Government Statistical Service

Using R for variance estimation in social surveys

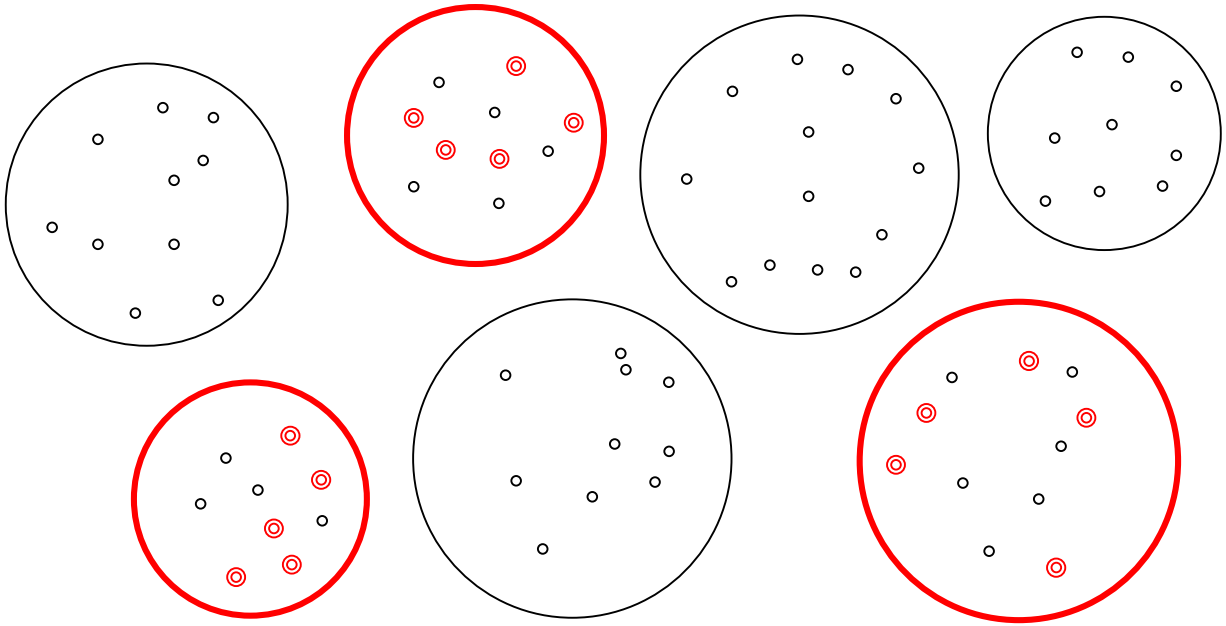
Eleanor Law, ONS

Outline

- Complex sample design and impact on variance of estimates
- The linearised jackknife method
- Implementation of the linearised jackknife in R
- Performance and testing of our package
  - Annual Population Survey
  - Wealth and Assets survey
- Future developments

Complex sample design

- Multistage sampling e.g. Wealth and Assets Survey
  - Primary sampling unit is a postcode sector
  - Systematic sampling after ordering by social demographic indicator/car ownership



Calibration

- Sampling frame for ONS social surveys is usually the postcode address file (PAF)
- No control over the composition of sex/age etc.
- Non-response rates differ between groups
- Weighting can compensate for over/underrepresentation of sex/age/region groups in the sample
- Calibration can reduce standard error of estimates if poststrata correlate with the variable of interest

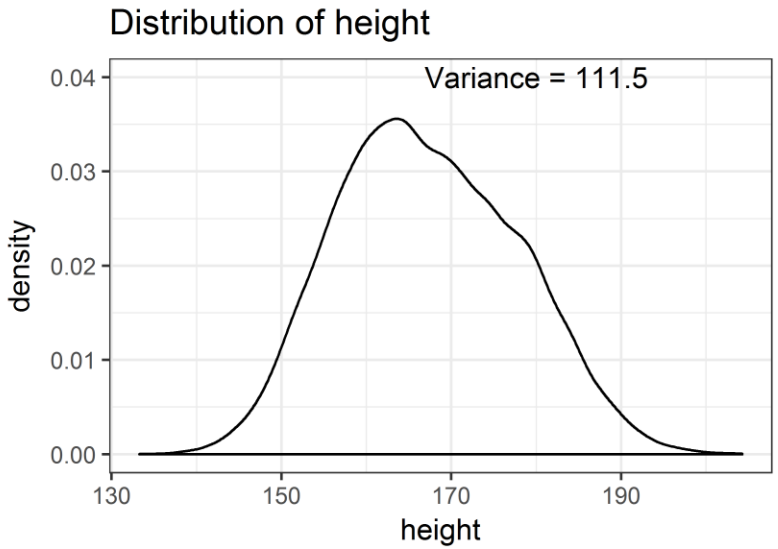
Variance in complex surveys

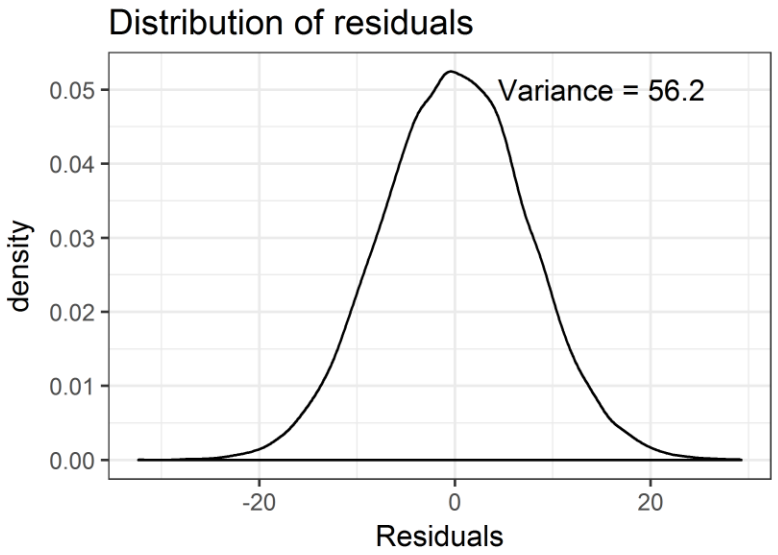
- R `survey` package `svydesign` object allows clusters and strata to be defined
  - Defined in this way, estimates of standard errors will not consider the effect of calibration
- Methods accounting for calibration and variation in the weights have been implemented, e.g. `ReGenesees`

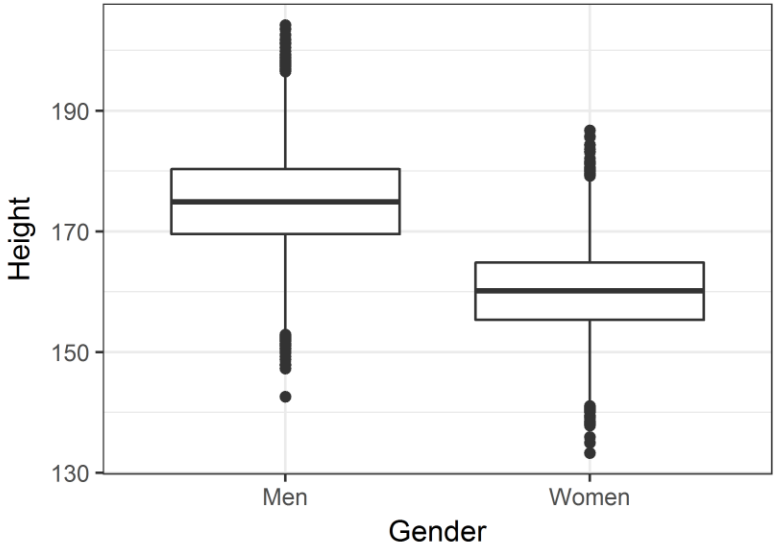
The linearised jackknife

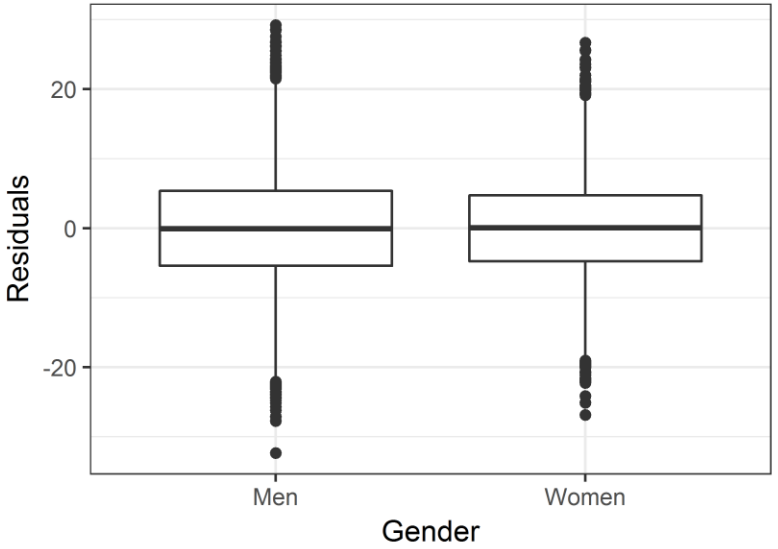
- Fit a linear model for the variable of interest as a function of the poststrata
- This establishes how much of the variance is accounted for by the poststrata as explanatory variables
- Variance that exists in the residuals, after the poststrata have been accounted for, is what we want to know

The linearised jackknife

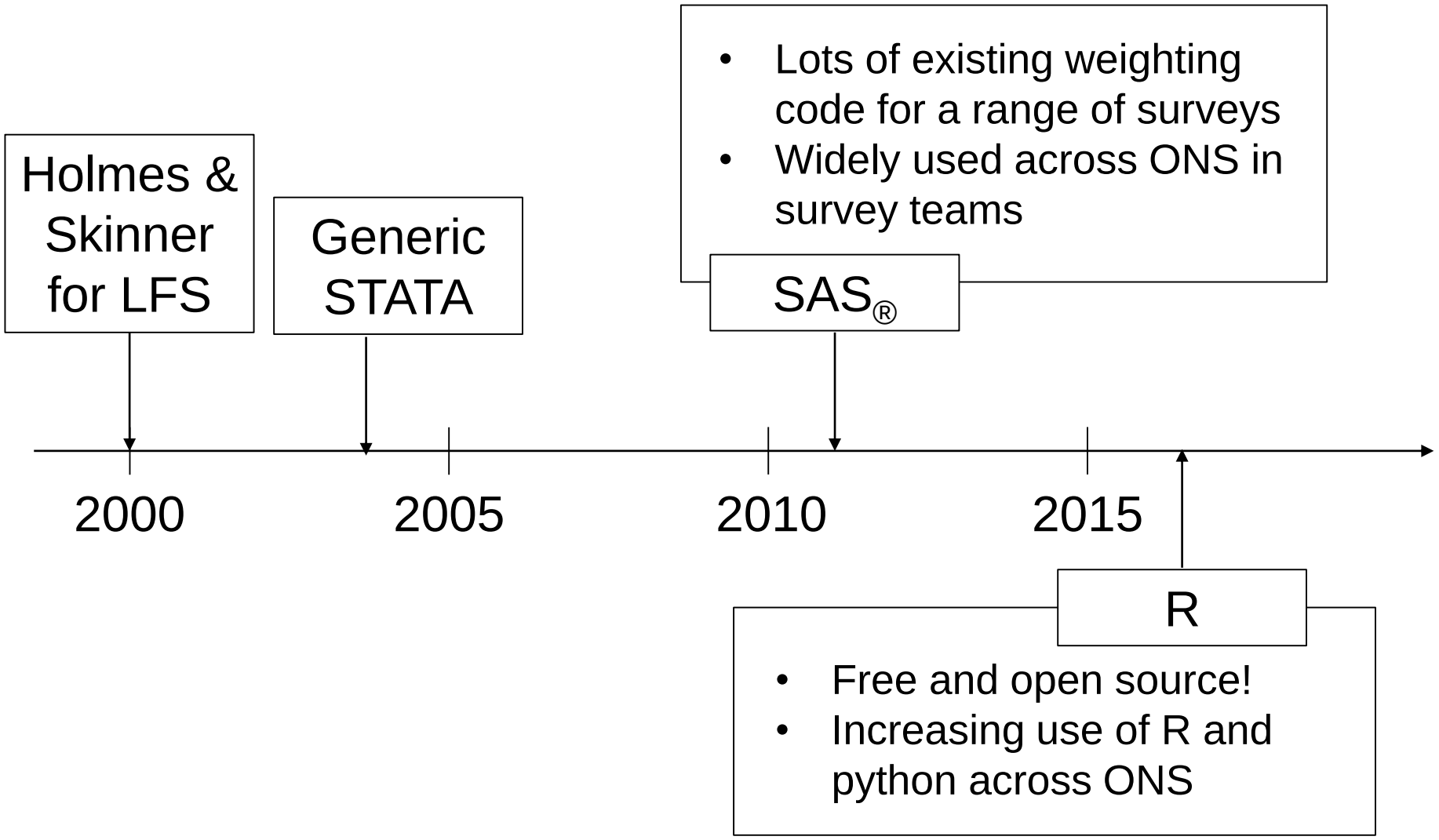








History of implementations in UK ONS



Implementation in R

glinjack {glinjack} R Documentation

Linearised jackknife sampling errors

Description

`glinjack` estimates linearised jackknife sampling errors for total, ratio and mean and returns results formatted in a data frame.

Usage

```
glinjack(df, estimator, y, v = NULL, poststrata = NULL,
  poststrata_h = NULL, deswt, wt, cluster = NULL, strata = NULL,
  domain = NULL, filter = NULL)
```

Arguments

`df`	A data frame which contains all of the variables and observations required for computation of the standard errors
`estimator`	Estimator: either `'total'`, `'ratio'` or `'mean'`
`y`	Main variable for estimation. `y` can be numeric or factor. If `y` is a factor, then a dummy for each category is created and estimation is run for each dummy.
`v`	Denominator (for ratio only)
`poststrata`	Vector of names of individual-level post-stratification variables used for calibration The post-stratification variables are coerced as factors
`poststrata_h`	Vector of names of household-level postratification variables; each variable should be numeric and will be included as a continuous variable
`deswt`	Design weight (used for the computation of the residuals)
`wt`	Post-calibration weight
`cluster`	Clusters used in sampling
`strata`	Strata used in sampling
`domain`	Estimator and standard errors are computed for each group defined by `domain`. The domain argument only allows for one element. If you want to get estimates by ethnicity and gender, you need to create an interaction term using the interaction() function
`filter`	A binary variable defining the sub-population for analysis. This is calculated in exactly the same way as domain, but gives only the results for one of the subpopulations.

Implementation in R

glinjack {glinjack}

Linearised jackknife sampling errors

Description

`glinjack` estimates linearised jackknife sampling errors for total, ratio and mean and retu

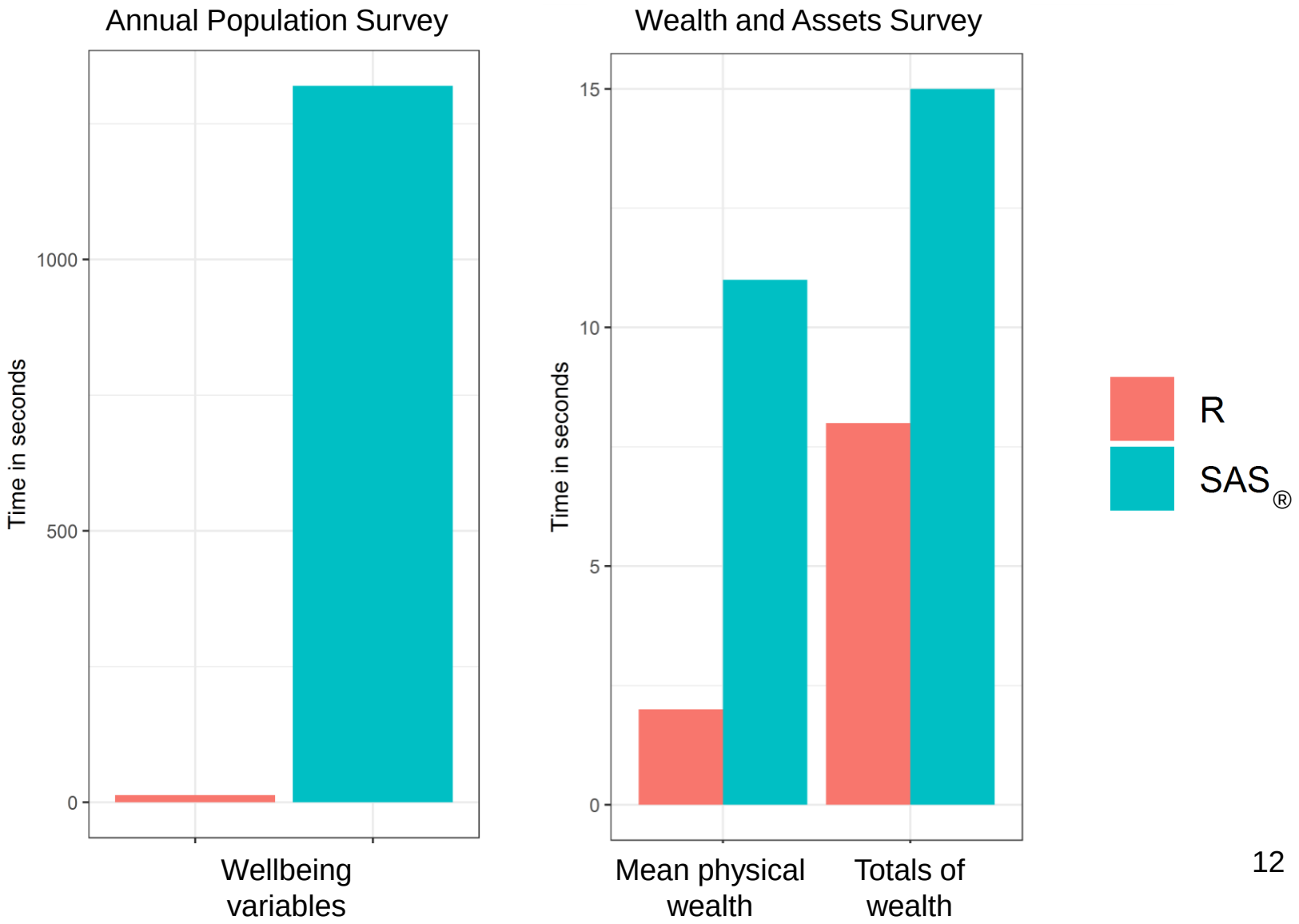
Usage

```
glinjack(df, estimator, y, v = NULL, poststrata = NULL,
  poststrata_h = NULL, deswt, wt, cluster = NULL, strata = NULL,
  domain = NULL, filter = NULL)
```

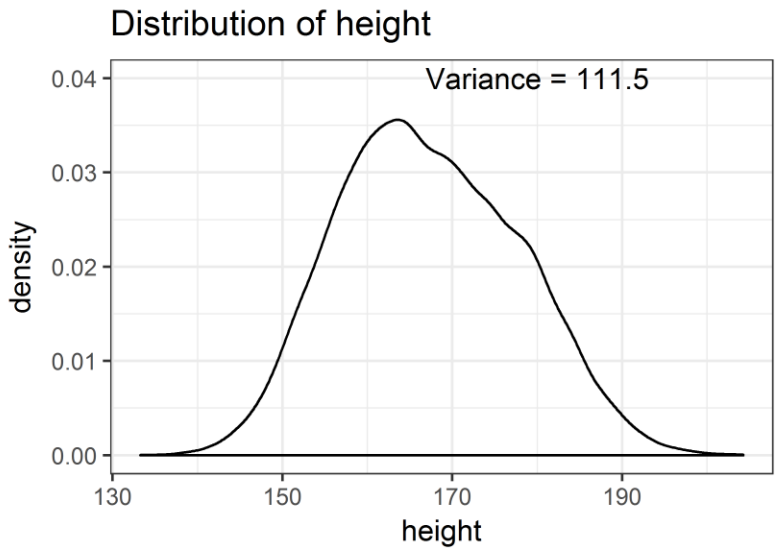
Reproducing standard errors - APS

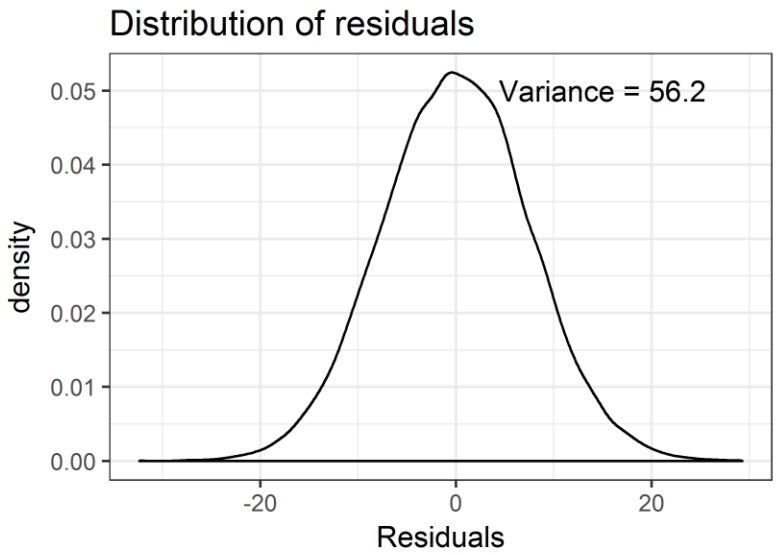
- Personal wellbeing in the UK
- Calibration to age x sex, local authorities
- Four wellbeing variables:
  - Life satisfaction, happiness, sense of worthwhileness and anxiety
- Estimates of average and percentage with very high/high/medium/low levels
- Estimates by age, gender, country and local authority
  - Very time consuming in SAS®

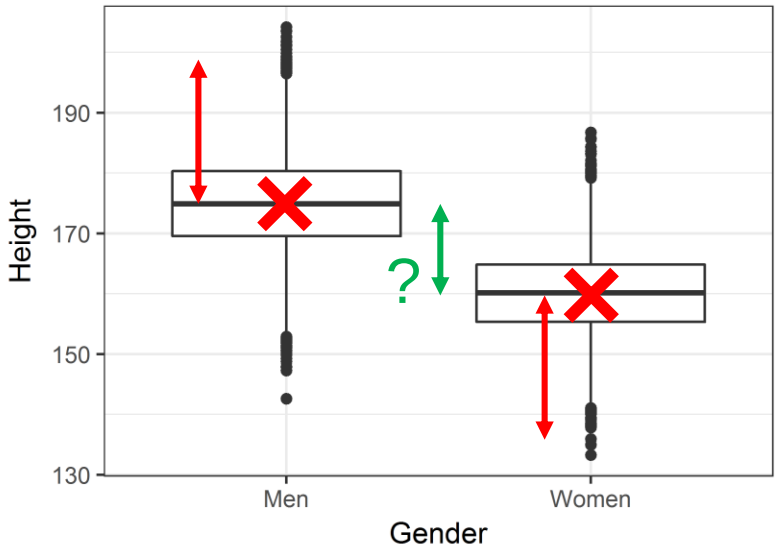
Computational efficiency

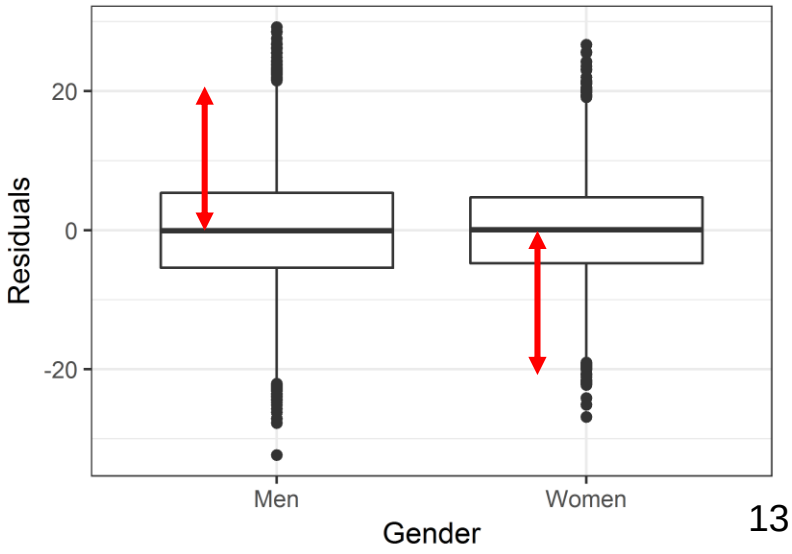


Computational efficiency

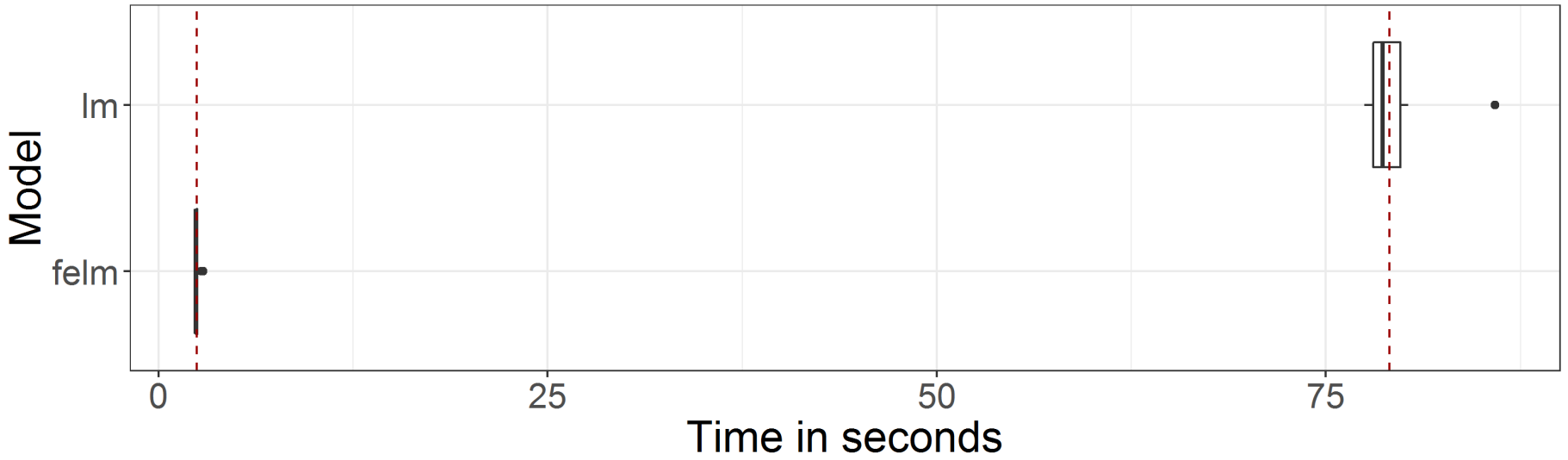








Importance of estimation methods



Variance estimation for households

- Poststrata are usually either
  - One categorical variable *OR*
  - Split into dummy binary variables
- Household level data are aggregated:

	Region 1	Region 2	Sex/age group 1	Sex/age group 2	Sex/age group 3
Person 1	0	1	0	0	1
Person 2	0	1	1	0	0
Person 3	0	1	0	0	1
Household total	0	3	1	0	2

Reproducing standard errors - WAS

- Wave 5 (2014-2016) estimates of
  - Total wealth
  - Financial wealth
  - Property wealth
  - Physical wealth
  - Pension wealth
- Standard Errors originally calculated in SAS®
- Quality assured by recalculation using R

Future Developments

- Automated unit testing using `testthat` package
- Further testing including collaboration within the UK Government Departments to get user feedback
- Aggregation over households within the R function
- Variance of change
  - Very similar method, using input of two datasets with overlapping samples
  - Intended to be a part of the same package, calling the main `glinjack` function

Acknowledgements

- Vahé Nafilyan
- Ria Sanderson
- SD&E(S) team