Methods for Selection Bias in the UK Biobank

Valerie Bradley

University of Oxford, Department of Statistics

15 October 2019

The UK Biobank

UK Biobank

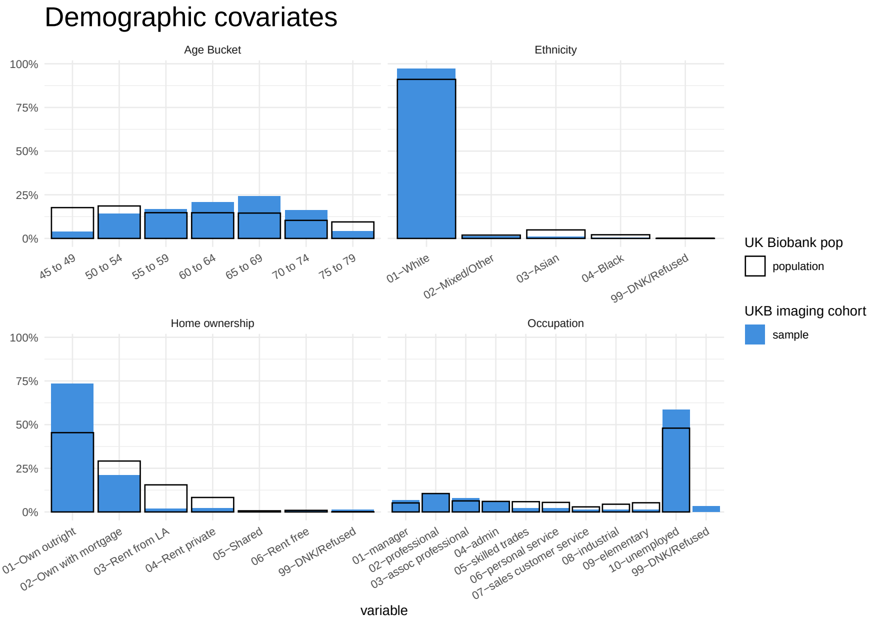
- Largest ever prospective health study
- Includes genetic sequencing, blood tests, physical exams, health history questionnaire
- 500,000 participants aged 40-70 when recruited between 2006 and 2010

UK Biobank imaging cohort

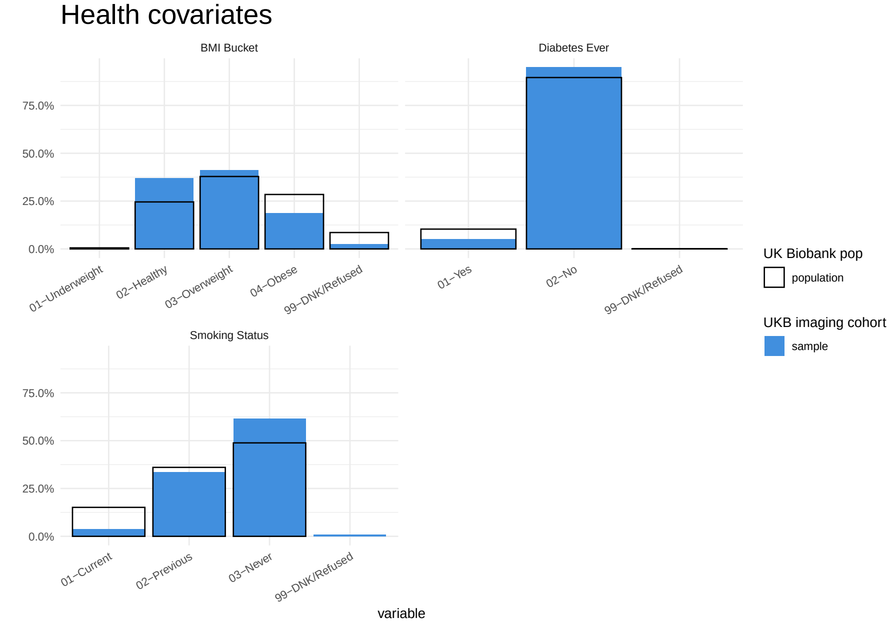
- Subset of UKB participants recruited to undergo additional imaging exams

Goal: Study exposures and outcomes that affect aging populations

Selection bias in the UK Biobank



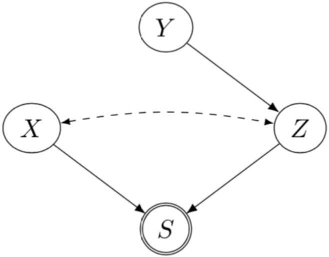
Selection bias in the UK Biobank



Why is this a problem?

For example,

- Y is an individual's hippocampal volume
- X represents a subject's socio-economic status (SES)
- Z represents the subject's level of dementia
- S is selection into the study



By conditioning on selection, if we know that someone is of low SES, they are less likely to show signs of dementia (than if we didn't know their SES)

Recovering from selection bias

If we can re-write our outcome of interest $\mathrm{P}(\mathbf{y}|\mathbf{x})$ as

$$\mathrm{P}(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{z}} \mathrm{P}(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)\mathrm{P}(\mathbf{z} \setminus \mathbf{z}^{\mathbf{T}}|\mathbf{z}^{\mathbf{T}}, S = 1)\mathrm{P}(\mathbf{z}^{\mathbf{T}})$$

then it's possible to recover from selection bias.

- $\mathbf{z}$ is a set of observed *auxiliary variables*
- $\mathbf{z}^T$ is the subset of $\mathbf{z}$ for which we have external, unbiased population data
- Key is that conditioning on $\mathbf{z}$ makes the outcome of interest, Y, *conditionally independent of selection*, S

Problem: can only condition on a limited number of discrete variables before this breaks down

Methods for selection bias seek to solve this in different ways

Methods: post-stratification

Intuition

Methods: summary

Classic weighting methods:

1. *Post-stratification*: adjust to the joint distribution of $\mathbf{z}$
2. *Raking*: iteratively adjust the marginal distributions of elements in $\mathbf{z}$
3. *Calibration*: raking, but with continuous variables as well as discrete variables

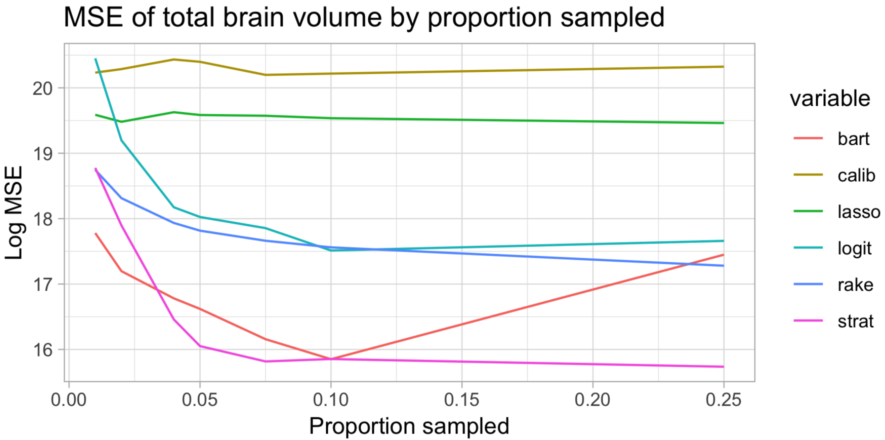
Less-common methods:

4. *LASSO*: use a LASSO to select variables and interactions for raking
5. *Logit*: estimate the probability of selection directly

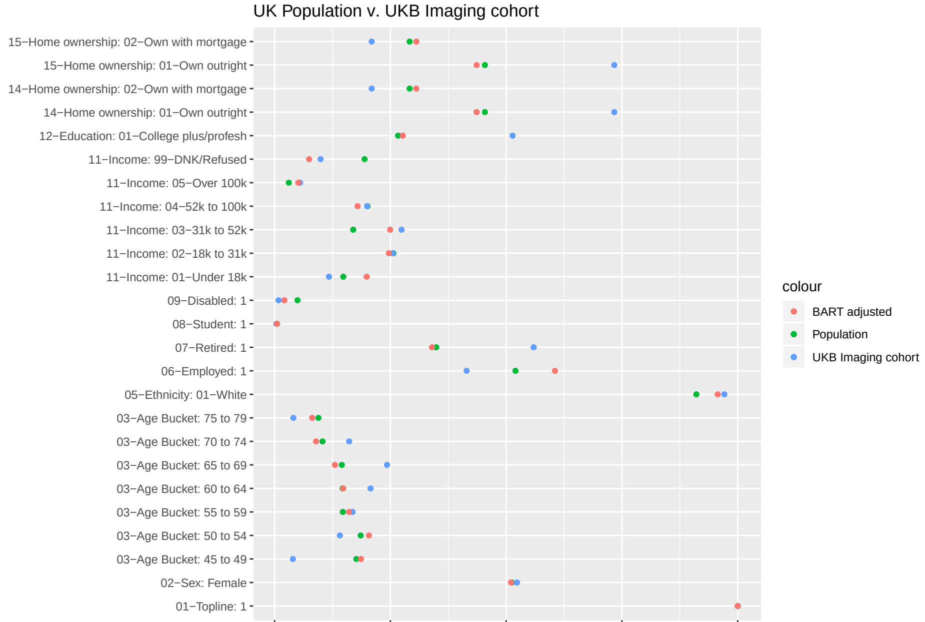
New method:

6. *BART + raking*: use a BART to estimate the probability of selection, then rake such that key marginal distributions match those of the population

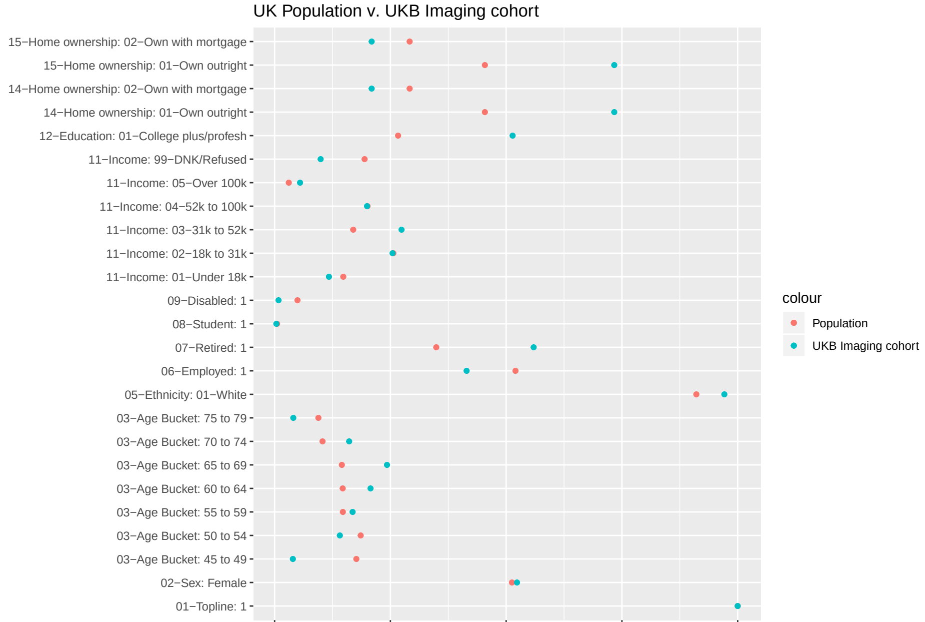
Simulation Results



Application to UK Biobank


UK Population v. UKB Imaging cohort

Selection bias in the UK Biobank


UK Population v. UKB Imaging cohort
15−Home ownership: 02−Own with mortgage
15−Home ownership: 01−Own outright
14−Home ownership: 02−Own with mortgage
14−Home ownership: 01−Own outright
12−Education: 01−College plus/profesh
11−Income: 99−DNK/Refused
11−Income: 05−Over 100k
11−Income: 04−52k to 100k
11−Income: 03−31k to 52k
11−Income: 02−18k to 31k
11−Income: 01−Under 18k
09−Disabled: 1
08−Student: 1
07−Retired: 1
06−Employed: 1
05−Ethnicity: 01−White
03−Age Bucket: 75 to 79
03−Age Bucket: 70 to 74
03−Age Bucket: 65 to 69
03−Age Bucket: 60 to 64
03−Age Bucket: 55 to 59
03−Age Bucket: 50 to 54
03−Age Bucket: 45 to 49
02−Sex: Female
01−Topline: 1
colour
● Population
● UKB Imaging cohort