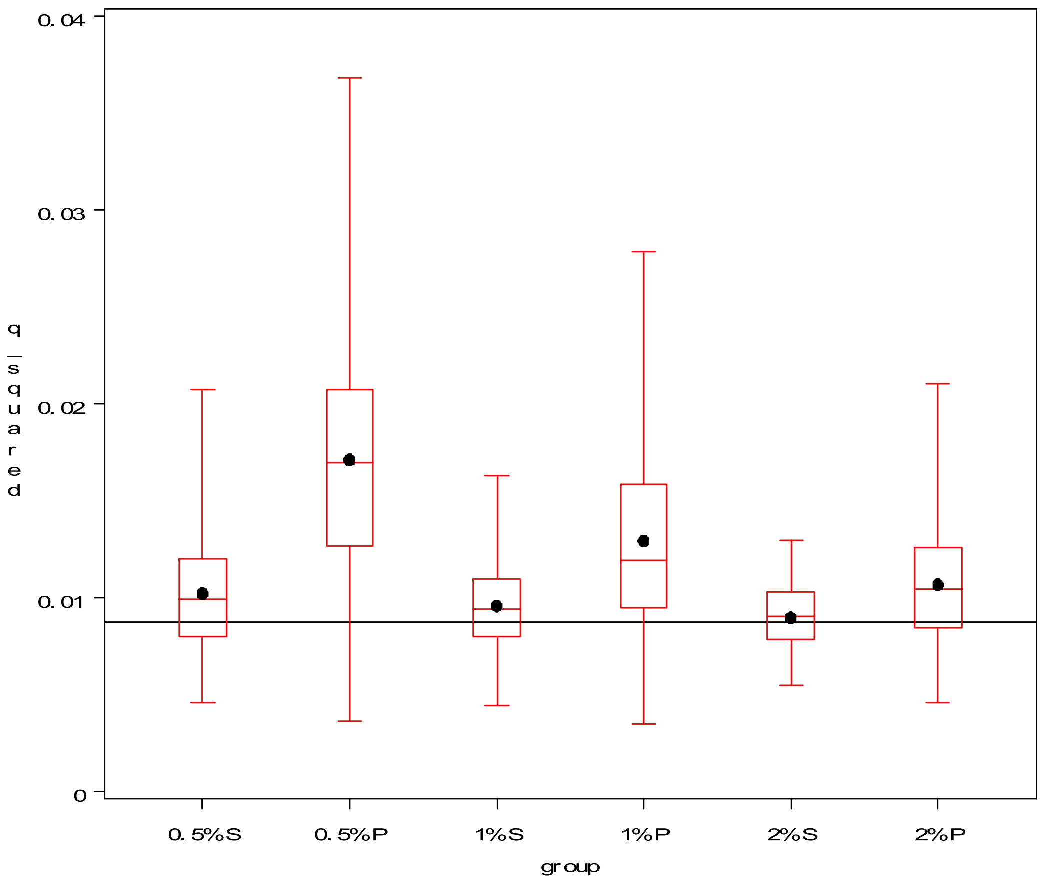


Simulation Means of q^2 for Sample (S) and Population (P) Based Auxiliary Variables - Less Complex Model



RISQ

q^2 for Sample (S) and Population (P) Based Auxiliary Variables - True Model



RISQ

Conclusions

- representativity indicators defined in terms of response propensities
- can estimate accurately given either sample-based or population-based information
- have also applied to surveys from RISQ countries
- and have examined bias-corrected estimators and estimators of standard errors of estimators