E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

UNIVERSITY OF Southampton

Southampton Statistical Sciences Research Institute

Assessing risk of nonresponse bias and dataset representativeness during survey data collection

Gabriele Durrant

Joint work with Jamie Moore, Solange Correa and Peter W.F. Smith
University of Southampton

University of Manchester, seminar series, 1 March 2016

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

The Research Project

- Research was funded by the
  - ESRC research grant on “Paradata”
  - ESRC National Centre for Research Methods, Workpackage 1 “Data Collection for Data Quality” and the
  - ESRC Administrative Research Centre for England (ADRCE).

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

My Current Research

- Analysis of length of response (response latencies) (with Olga Maslovskaya and Patrick Sturgis)
- Consent to data linkage; consent and data linkage representativeness (with Jamie Moore and Peter Smith)
- Interviewer effects on nonresponse bias (with Denize Barbosa and Peter Smith)
- Use of a Bayesian approach to adaptive survey design (with Eliud Kibuchi, Natalie Shlomo, Patrick Sturgis)
- Use of mobile devices (not properly started)
- Analysis of environmental behaviour (household effects) (with Vivian So and Peter Smith)

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Outline

- Background and Introduction
- Measuring risk of nonresponse bias during data collection: Representativeness Indicators
- Data
- Application and Results
- Summary
- Implications for Survey Practice

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Background

- Aim previously: maximise response rates
- But response rates decreasing, also low association with bias
- Also data collection costs increasing
- Focus now on nonresponse bias

Introduction

- Key question: How to monitor, assess and minimise (risk of) nonresponse bias?
  - Post or during data collection

- Questions from survey practice: when to stop calling and who best to follow up?

- Aim here: Trajectories of (risk of) nonresponse bias or representativeness over data collection, to inform adaptations to maximise quality and / or minimise costs

Introduction

- Fully observed information on both respondents and nonrespondents may be necessary. Sample frame information from
  - register / Census
  - administrative data
  - previous wave

- Datasets (face-to-face surveys):
  - Census nonresponse link study
  - Understanding Society

UNIVERSITY OF Southampton

Southampton Statistical Sciences Research Institute

Representativeness Indicators

How to assess the risk of nonresponse bias?

- Main idea: measure similarity between sample data obtained and frame data in terms of variation in response rates
- Use of a response propensity model to obtain estimated response propensities (information needed on entire sample frame)
- Representativeness indicators: estimate variation in these response propensities (SD = Standard deviation of the response propensities)
- Low variability in response propensities imply high representativeness

Representativeness Indicators

- R indicator: $R = 1 - 2SD$

  Ranges between 0 and 1

  Close to 1 indicates high representativeness

- CV (Coefficient of Variation): $CV = \frac{SD}{r}$

  SD= standard deviation of response propensities

  r = response rate

  CV close to 0 indicates high representativeness

- Here computed at each call

Considerations and Properties

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- Indicators (specific to attribute variables) decomposable to assess variation (conditionally) associated with different variables/ categories
- Monitoring during data collection: at each call response propensity model must be fitted
  - Indicators are covariate specific; same covariate set must be used in the underlying response propensity model
  - However, at the beginning of data collection lower response rates, hence may expect model selection to retain fewer covariates, hence the 'optimal' model might change during data collection

Considerations and Properties

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- We investigate the sensitivity to different model specifications. Select covariates in the model
  - after 5 calls
  - after 20 calls (at end of data collection)
  - no model selection (use of all variables)
- Monitoring of response within and between sample frame subgroups

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Applying these Methods – Key Research Objectives

1. Visualise dataset representativeness trajectories for different surveys of the same population
2. Representativeness trajectories may differ between surveys- are they generalizable across surveys?
3. Evaluate the utility of census derived attribute information for assessing survey dataset representativeness.
4. Are indicators sensitive to different model specifications?
5. Can adaptive collection strategy stopping points be generalised?

UNIVERSITY OF Southampton

Southampton Statistical Sciences Research Institute

Data

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Data

- ONS 2011 Census Non-Response Link Study (CNRLS)
- Uniquely, linking response indicator from three UK social surveys to survey call record data and census household (HH) attribute information on sample frames
- Information on both respondents and nonrespondents available
- 3 (cross-sectional) surveys:
    - Labour Force Survey (LFS) (wave 1).
    - Life Opportunities Survey (LOS) (wave 1).
    - Opinions Survey (OPN).

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Data

- Automated and (if imperfect) clerical linkage of HHs to 27th March 2011 Census records.
    - linkage rates >95%, so high sample frame coverage
    - rich suite of attribute variables available
    - interviewing took place within 3 months of census day

Quantifying representativeness trajectories

- Interviews face to face, up to 20 calls to a household
- Successful interview = response (else non-contact or refusal).
- For each survey, visualise *CV*s and partial variants at each call (1 to 20).
- Estimate response propensities using 10 HH attribute variables (also compute partial indicators for):

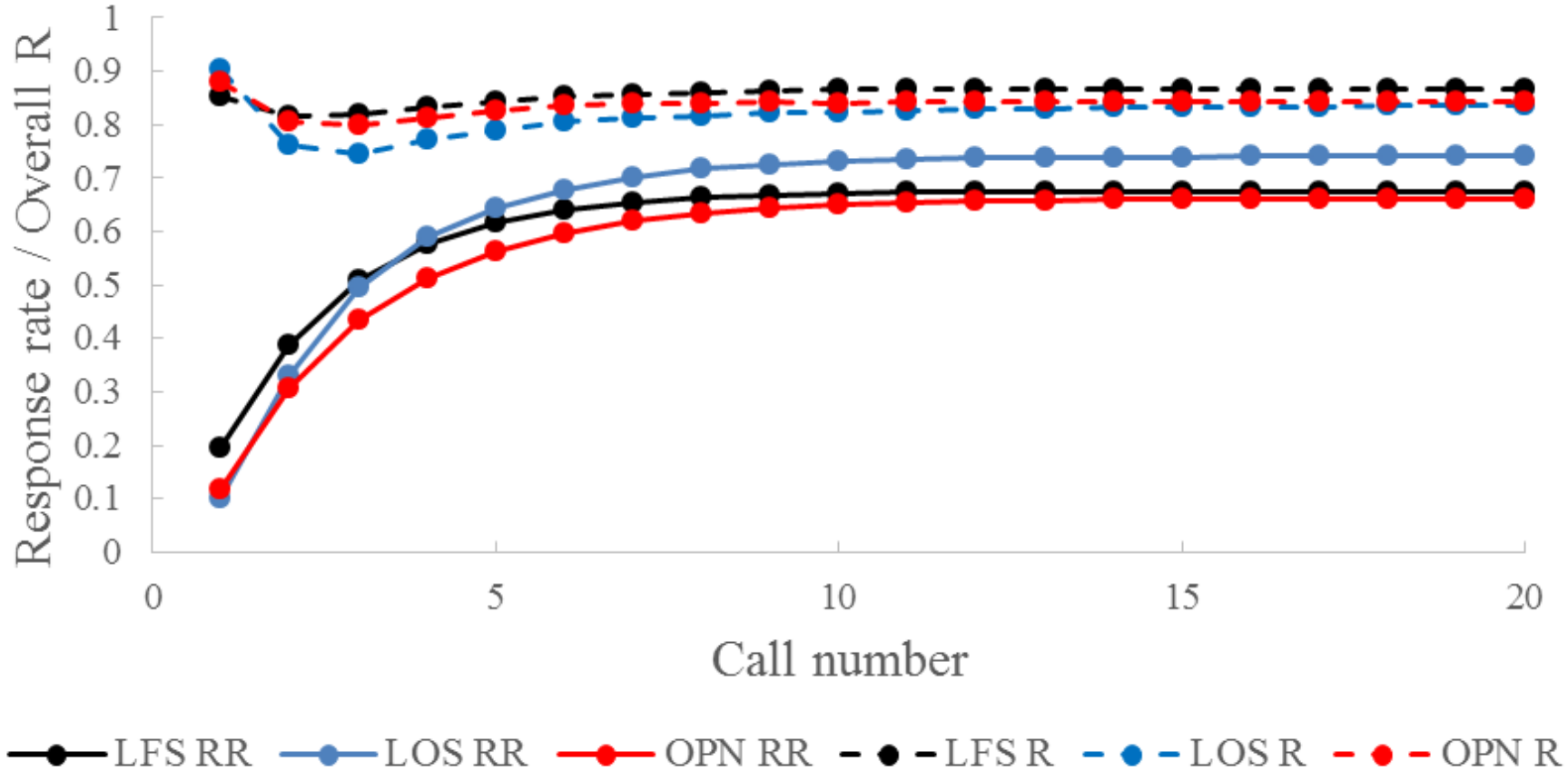
HH Economic Status, HH Structure, Accommodation type, Tenure type, Cars available, Ill Health individual in HH, Impaired individual in HH, Retiree in HH, English fluency in HH, Located in London / SE

UNIVERSITY OF Southampton

Southampton Statistical Sciences Research Institute

Application and Results

R indicators

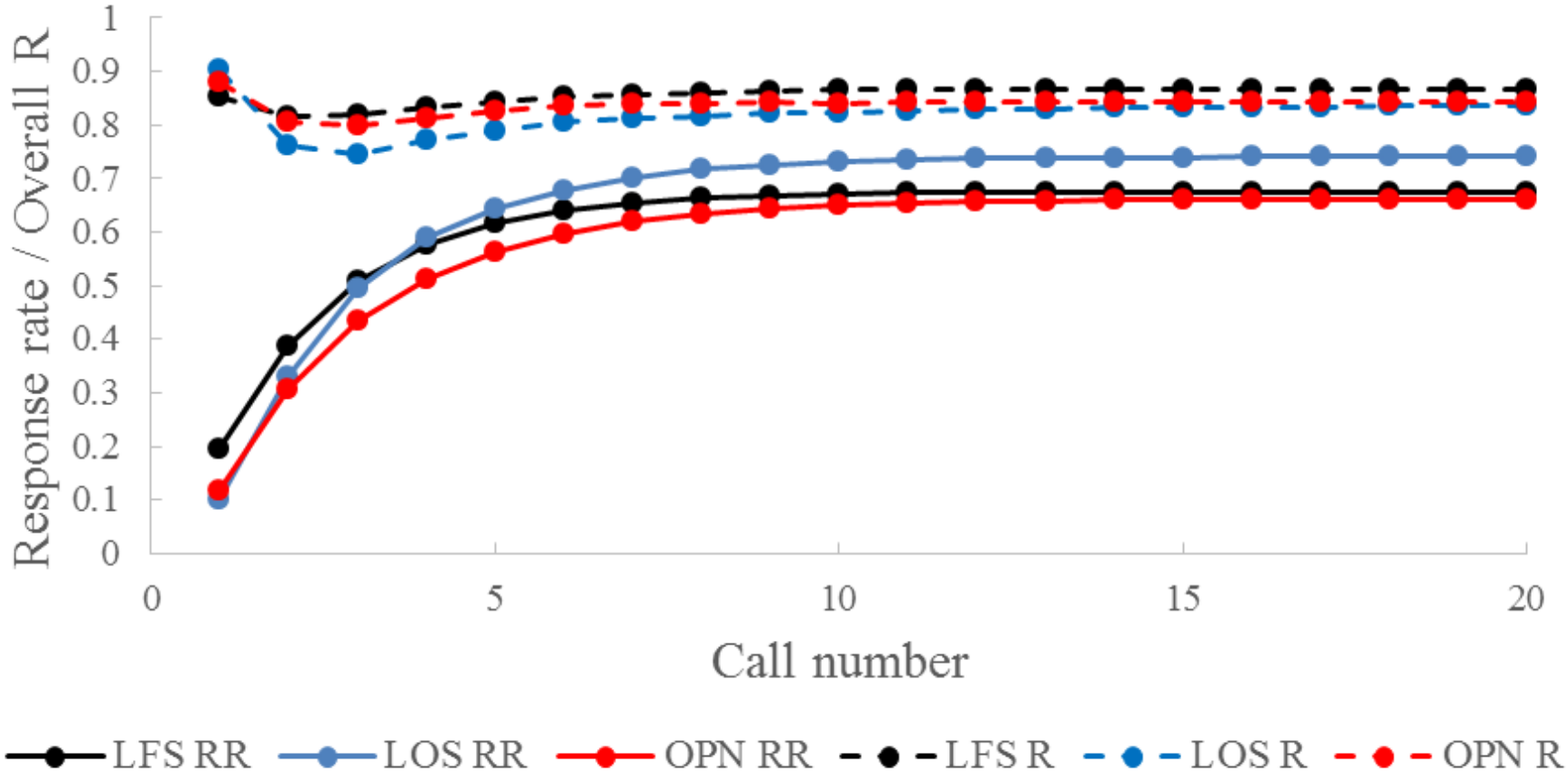


LFS: N = 18,997 final r = 65.7%

LOS: N = 6,469 final r = 70.1%

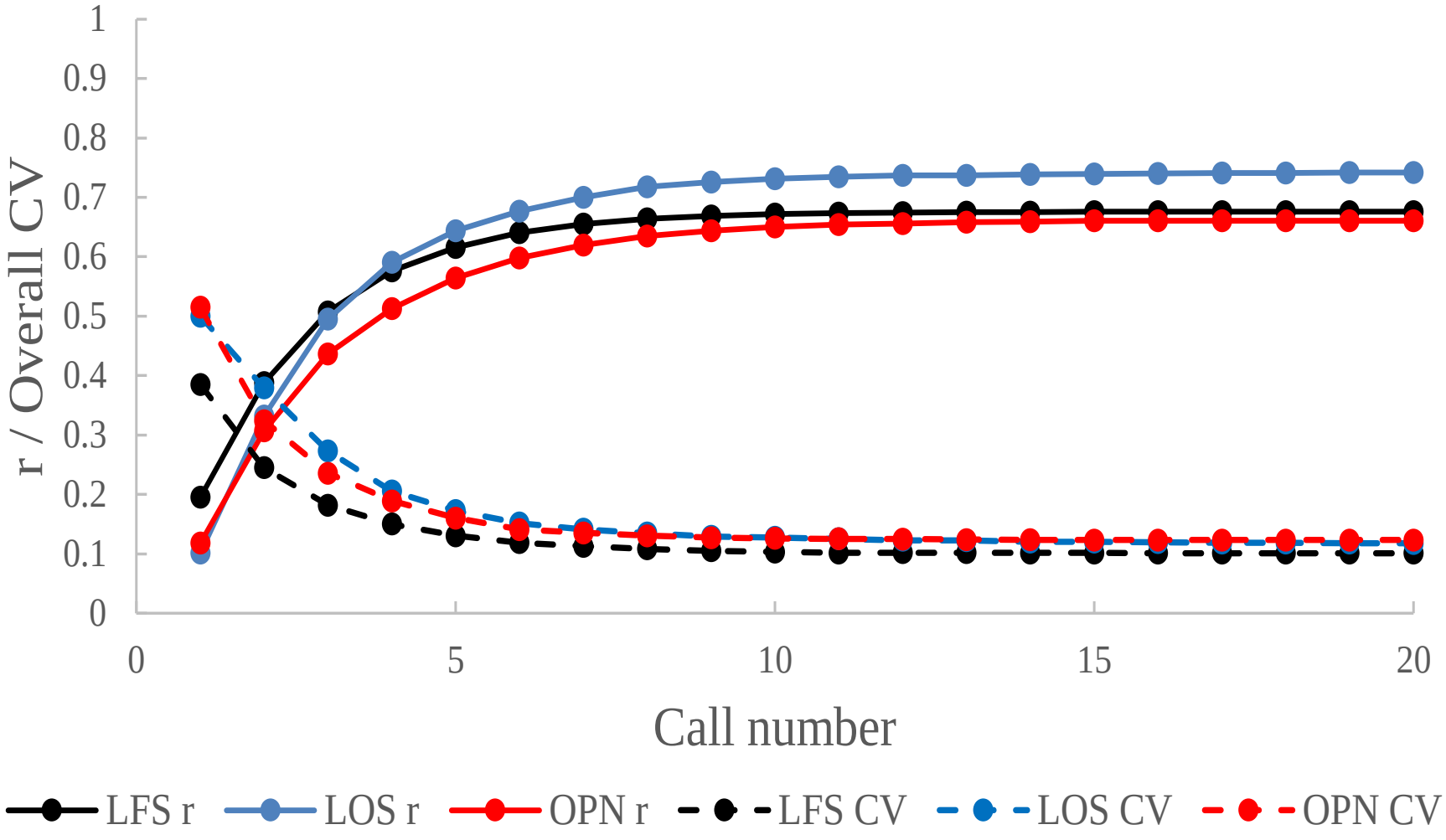
OPN: N = 6,249 final r = 64%.

R indicators



- In case of low response rates (as is the case early on in data collection) small response propensity variation, limited potential for response propensity divergence
- R indicators close to 1, falsely indicating high representativeness
- R-indicator can be misleading in this case

CV (Coefficient of Variation)



- CV standardises SD by r; overcomes the problem of the R indicator
- CV decreasing, close to 0 indicating high representativeness

(Unconditional) Partial Indicators

- Aim: estimate the extent to which response is representative with respect to a covariate or a particular category

(Unconditional) Partial CVs

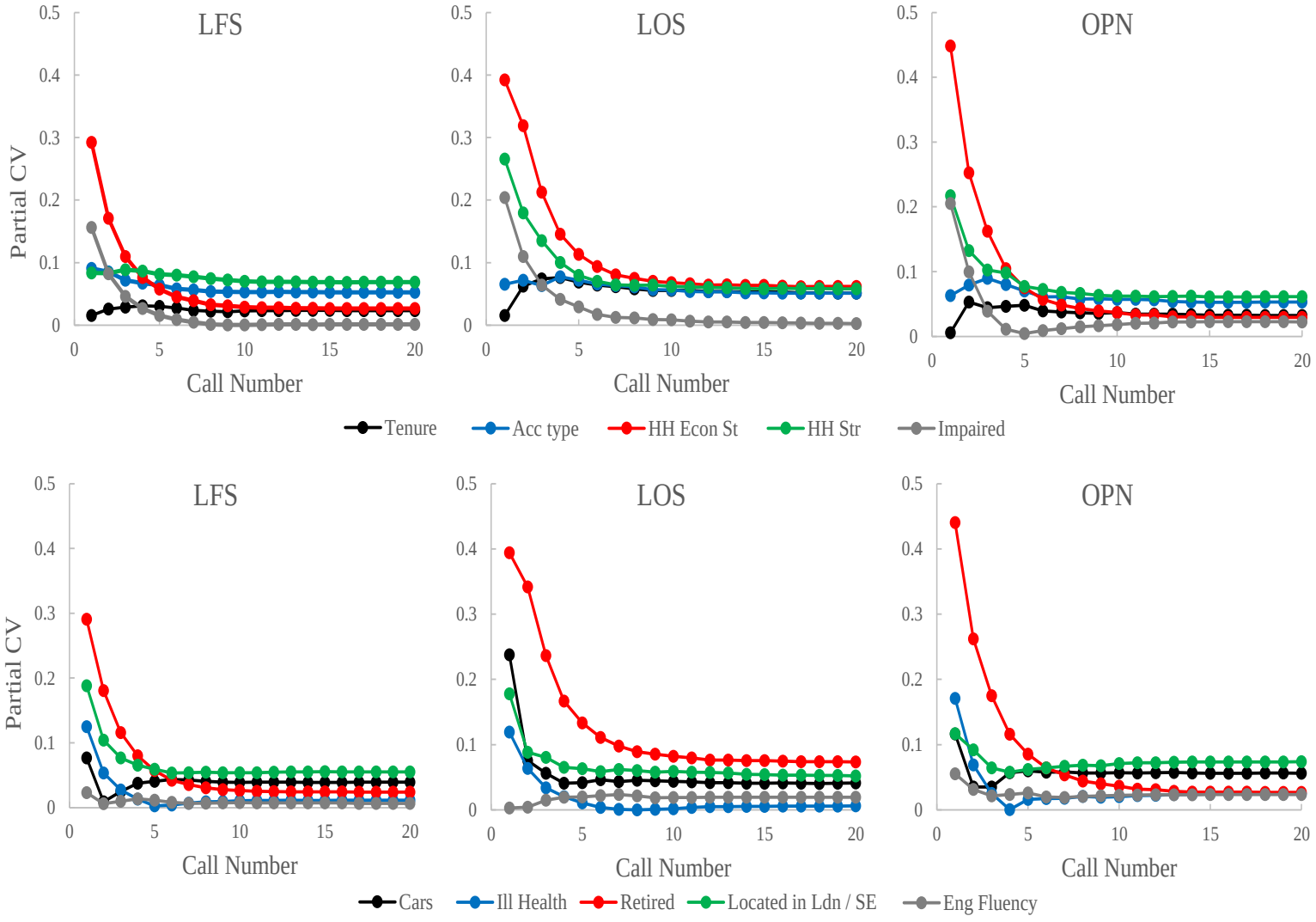
- By covariate Z (with K categories):

$$\widehat{CV}_u(Z, p_x) = \frac{\sqrt{\frac{1}{n}\sum_{k=1}^{K} n_k(\hat{\bar{p}}_k - \hat{\bar{p}})^2}}{\hat{\bar{p}}}$$

- By category k (of covariate Z):

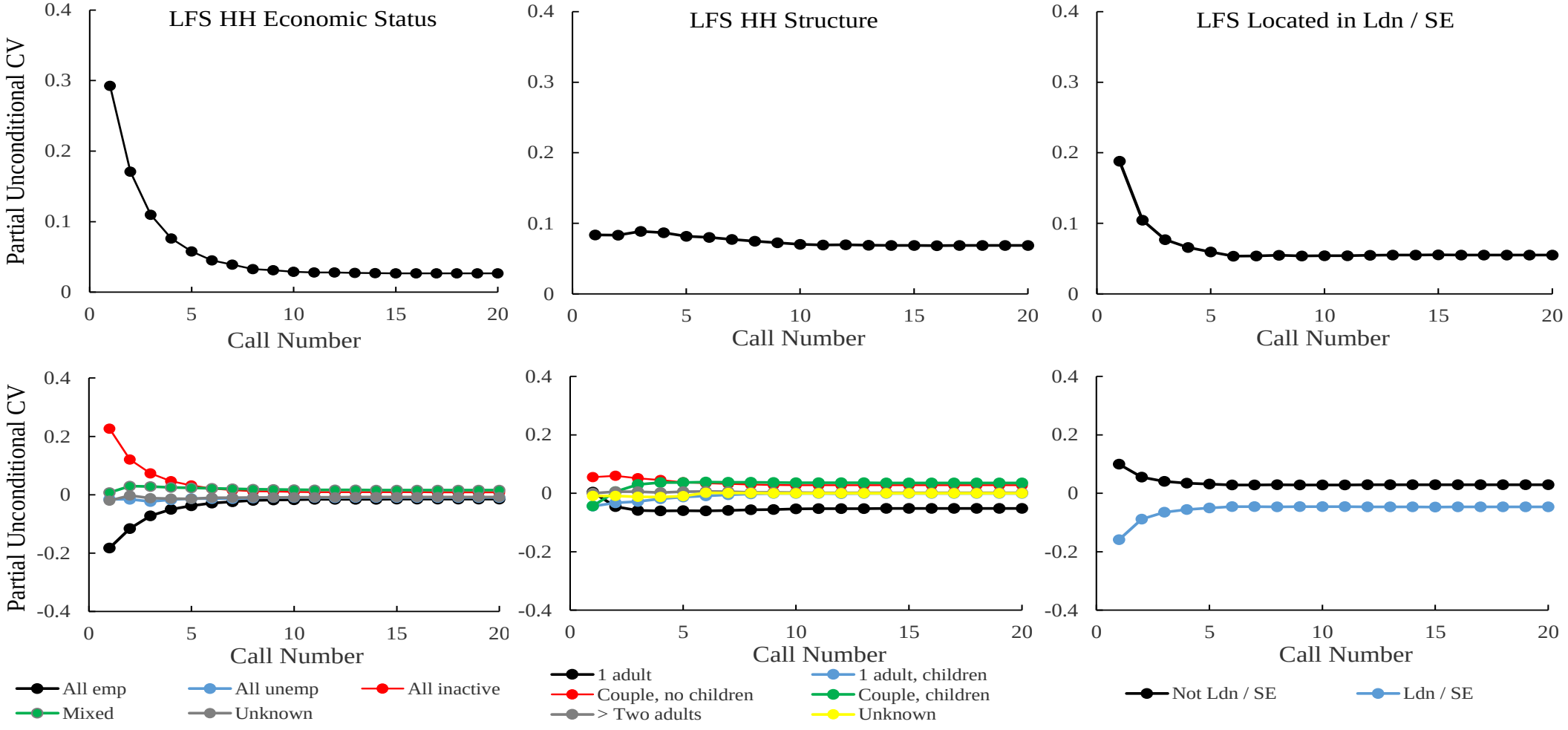
$$\widehat{CV}_u(Z_k, p_x) = \frac{\frac{n_k}{n}(\hat{\bar{p}}_k - \hat{\bar{p}})}{\hat{\bar{p}}}$$

(Unconditional) Partial by Variable CVs



- Surveys similar

Variables with Substantial Impacts (in LFS) - (Unconditional) Partial by Category CVs



UNIVERSITY OF Southampton

Southampton Statistical Sciences Research Institute

Phase Capacity or Stopping Points

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Stopping or Phase Capacity Points

- When to change a survey data collection method?
- When to stop calling?

Stopping or Phase Capacity Points

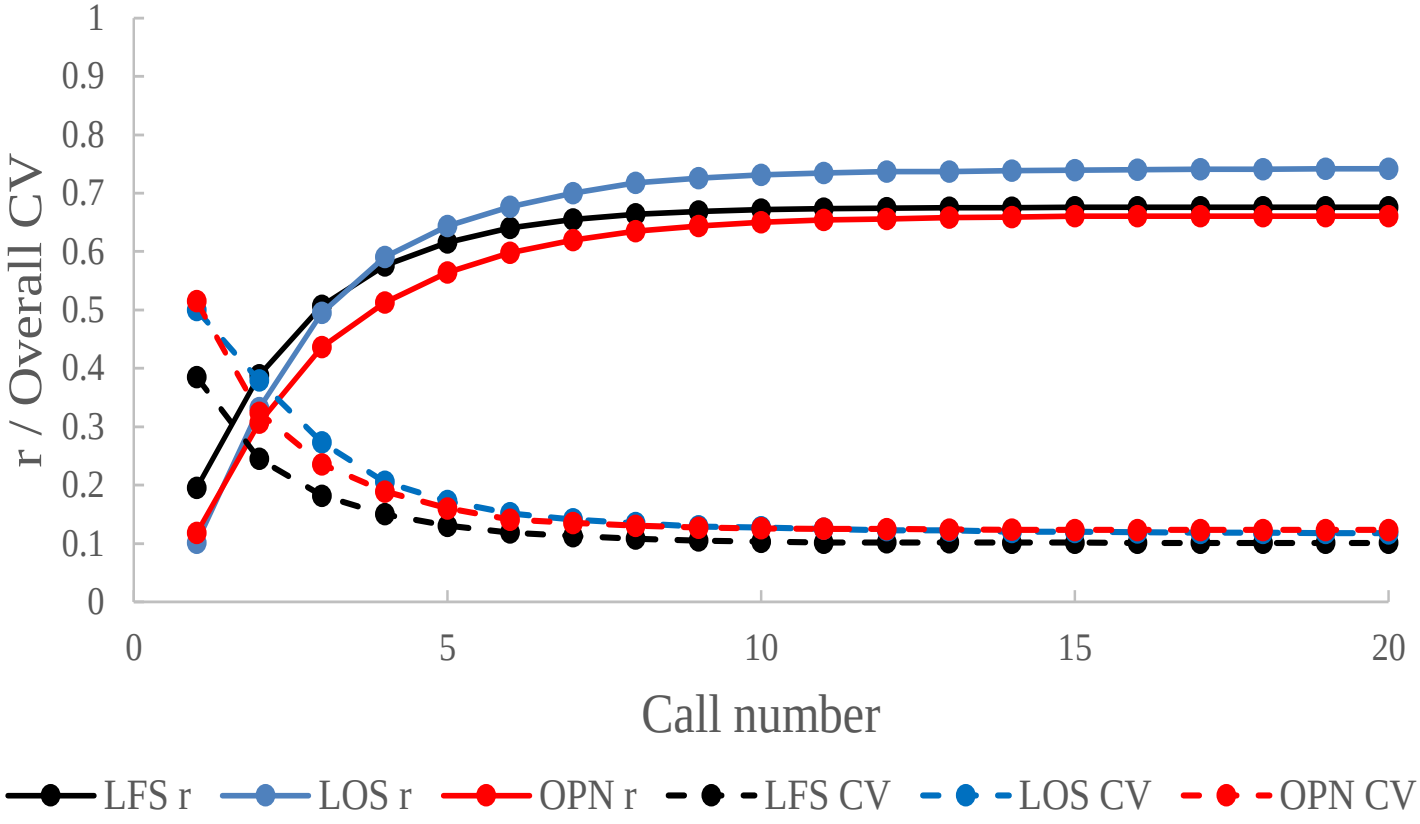
UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- Monitoring response during data collection to identify points when benefits of continuing with the existing data collection method are minimal
  - to modify the data collection method (phase capacity point) or
  - to stop data collection completely (stopping point)

- Either post data collection to inform future data collection methods (adaptive strategy) or to optimise the current data collection period (responsive strategy)

- Can points be generalised?

Stopping or Phase Capacity Points



- Adaptive Strategy: stop when indicator within 0.02 of minimum value (points later when threshold decreased)
- Responsive strategy: stop when indicator within 0.02 of previous value

Stopping or Phase Capacity (PC) Points

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- Overall:

Survey	PC point (adaptive)	% calls saved	PC point (responsive)	% calls saved
LFS	6	8%	5	12%
LOS	8	15%	7	18%
OPN	6	13%	6	13%

- By variable (adaptive):

Survey	HH Economic Status	Located in Ldn / SE	HH structure	Retiree in HH
LFS	6	4	1	6
LOS	7	4	6	8
OPN	7	4	5	8

UNIVERSITY OF Southampton

Southampton Statistical Sciences Research Institute

Further Evidence from Understanding Society

Understanding Society Data

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- Analysis sample
  - Individuals who responded at wave 1 and are eligible at wave 2 (47,000 individuals)
  - Use wave 2 call outcome to assess nonresponse bias at each call for wave 2 for a range of survey variables as measured at wave 1
- Risk of nonresponse bias measured with respect to Wave 1 variables
- Response indicator=1 if call outcome is 'any interviewing done', and 0 otherwise

Survey Quality Indicators

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- Commonly used for nonresponse bias monitoring
  - Response rate
  - Nonresponse bias (absolute and relative)
- Proposed approach
  - Dissimilarity indices (Delta index)
  - Basic idea: compare two distributions (those for respondents and those if everyone had responded)
- Comparison to
  - Coefficient of Variation (CV) (also partial CV)

Dissimilarity Index: Categorical

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- Delta index (Agresti 2013)

$$\Delta_z = \sum_{k=1}^{K} \left| \hat{\pi}_{z,k} - \pi_{z,k} \right| / 2$$

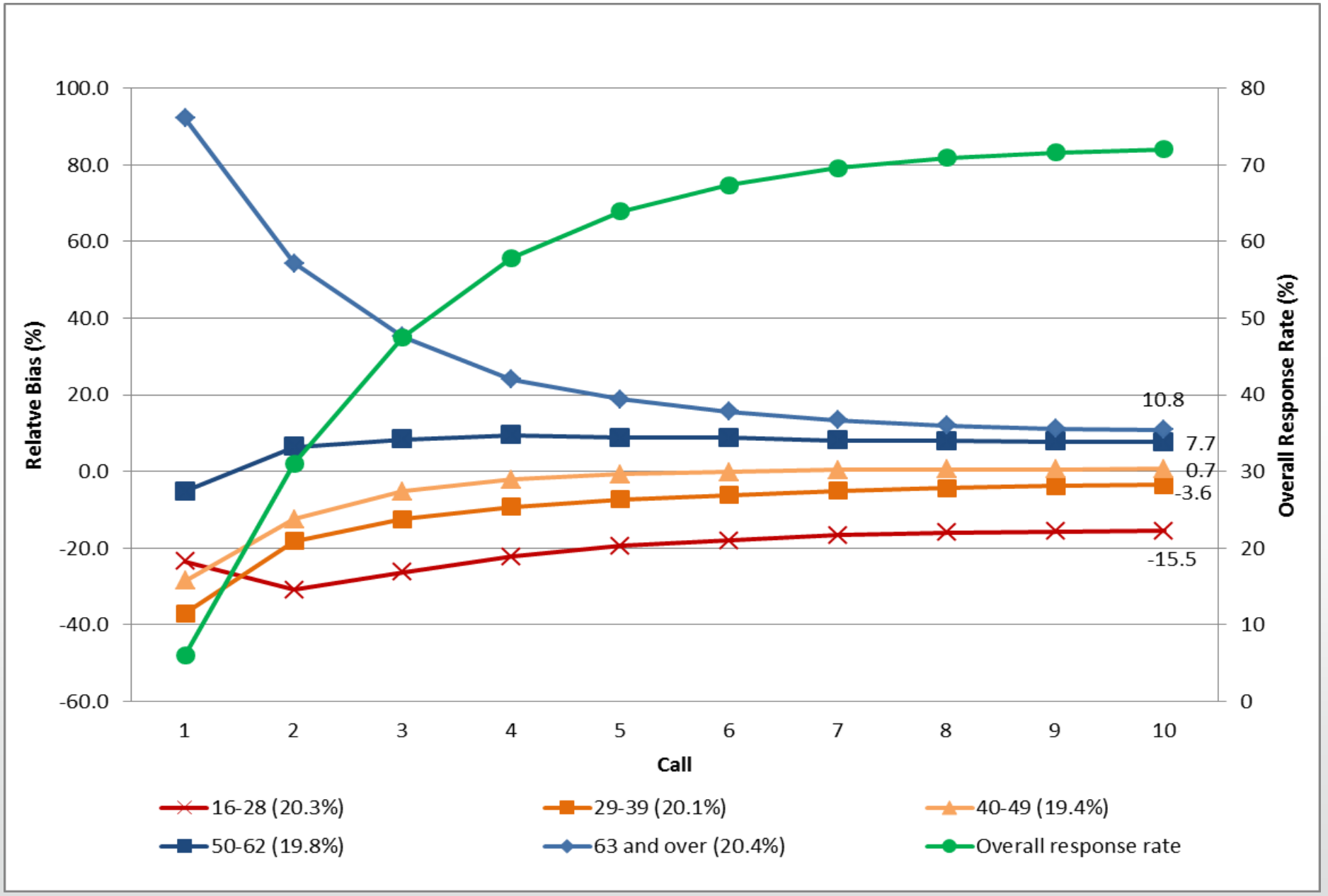
$\hat{\pi}_{z,k}$ observed proportion in category k of survey variable z

$\pi_{z,k}$ corresponding expected proportion

- ranges from 0 to 1
- the higher the delta index the more dissimilar is the estimated distribution to the true distribution
- values below 0.03 may indicate similarity (negligible nonresponse bias)
- no model required

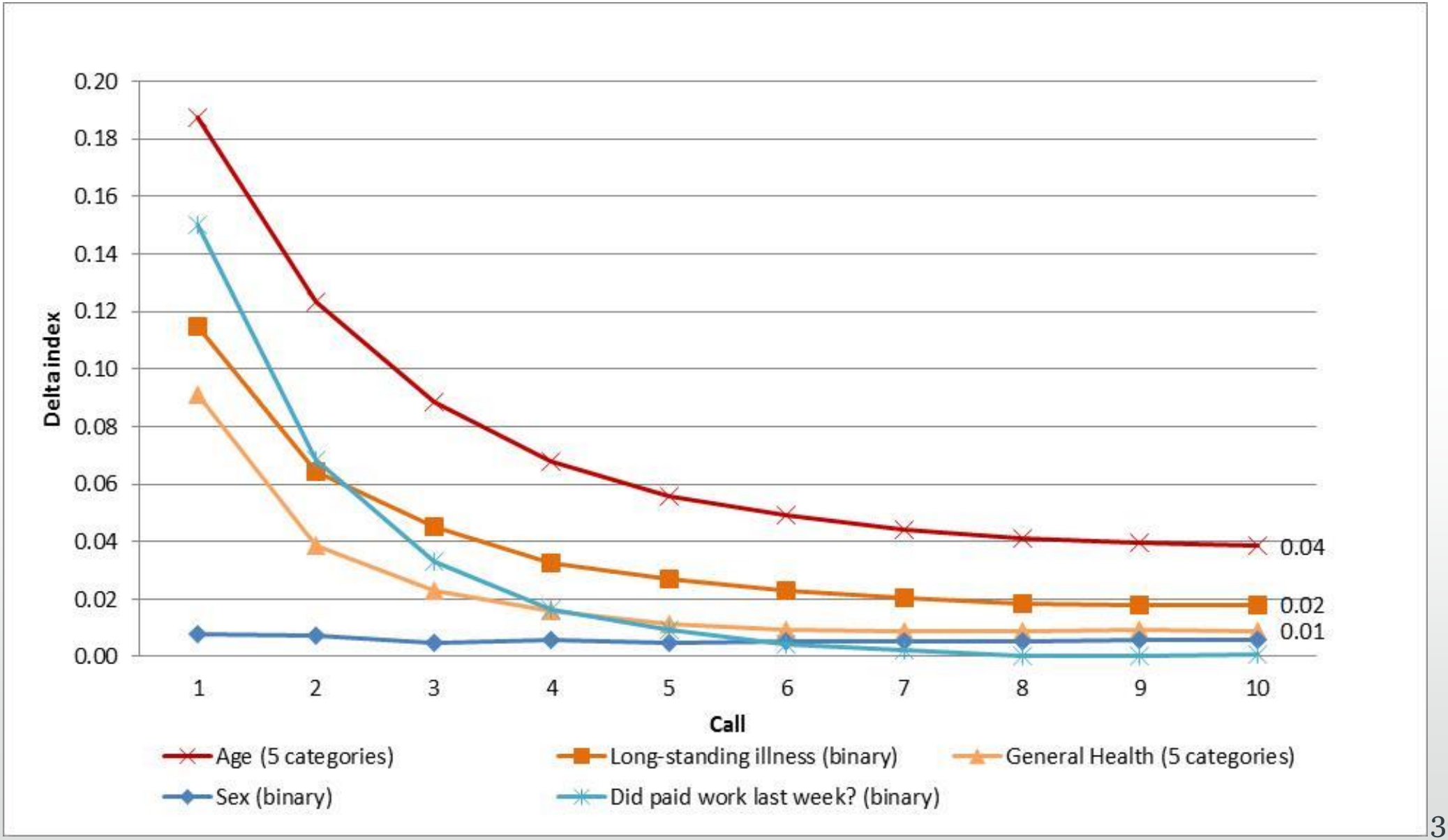
Relative Bias: Age group

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute



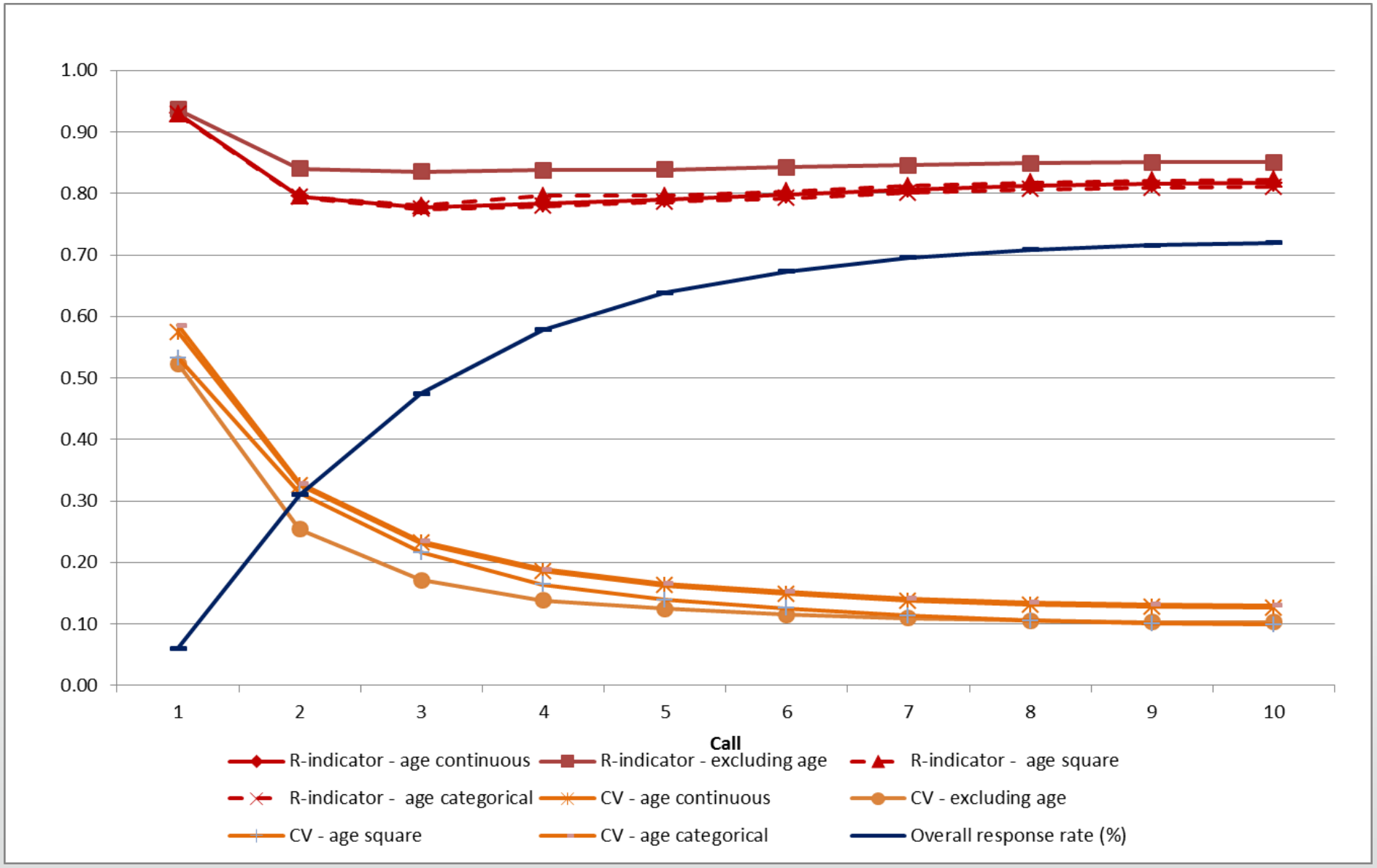
Delta Index
Binary and Categorical Variables

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute



Response Rate, R-indicator and CV

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute



UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Results

- Can monitor categorical variables with several categories
- Allow monitoring of several variables in the same graph
- Dissimilarity index (Delta index) very similar to CV
- Delta index does not require the fit of a model at every call

Summary

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- Representativeness increases at a decreasing rate over call records similarly in the three surveys
  - Sources of non-representativeness are under-representation of economically active HHs, HHs located in London / SE, and single adult HHs
- Despite the above, data collection stopping points differ (slightly) between surveys
- The census is a good source of attribute information for assessing dataset representativeness (when surveys are within 3 months of its date)
  - Its utility further from this date would need to be studied; hence, in this case we may not recommend that strategies can be generalised

Summary

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- Effects of covariates in the model:
  - Empirical results indicate (as would be expected) that there is an effect on the CV depending on underlying model
  - If model selection after 5 calls then generally less variables included, CVs smaller, indicating higher representativeness (in comparison to no model selection)
  - Also possible that variables significant at the beginning but insignificant at the end
  - Might be best to include as many variables as possible and be consistent in including them, to allow CV comparisons
- Results for CV very similar to Dissimilarity Indices – reassuring

Implications for Survey Practice

- Number of calls could be reduced
- Implications for cost savings without potentially much loss of data quality

Future Work

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- Explore the use of a multilevel model (e.g. taking into account interviewer clustering) – does this improve the response propensity model?
- Extend the use of representativeness indicators to linked datasets
  - Estimating the risks of non-consent / non-linkage biases
  - Estimating the risk of non-consent / non-linkage bias in biosocial data

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Upcoming RSS event

- RSS event: Thursday 3 March 2016, 5.00pm - 6:30pm
- Title: Maintaining high response rates - is it worth the effort?
- Presenters: Patrick Sturgis, Gabriele Durrant and Joel Williams
- Organiser: Ian Brunton-Smith

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Thank you.

g.durrant@southampton.ac.uk

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Acknowledgements

This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

Conditional Partial CV

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

- Effect of Z when already controlled for all other variables
- The estimator for covariate Z is:

$$\widehat{CV}_c(Z, p_x) = \frac{\sqrt{\frac{1}{n}\sum_{l=1}^{L}\sum_{i\in l}\left(p_i - \hat{\bar{p}}_l\right)^2}}{\hat{\bar{p}}}$$

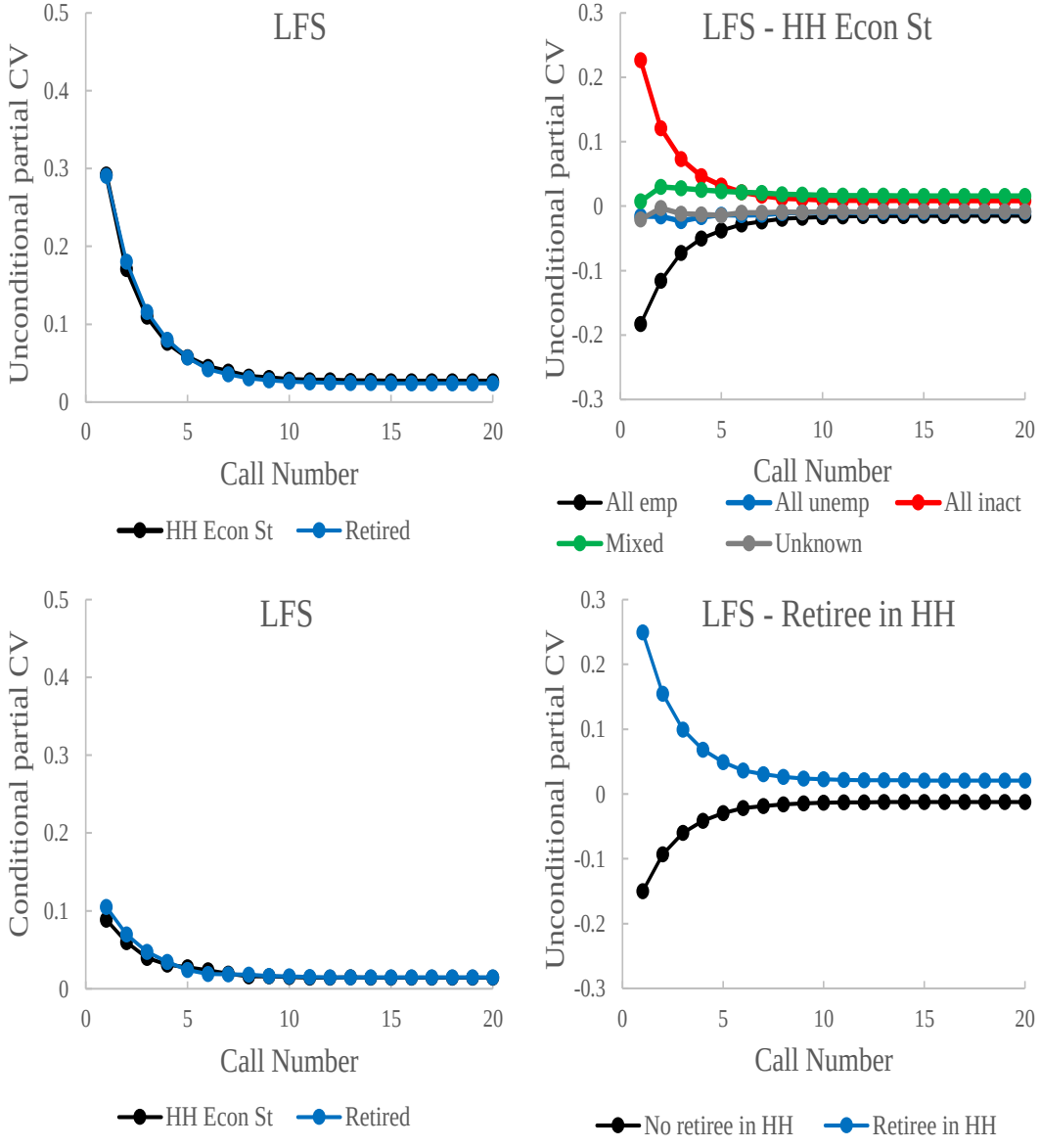
where $\hat{\bar{p}}_l$ is the average estimated response propensity of the *l*th of *L* cells resulting from cross-classification of *x* excluding Z and response propensity modelling given this covariate subset.

- The estimator for category *k* of covariate Z is:

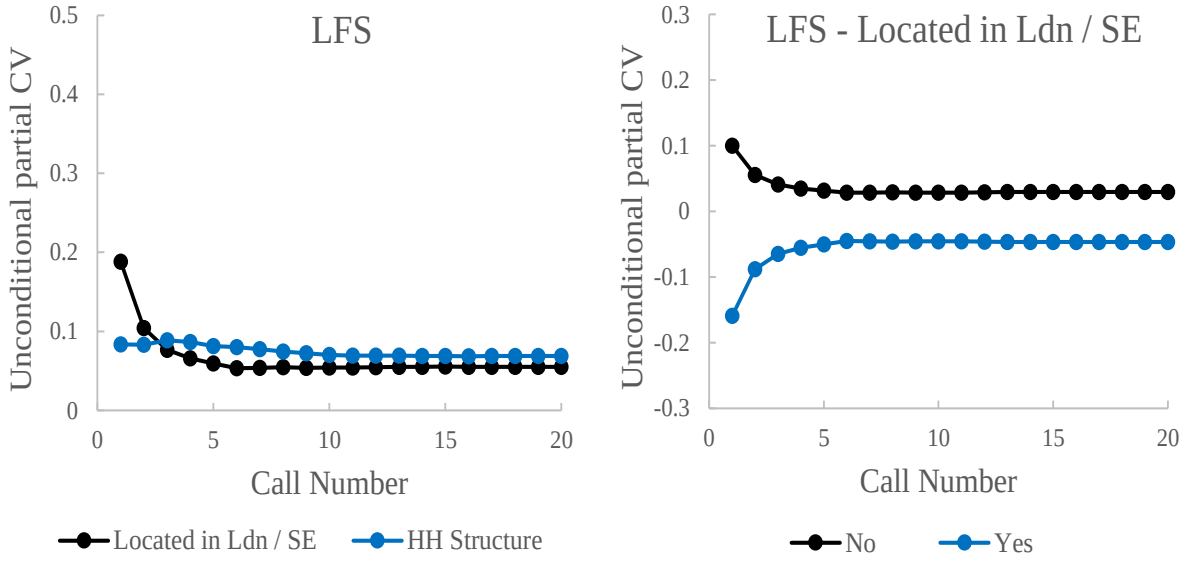
$$\widehat{CV}_c(Z_k, p_x) = \frac{\sqrt{\frac{1}{n}\sum_{l=1}^{L}\sum_{i\in l} h_i\left(p_i - \hat{\bar{p}}_l\right)^2}}{\hat{\bar{p}}}$$

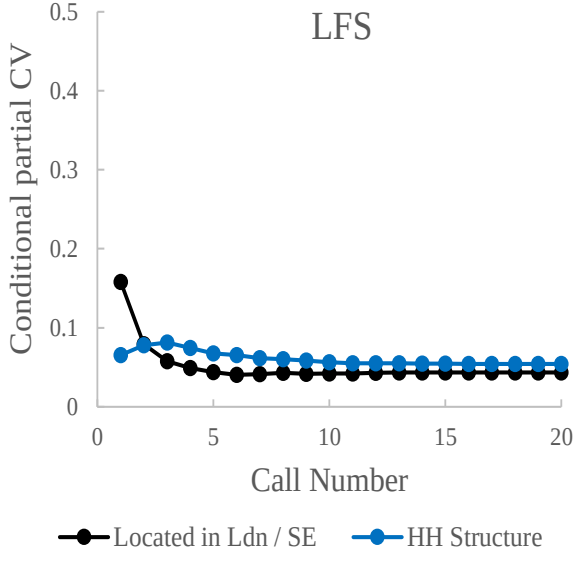
where h_i is an indicator detailing whether participant *i* is in category *k*

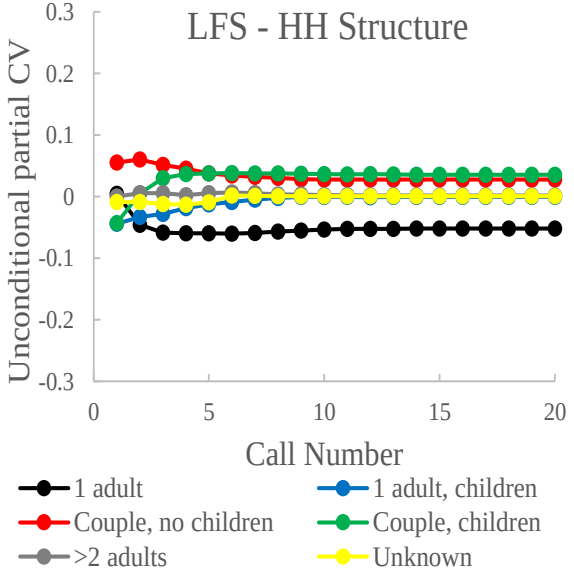
Variables with substantial impacts 1: (correlates of) HH Economic activity



Variables with substantial (uncorrelated) impacts (LFS): Located in London / SE and HH structure

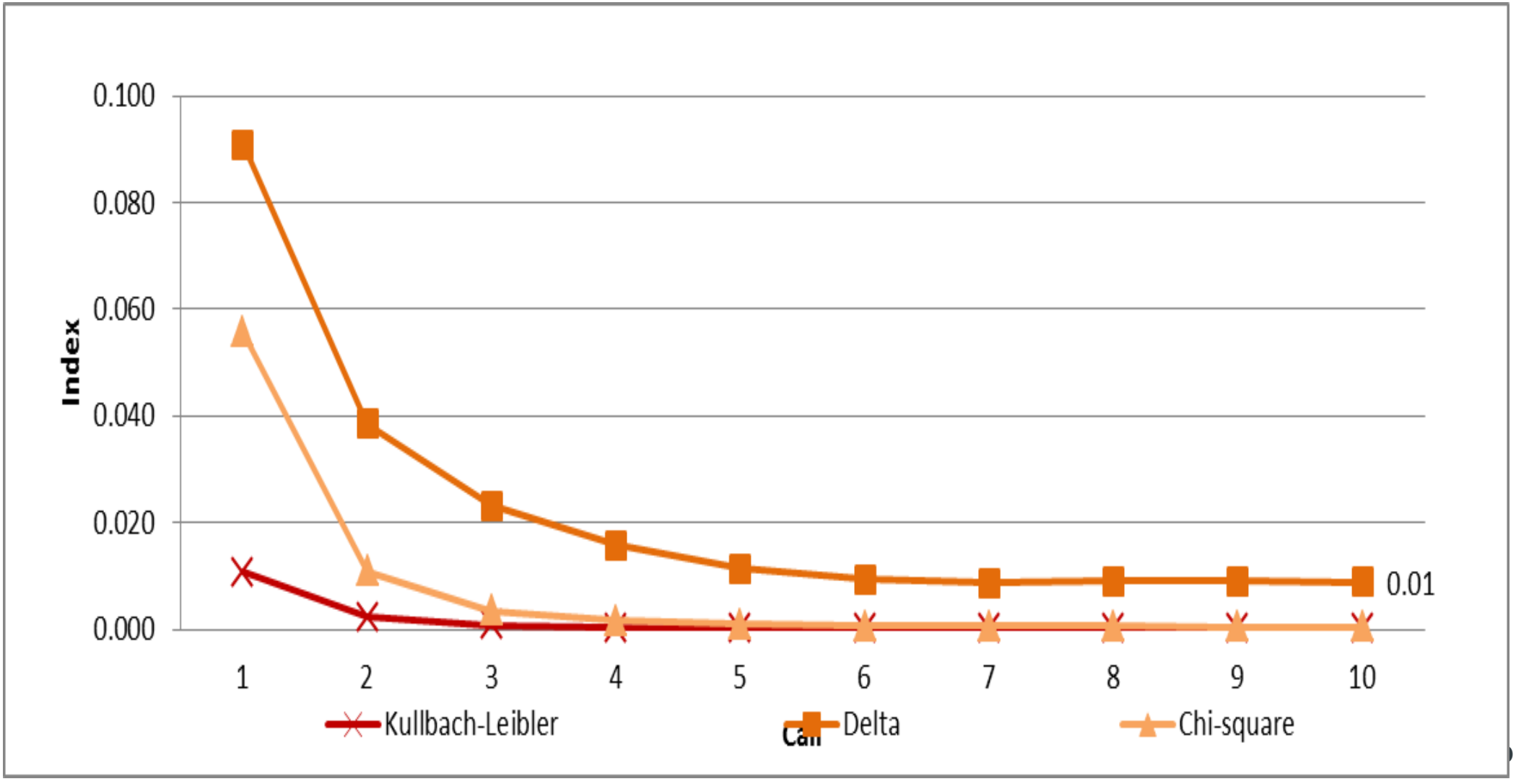






Dissimilarity indices

General Health (categorical variable)



Dissimilarity Index: Categorical

- Kullback-Leibler index:

$$L_z = \sum_{k=1}^{K} \pi_{z,k} \log\left(\frac{\pi_{z,k}}{\hat{\pi}_{z,k}}\right)$$

- Chi-square statistic

$$\frac{(O-E)^2}{E}$$