Utilising representativeness indicators to evaluate non-response and non-linkage biases

Jamie C. Moore, Gabriele B. Durrant & Peter W. Smith

Department of Social Statistics and Demography, University of Southampton, UK.

Emails: j.c.moore@soton.ac.uk

g.durrant@soton.ac.uk

p.w.smith@soton.ac.uk

The problem of survey non-response

- Non-response by some survey sample members ubiquitous.
- More prevalent over the last 30 years.
- Reduces survey sample size.
- Plus, if response not MCAR (respondents and non-respondents do not differ), survey estimates may deviate from sample values.
- These ‘non-response biases’ can have major impacts on inference from datasets.

Reducing non-response bias risks

- Weight survey subjects so they reflect population (sample) using respondent auxiliary attribute covariates and population information?
- Assumes responses MAR (respondents and non-respondents with the same attributes respond similarly) given auxiliary covariates used.
- Tenability? For all survey responses? Responses NMAR? Return to weighting later..

- Alternatively, seek to increase response rates.
- Multiple contact attempts, modes of response (f2f, phone, web).
- But no correlation between response rates and non-response biases.
- Even at high rates respondent – non-respondent differences exist.

For lack of response rate – non-response bias correlation, see Groves & Peytcheva (2008) *Pub. Op. Quart.*, 72, 167–189.

Modern dataset collection methods

- Realisation that high response rates do not preclude biases has led to more nuanced approaches.
- Seek to reduce variation in response between sub-groups defined by attributes correlated with survey responses i.e. make dataset more MCAR.
- Target under-represented sub-groups?
- More contact attempts? How many?
- How to quantify (changes in) impacts on datasets, and bias risks more generally?
- With data collection, all responses of interest, so overall measure required.

e.g. Groves & Heeringa (2006) *JRSSA,* 169, 439-457.

Representativeness indicators

- Quantify sample – respondent similarity using response propensities estimated given a set of auxiliary attribute covariates observed for both respondents and non-respondents.
- One indicator the R-indicator:

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\rho}_i - \hat{\bar{\rho}})^2}$$

- where $\hat{\rho}_i$ is the estimated probability of response for sample unit i and $\hat{\bar{\rho}} = (1/n)\sum_{i=1}^{n}\hat{\rho}_i$.
- Ranges from 0 to 1 (dataset representativeness); Sensitive to auxiliary covariates used and sample size.

See Schouten et al. (2011) *J. Off. Stats.,* 27, 231-253.

The Coefficient of Variation of response propensities (CV)

- Another representativeness indicator is the CV:

$$\widehat{CV}(\rho) = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\rho}_i - \hat{\bar{\rho}})^2}}{\hat{\bar{\rho}}}$$

- Divides response propensity standard deviation by its mean; zero indicates dataset representativeness.
- Quantifies absolute standardised maximal bias of a response mean when response is maximally correlated with the auxiliary covariates.
- HT estimate of survey estimate mean bias is:

$$Bias(\hat{\bar{y}}_r) = \frac{Cov(\hat{\bar{y}}_r, p_R)}{\widehat{\bar{p}_R}}$$

- Standardising with $S(y)$, and replacing covariance with its maxima (the Cauchy Schwartz inequality):

$$\frac{Bias(\hat{\bar{y}}_r)}{S(y)} = \frac{Cov(y, p_R)}{\widehat{\bar{p}_R}\, S(y)} = \frac{Cov(y, p_x)}{\hat{\bar{p}}\, S(y)} \leq \frac{SD\ S(y)}{\hat{\bar{p}}\, S(y)} = \frac{SD}{\hat{\bar{p}}}$$

See Schouten et al. (2011) *J. Off. Stats.,* 27, 231-253.

Partial CVs

- Overall CVs can also be decomposed quantify associations with auxiliary covariates.
- Partial Unconditional CVs:

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

- Test whether MCAR with respect to focal auxiliary covariate.
- Partial conditional CVs:

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

- Test whether MAR with respect to covariate.
- Category level decompositions also possible.

Partial CV functionality

- Comparable, with SEs, so can be used to identify major impacts on datasets.
- Target groups with large conditional CVs: uncorrelated impacts.
- Covariate CVs also identify auxiliary covariates for weighting adjustments.
- Plus, make predictions about auxiliary covariate mean standardised biases.

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

- Covariate CVs predict two category covariate (conditional) bias, multi-category covariate CVs bias maxima.

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

- Category CVs predict (two category) covariate bias minima.
- Predict biases in survey estimate analogues?

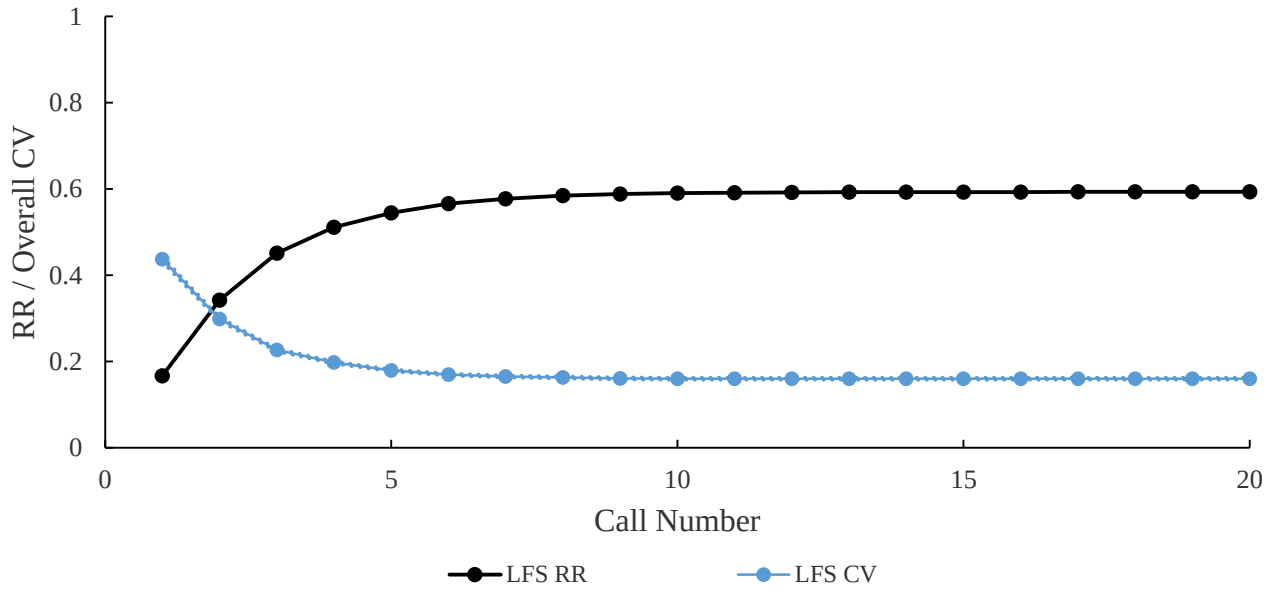
See Moore et al. (submitted) JRSSA

Using CVs to monitor LFS data collection

- The Labour Force Survey (LFS) a major UK government social survey.
- Treated as cross-sectional here.
- Repeated attempts to interview non-respondents.
- Do they improve datasets? What are major impacts on datasets?
- 2011 UK CNRLS links sample households (HHs) to their census information.
- We also link call record paradata on interview attempts (up to 20) to HHs.
- Use this data and CVs to evaluate dataset representativeness over call record.
- 8 auxiliary covariates used to model response propensities: Gender, Age, Quals, Tenure, HH Structure, Located in Ldn / SE, Ethnicity, Activity last week.

See Moore et al. (submitted) JRSSA

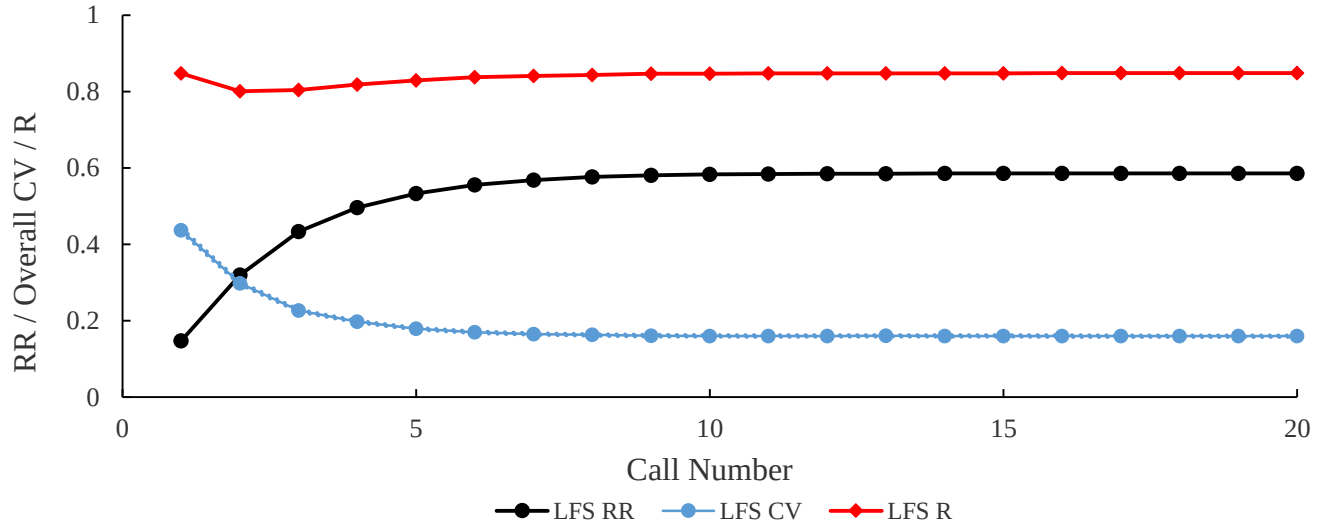
Overall LFS dataset representativeness over calls



- N = 21150.
- Response rate increases at decreasing rate over calls.
- At call 20 (16) = 58.7%.
- Overall CV decreases (representativeness increases) at a decreasing rate over calls.

See Moore et al. (submitted) JRSSA

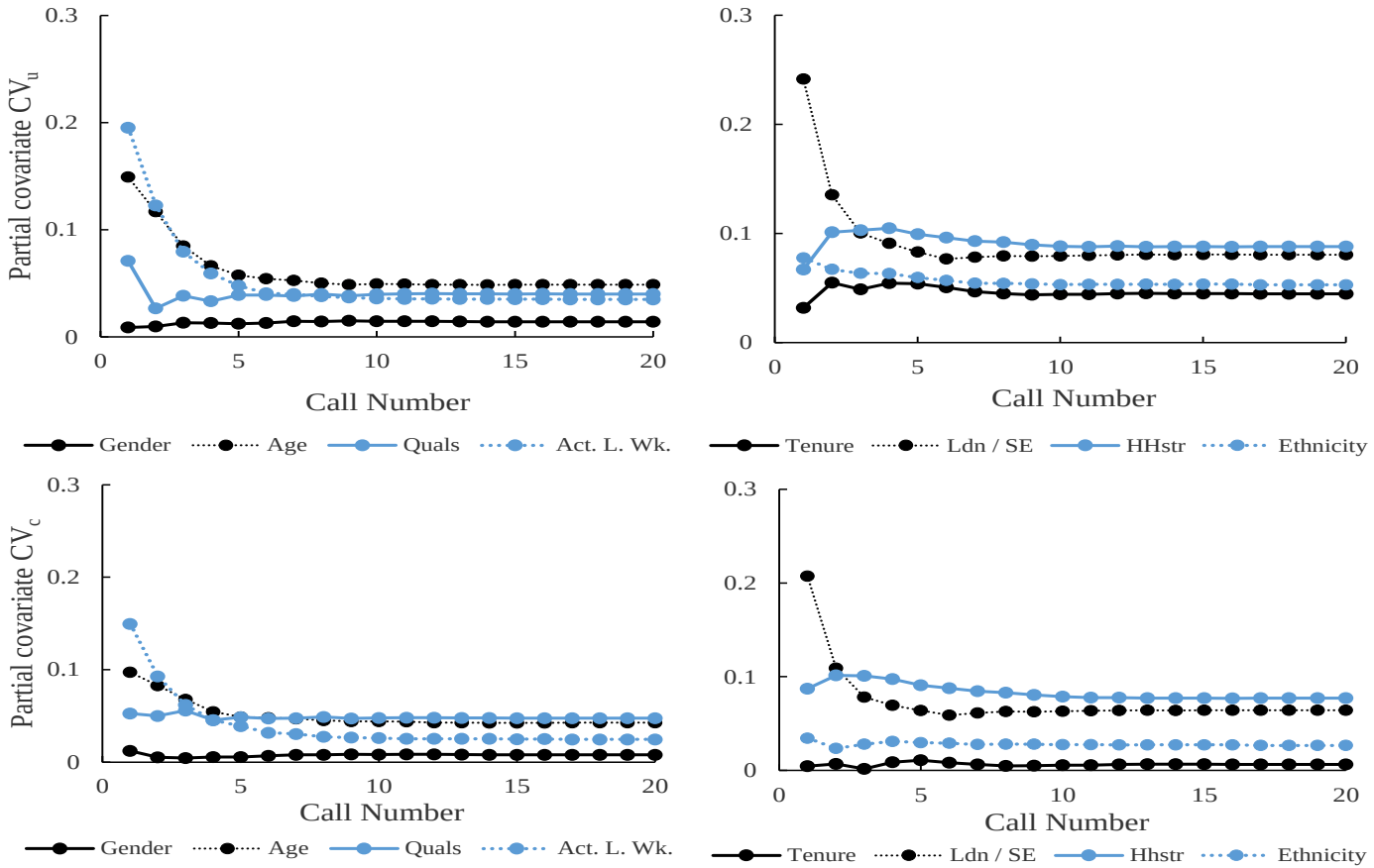
CVs vs. R indicators



- R indicators can falsely suggest representativeness is highest when response rates are low at early calls when there is limited scope for response propensity variation.

See also Moore et al. (2018) *JRSSA* 181, 229-248.

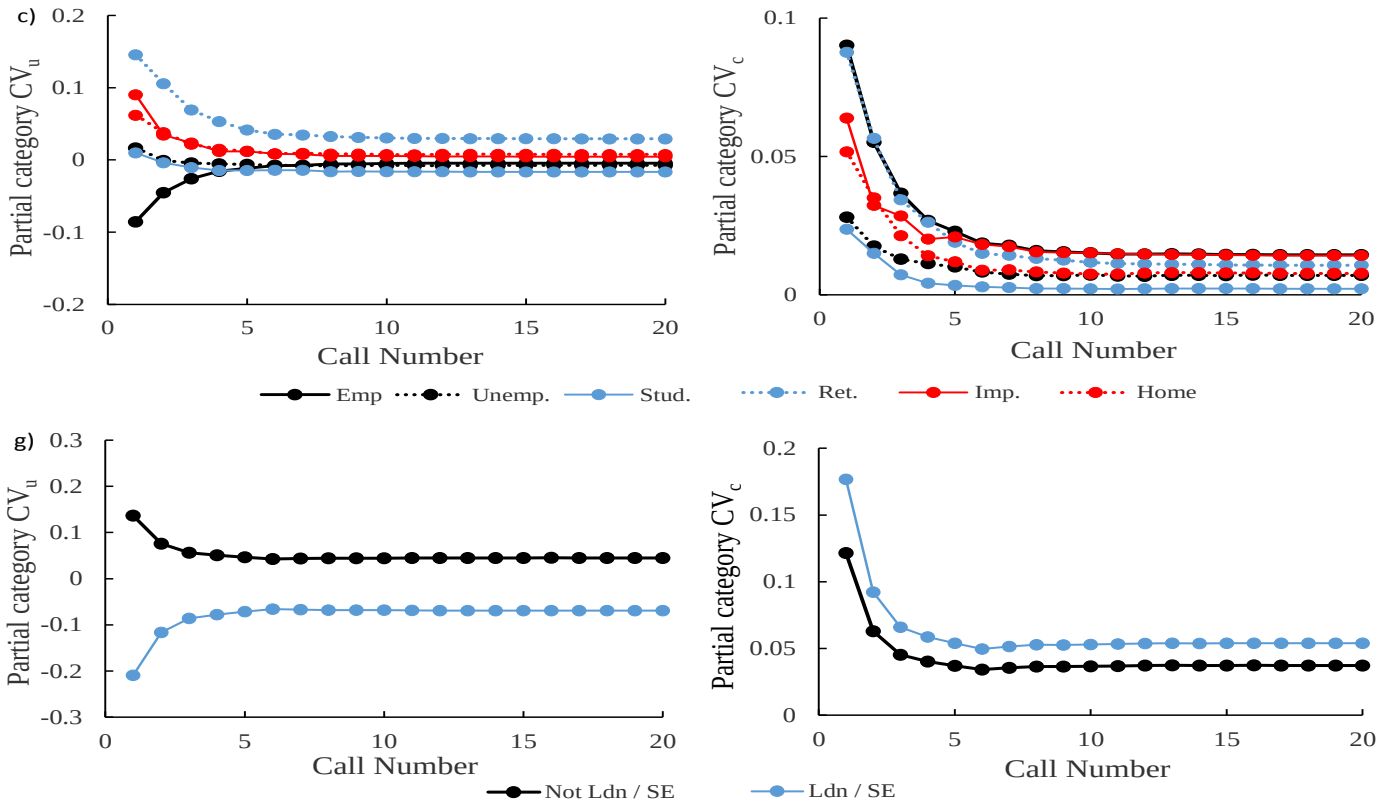
Auxiliary covariate impacts



- All covariates except Gender statistically significant, and should be utilised in computing weights to adjust for biases.

See Moore et al. (submitted) JRSSA

Act. last week & Loc. in Ldn /SE category CVs (major impacts)



- Employed and Ldn / SE ‘Yes’ targets for collection method modifications to improve sub-group response and hence representativeness.

See Moore et al. (submitted) JRSSA

Phase capacity (PC) points

- Given representativeness trajectories over call record, could fewer attempts to interview non-respondents be made, and costs reduced?
- Do Phase Capacity (PC) points beyond which further similar efforts bring minimal increases in dataset quality exist?
- Compute after collection in adaptive strategies, during collection in responsive strategies.
- Numeric (does CV fall with threshold of best / previous value?).
- Or inferential (does CV 95% CI fall within best / previous value interval?).

See Moore et al. (submitted) JRSSA

LFS PC points

- Overall CV:

	During	After
Numeric	4	5
Inferential	5	5

- Act. Last week CV_u:

	During	After
Numeric	5	5
Inferential	4	4

Located in Ldn / SE CV_u:

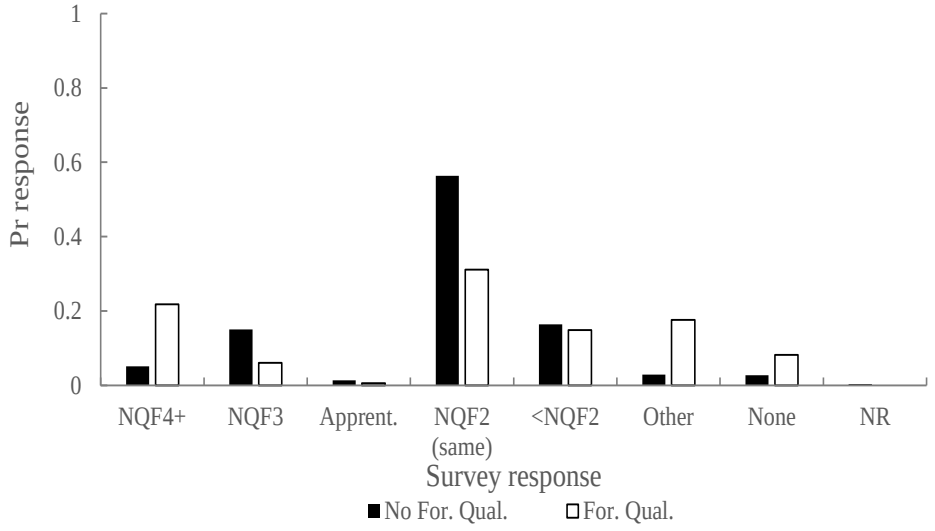
	During	After
Numeric	4	4
Inferential	4	3

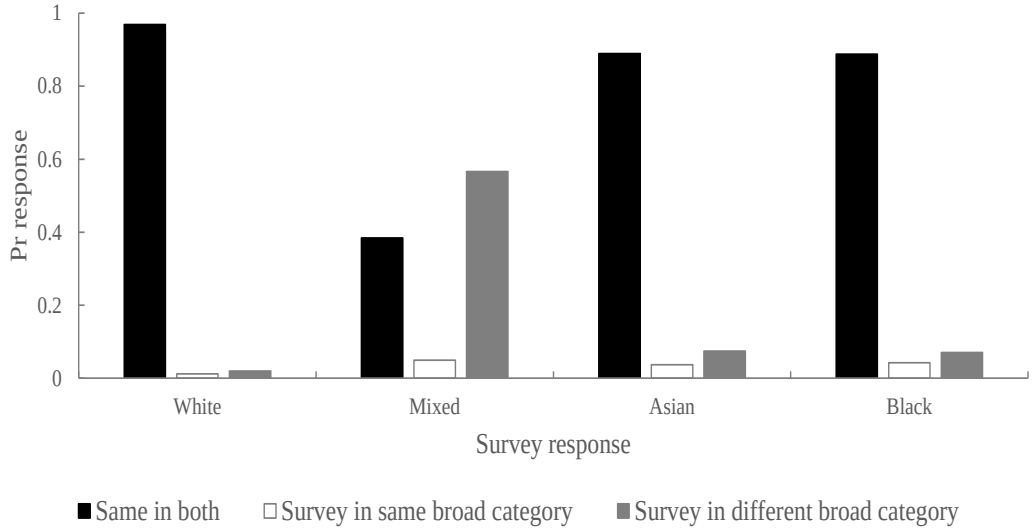
- Overall CVs combine multiple inequalities, so covariate / category points may be earlier or later.
- Inferential points affected by responding proportion and its impact on CV SEs.

See Moore et al. (submitted) JRSSA

Auxiliary covariate biases: a cautionary note

- Partial CVs may predict survey estimate analogue biases for some characteristics.
- But not for others, even those that are fairly static.
- Differential reporting of Qualifications and Ethnicity in census and survey.



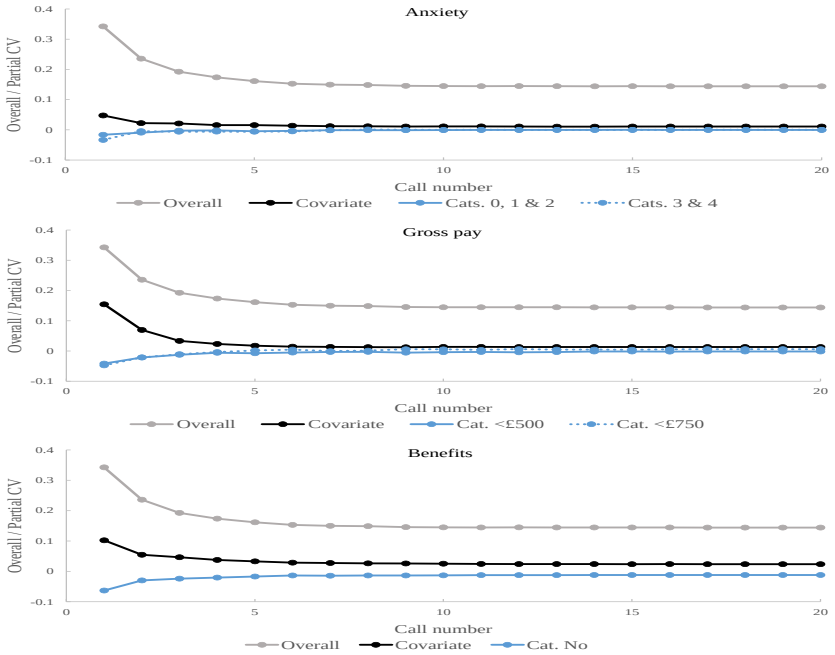


- Due to survey interviewers eliciting more accurate (?) responses.

See Moore et al. (2018) JRSSA 181, 584-585.

Non-auxiliary covariate analogue survey estimates

- Use MI to impute (final dataset) non-respondent values given auxiliary covariates, then compute CVs (Rubin's rules).



CV type	PC point
Overall CV	5
'Anxiety':	
Covariate	2
Cat. 0, 1 & 2	1
Cat. 3 & 4	2
'Gross pay':	
Covariate	4
Cat. <£500	2
Cat. <£750	2
'Benefits':	
Covariate	3
Cat. No	2

- Partial covariate CVs ((maximal) biases) smaller than overall CV, and PC points before overall CV point, but assumes MAR.

Moore et al. (in prep)

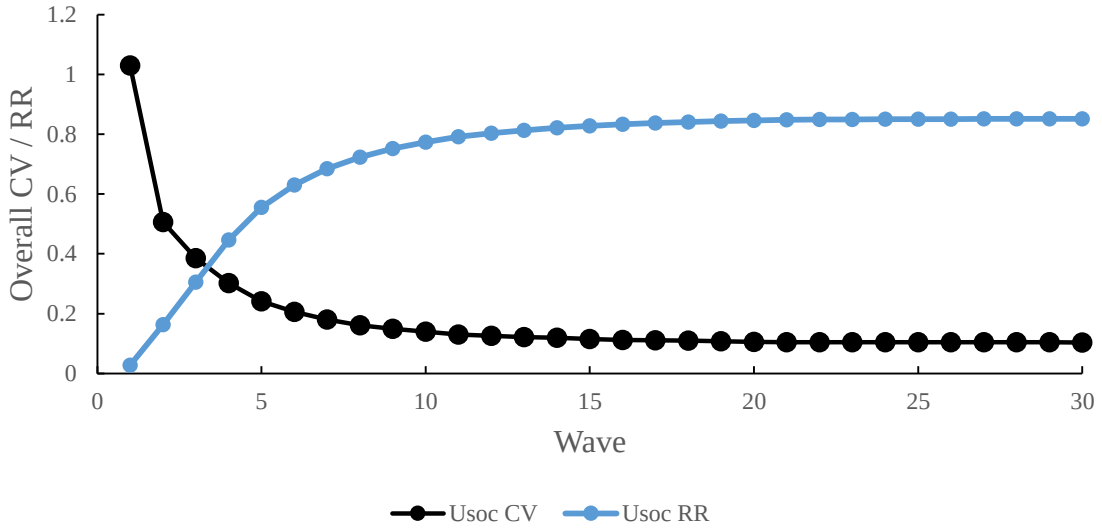
Using representativeness indicators without complete sample level non-respondent information

- Indicators as used with the LFS require auxiliary covariate information for all sample members.
- What to do when this is not available?
- An example is Understanding Society – the UK Household Longitudinal Study (Usoc).
- We've recently evaluated USoc datasets using similar CVs.
- Previous (focal) wave responses as auxiliary covariates.
- Weights to account for subjects also not responding previously and unequal selection probabilities.

Moore et al. (in prep)

Usoc dataset representativeness over collection: wave 7

- Up to 38 HH interview attempts, N = 40938, w6 to w7 call 30+ RR rate = 85%.
- Auxiliary covariates: Gender, Age, Qualifications, Ethnicity, Act. last week, Proxy, Region, Tenure, HH structure.



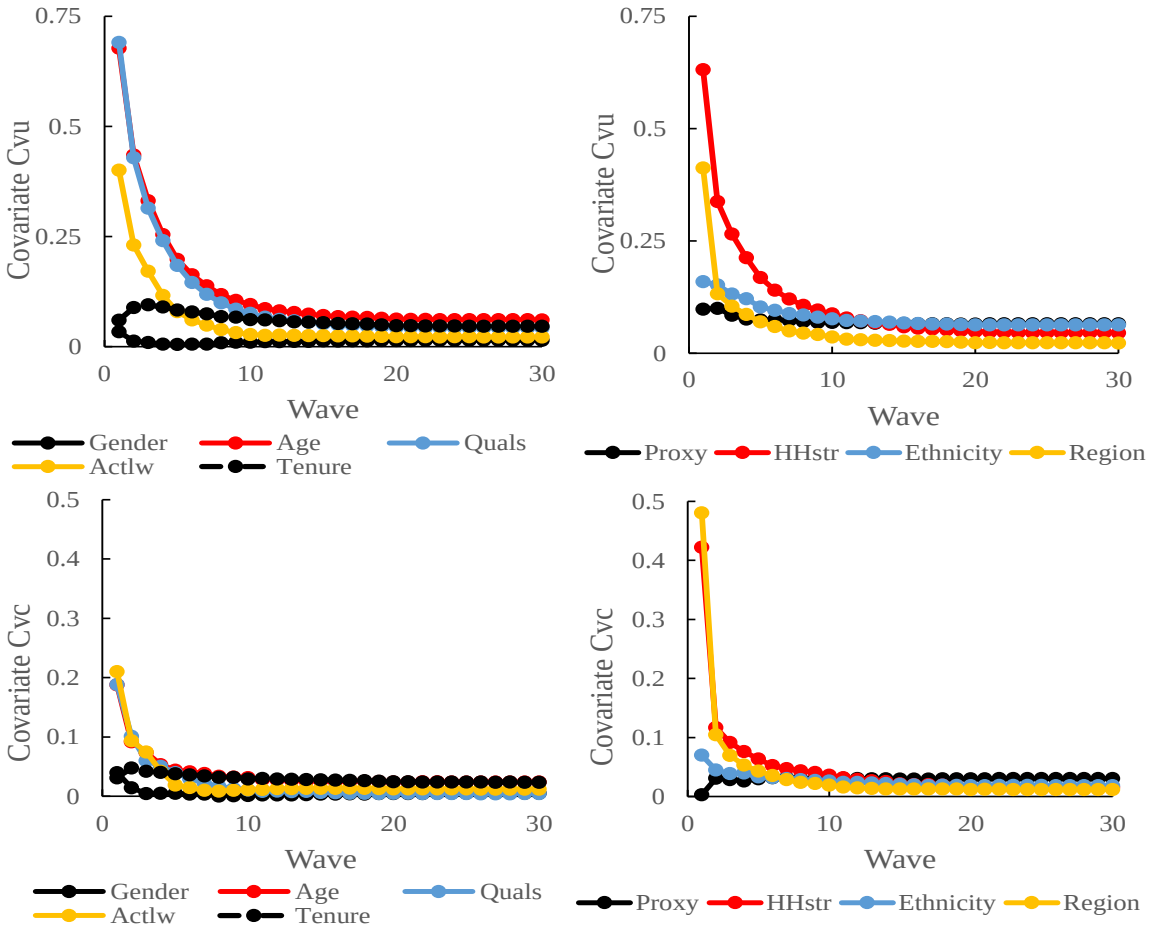
PC points:

	During	After
Numeric	8	13
Inferential	10	16

- Again, representativeness increases at decreasing rate over call record.
- PC points later than for LFS: inferential partly due to larger sample size.

Moore et al. (in prep)

Auxiliary covariate impacts

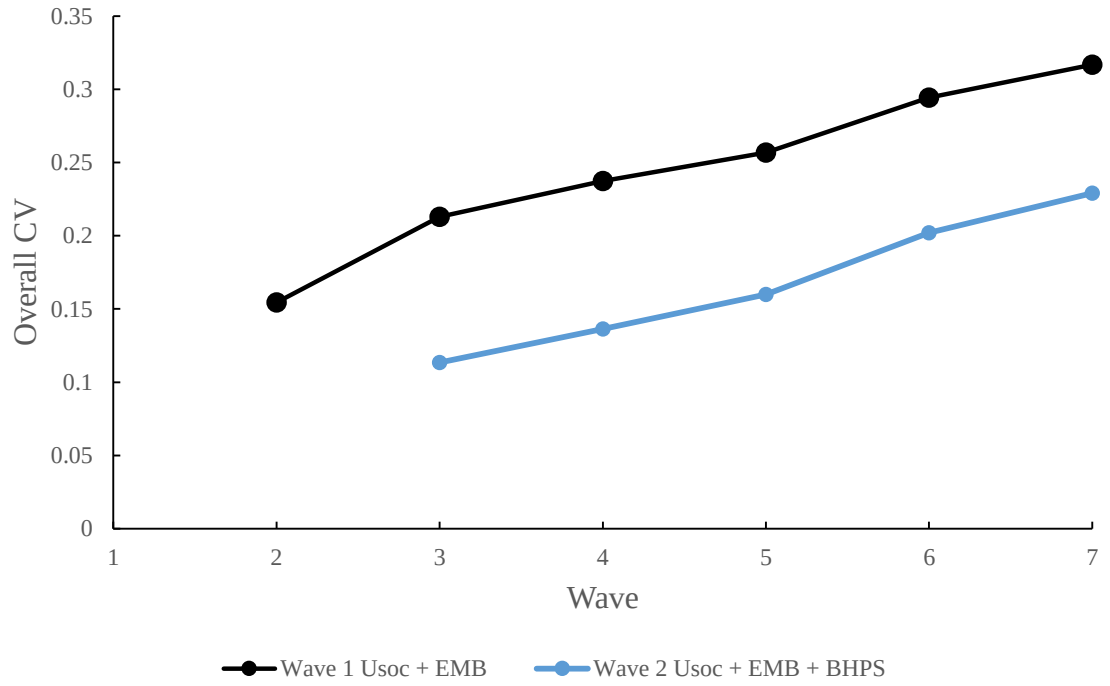


- All except Gender significant, major category level impacts at end include Age ‘16 to 17’, Tenure ‘Not owned’, Proxy ‘Yes’.

Moore et al. (in prep)

USoc sample attrition across waves

- Can also evaluate attrition across survey waves.
- In Usoc (auxiliary covariates as before, plus >6 calls & Health linkage consent):



USoc + EMB w1 to w2 on.

w1 N = 50283, w7 RR = 46%

Usoc + EMB + BHPS w2 to w3 on.

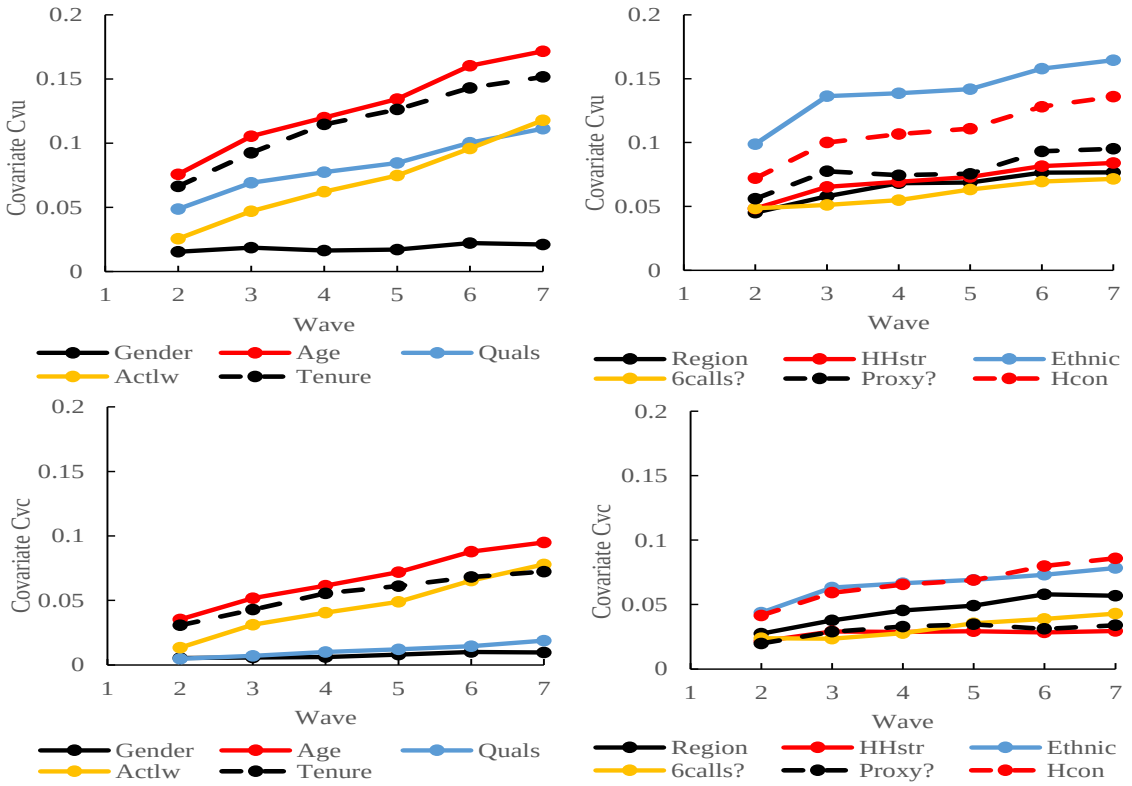
w2 N = 52985, w7 RR = 57%.

- Dataset representativeness decreases over waves.

Moore et al. (in prep)

Auxiliary covariate impacts

- USoc + EMB (USoc + EMB + BHPS similar):



- Gender again NS, impacts increase – a ‘propensity’ to respond or not?
- Note Health linkage consent impact – new; correlated with responses?

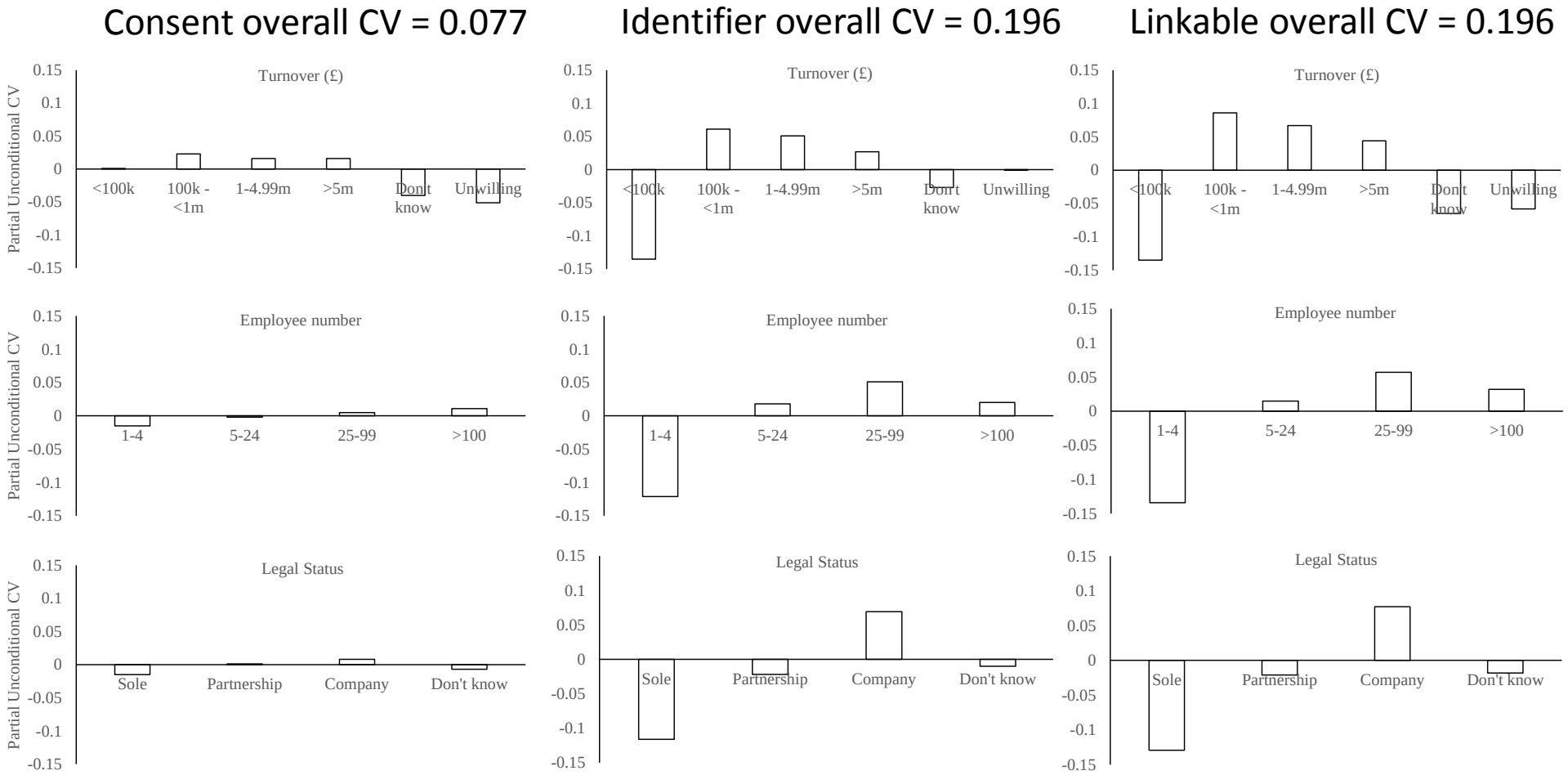
Moore et al. (in prep)

Using representativeness indicators in other scenarios

- Indicators also usable in other scenarios where there are missing data and (some) sample level information on subjects without it.
- e.g. The impacts of non-linkage on linked datasets.
- Analogous to non-response, can cause biases.
- We evaluated linkable UK Small Business Survey datasets.
- Subjects must consent to linkage, and an IDBR identifier be appended.
- Is there non-representativeness? Which auxiliary covariates? The consent or identifier component?
- N = 4850, auxiliary covariates used: Region, Multi-site, Employee number, Legal status, Sector, Exporter, Turnover & Expected performance.

See Moore et al. (2018) *JRSSA* 181, 1211-1230.

Small Business Survey linkable dataset representativeness



Summary: Using representativeness indicators

- Given (some) sample level information on non-respondents, indicators enable efficient evaluation of survey dataset non-response risks.
- Decomposable to identify (and quantify) major impacts on representativeness.
- Can be used to monitor data collection and sample attrition across waves.
- If weights exist, also usable without information on all non-respondents.
- And for other missing data problems e.g. non-linkage in linked datasets.
- However, recall indicator assumptions.
- Predictions may not hold if responses / estimates (weights) are not MAR given the utilised auxiliary covariates.
- In addition, auxiliary covariates from other source / previous wave – timeliness?

Future work

- Timeliness issue potentially addressed by population level representativeness indicators.
- Estimate response propensities given collected respondent information (auxiliary covariates) and external (current?) auxiliary covariate sub-population totals.
- Especially useful for USoc?
- However uncertainty associated with indicator estimates high, and obtaining sufficiently detailed external information may also be problematic, so confirmatory only?
- In addition, only overall indicators available.
- Am currently developing partial versions..

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.
- The same probably holds concerning ISER and the Understanding Society data…