Variance estimation and weighting in regression

Jean Monet Lecture in Pisa – 2018

Ralf Münnich
Trier University, Faculty IV
Chair of Economic and Social Statistics

Pisa, 08st May 2018

5. Model versus design

6. Weighting in regression – a different view

7. Boostrap reconsidered and use of replicate weights

8. And finally ...

Survey Econometrics

Appropriate methods of inference in the context of econometric analysis of survey data should encompass survey statistical methods such as

- (Survey sampling)
- Non-response handling
- Weighting
- Incorporation of para data
- Variance estimation
- Novel estimation methods

Design-based vs. model-based inference

- The use of survey weights is common practice in the traditional survey statistics context, e.g. in the descriptive analysis of survey data.
- In contrast, the pros and cons of using survey weights when analysing data using statistical models are still discussed in the literature.
- Design-based vs. model-based approach

Pros and cons of the design-based approach

- Advantages
  - No need for assumption of random errors
  - Estimators are constructed so that they are design-unbiased
  - No underestimation of variances of point estimates
  - Robust against model misspecification and heteroscedasticity.
- Disadvantages
  - Possibly inefficient, especially under design ignorability and small sample sizes
  - *"'Survey weighting is a mess.'"*

    Gelman, A. (2007): Struggles with Survey Weighting and Regression Modeling.

    Statistical Science, 22(2), pp. 153-164, p. 153.

Design ignorability:
Probability of inclusion into the sample only depends on known and/or observed information.

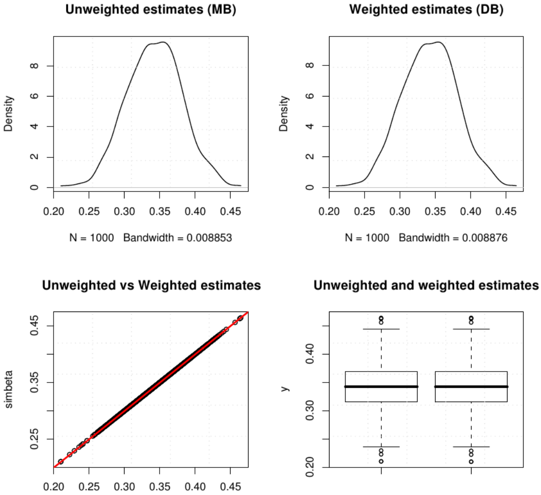
Pros and cons of the model-based approach

- Advantages
  - If model is correctly specified, the unweighted estimators are efficient
- Disadvantages
  - Assumptions about errors are restrictive
  - Model misspecification may lead to bias and even to design-inconsistent estimators
  - ”“*Essentially, all models are wrong, some are useful.*”’

    Box, G.E.P. a. N.R. Draper (1987): Empirical Model-Building and Response Surfaces. p. 424.

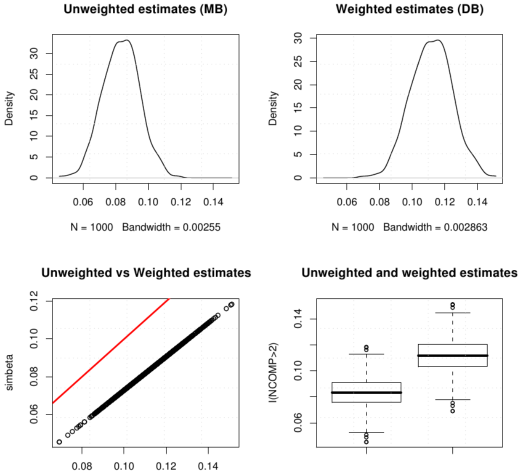
Example 1 - SHIW 2006

Figure 1. Distribution of the parameter: (log) household income (y)



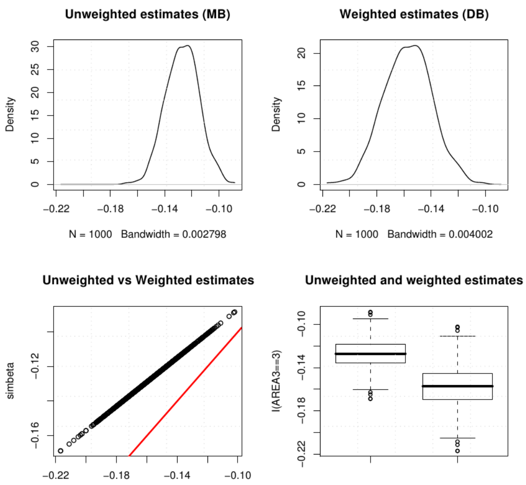
Example 2 - SHIW 2006

Figure 3. Distribution of the parameter: number of household members greater than 2 ($I(NCOMP > 2)$)

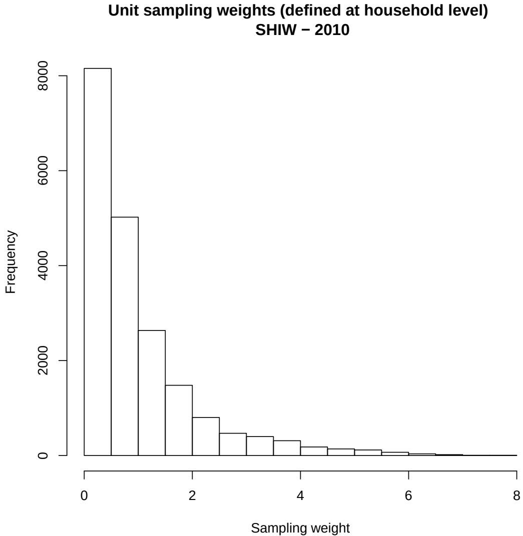


Example 3 - SHIW 2006

Figure 4. Distribution of the parameter: household residing in the South and Islands ($I(AREA3 == 3)$)



Example 4 - SHIW 2002



Reasons for weighting

- Purpose: Estimate population descriptive statistics.
  Correct for:
  - Sampling Design (e.g. oversampling)
  - Non-response
  - Frame errors
  - Fitting to known marginal distributions
    (e.g. calibration or post stratification)
- Purpose: Estimate conditional expectations, maybe causal effects.
  - correct for heteroskedasticity
  - correct for endogenous sampling / informative design
  - identify average partial effects in the presence of unmodeled heterogeneity of effects

Reasons for and the choice of weighting methods are not clearly seperable. Many problems have multiple possible occurences. E.g. sampling design and informative samples

Reasons for weighting

- Purpose: Estimate population descriptive statistics.
  Correct for:
  - Sampling Design (e.g. oversampling)
  - Non-response
  - Frame errors
  - Fitting to known marginal distributions
    (e.g. calibration or post stratification)
- Purpose: Estimate conditional expectations, maybe causal effects.
  - correct for heteroskedasticity
  - correct for endogenous sampling / informative design
  - identify average partial effects in the presence of unmodeled heterogeneity of effects

Reasons for and the choice of weighting methods are not clearly seperable. Many problems have multiple possible occurences. E.g. sampling design and informative samples

Oversampling

Analogous to πps sampling, where a certain variable y is of central importance, in many cases certain subpopulations play a major role in the estimation of the population parameters of interest.

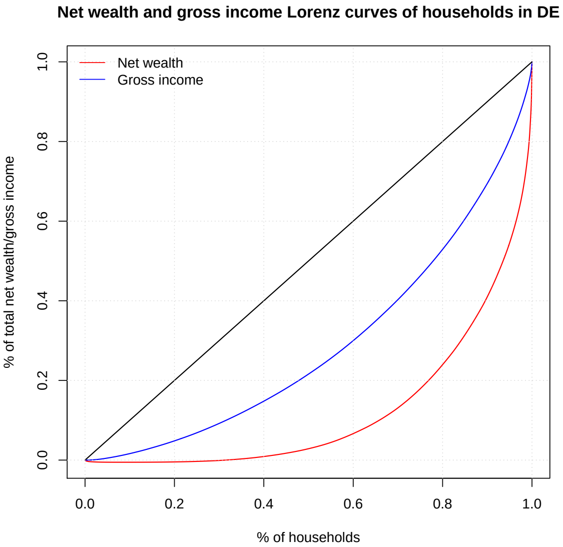
In this case a disproportionately high inclusion probability may be assigned to units within relevant subpopulations. This approach is called oversampling.

Oversampling, if applied correctly, may lead to more precise estimates or lower costs (through reduced sample sizes). The correction of non-response bias might be facilitated as well.

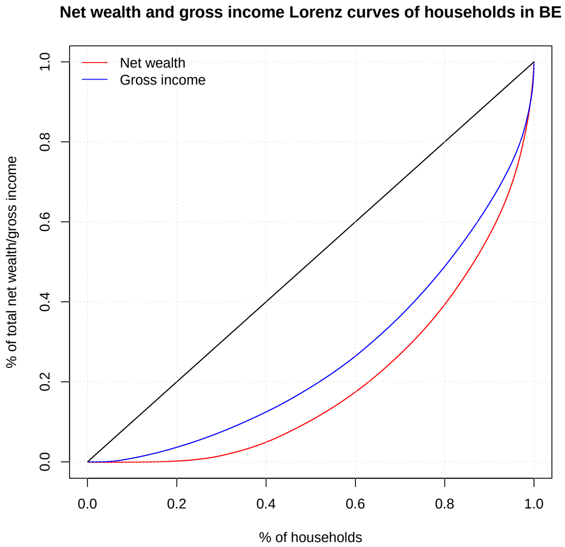
As a prerequisite, auxiliary information is needed to identify the relevant subpopulation. Ideally, as in πps sampling, the auxiliary information and the variables of interest are highly correlated.

Oversampling in wealth surveys

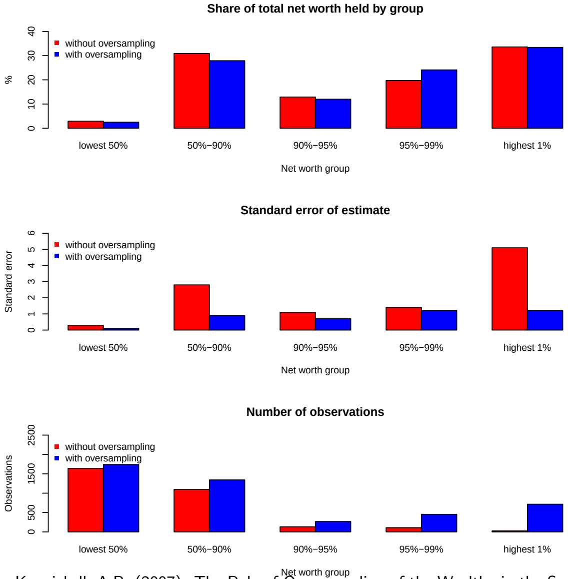
- Wealth distribution highly skewed
  - Gini coefficient of household wealth in Italy (2010): 0.624
  - Gini coefficient of household income in Italy (2010): 0.351
  - Banca d'Italia (2012):
    Household Income and Wealth in 2010. Supplements to the Statistical Bulletin – Sample Surveys 6.
- Very few households hold majority of wealth
- Concentration may well keep on rising
  (see discussions about Thomas Piketty's book)
- Wealth surveys aim at precise estimation of the wealth distribution
- Wealth surveys cover asset holdings as well
- Estimation of financial asset holdings needs good coverage of right tail of wealth distribution
- Non-response typically much higher in wealth surveys
  (sensitivity of subject and complex questions)
- Wealth surveys like the Survey of Consumer Finances (SCF)



Data source: Eurosystem Household Finance and Consumption Survey (2013).



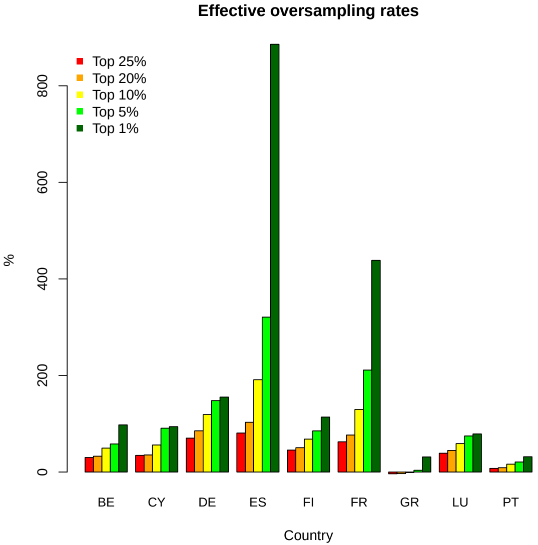
Data source: Eurosystem Household Finance and Consumption Survey (2013).



Data source: Kennickell, A.B. (2007): The Role of Over-sampling of the Wealthy in the Survey of Consumer

Oversampling in the HFCS

- Oversampling of wealthy households is implemented in 9 out of 15 HFCS country surveys
  (BE, CY, DE, ES, FI, FR, GR, LU and PT)
- Different auxiliary information used for routine
  (partially depending on institutional setting)
  - Electricity bill: CY
  - Geographical area: BE, DE, GR, PT
  - Income: FI, LU
  - Wealth: ES, FR
- Oversampling seems to have worked quite well in first wave of HFCS
- Partially very complicated designs
- High variation of final household weights
- How to deal with such designs when analysing survey data?



Data source: Eurosystem Household Finance and Consumption Survey (2013).

Survey Weighted Regression I

In contrast to weighting for heteroscedasticity, weighting for survey issues has the aim to *expand the sample to a finite population*.

That is, the units in the data set are weighted according to their relative importance for the estimation of the population parameters.

Major reasons for the variation of weights in a data set:

- Sampling Design
- Non-response
- calibration to known marginals

Survey Weighted Regression II

The estimation can be performed according to the WLS estimator under heteroscedasticity. E.g.,

- compensating for different inclusion probabilities π_i can be done by using $w_i = \dfrac{1}{\pi_i}$.
- compensating for non-response via the modeling of response propensities which can be used to construct weights.
- accounting for known marginals via post-stratification weights.

Survey Weighted Regression III

As we know from generalized least squares the coefficients b may be estimated via:

$$b^w = (X'WX)^{-1}X'Wy$$

In contrast to the case of generalize least squares for known covariance matrix, the σ^2 has to be estimated, as variances are not known.
An approximately unbiased estimator for σ^2_w is :

$$\widehat{\sigma^2_w} = \sum_{i=1}^{n} \frac{w_i e_i^2}{\sum\limits_{j=1}^{n} w_j - K - 1}$$

Survey Weighted Regression IV

And the variance of the b can be estimated design consistent under a single-stage, unstratified and unclustered design where units are selected with probabilities $\pi_i = 1/w_i$ with replacement.

$$\widehat{\mathrm{V}}(b^w) = (X'WX)^{-1}(\sum_{i=1}^{n} x_i' w_i e_i^2 w_i x_i)(X'WX)^{-1}$$

As can rapidly be seen, under more complex designs, the formulas for the variance estimation get much more complicated.

Under complex survey designs, often the easiest solution is to use resampling techniques.

Li, Jianzhu, and Richard Valliant. “Influence analysis in linear regression with sampling weights.“ 2006. Proceedings of the Section on Survey Research Methods (2006).

Residual bootstrap

Consider the linear regression

$$y_i = \hat{y}_i + \varepsilon_i,$$

then instead of resampling the observations the residuals may be resampled.

1. Fit the linear regression model, and store $\widehat{y}$ and e.
2. Generate $y_i^{*(r)} = \widehat{y}_i + e_i$, where e_j is drawn with replacement from e.
3. Fit the linear regression model as before, just replace y with $y_i^{*(r)}$. Extract the information of interest $h^{*(r)}$ from the regression model (usually the coefficients).
4. Repeat steps 2 and 3 R times and compute afterwards the bootstrap estimates.

Usually the choice of the residual type has no large impact on the results. If in doubt use studentized residuals.

Wild Bootstrap I

The Wild Bootstrap is constructed to be used under heteroskedasticity.

1. Fit the linear regression model, and store $\widehat{y}$ and e.
2. Generate $y_i^{*(r)} = \widehat{y}_i + \nu_i^{*(r)} e_i$, where $\nu_i^{*(r)}$ is a realization of a random variable.
3. Fit the linear regression model as before, just replace y with $y_i^{*(r)}$. Extract the information of interest $h^{*(r)}$ from the regression model (usually the coefficients).
4. Repeat steps 2 and 3 R times and compute afterwards the bootstrap estimates.

Wild Bootstrap II

Popular choices of ν_i are

- standard normal distribution
- Radermacher distribution

$$\nu_i = \begin{cases} -1 & \text{probability of } 0.5 \\ 1 & \text{probability of } 0.5 \end{cases}$$

- Mammen's two-point distribution

$$\nu_i = \begin{cases} -\dfrac{\sqrt{5}-1}{2} & \text{probability of } \dfrac{\sqrt{5}+1}{2\sqrt{5}} \\ \dfrac{\sqrt{5}+1}{2} & \text{probability of } \dfrac{\sqrt{5}-1}{2\sqrt{5}} \end{cases}$$

Wild Bootstrap III

- Radermacher Distribution seems to outperform in many cases Mammen's two-point distribution.
- Radermacher Distribution assumes symmetries, but Mammen's two-point distribution gets the fourth moment wrong.
- In the linear regression, depending on the choice of ν the Wild Bootstrap for variance estimation of the regression parameters converge to robust standard errors.

Parametric bootstrap

In the case of parametric bootstrap the empirical distribution function is replaced by an estimated parametric distribution. Especially in small sample cases this might give better results.

Considering the Design

- When making bootstraps the design of the sample has to be considered.
- That means, if n_h elements were drawn from the h-th stratum, then it should follow $n_h^* = n_h$
- If the observed units were drawn in clusters, e.g. several persons per household, where the household are the sampling units, then the sampling units have to resampled.
- This has also to be considered in multi-stage designs.
- However, in many applications the information on clusters and strata may be retained due to disclosure risks.
- Is there a way to do the *right* bootstrap, without having the full design information?

Replicate Weights

If the data producer can not give the necessary information in order to allow for correct resampling, he can provide a set of replicate weights.

- Use the appropriate resampling techniques, obtaining R resamples $s^{*(r)}$, $r = 1 \ldots R$.
- Recalculate the weights $w^{*(r)}$, in the same manner it was done for the full sample $s^{*(r)}$.
- Use the sample $s^{*(r)}$ with the weights $w^{*(r)}$ to compute the estimate of interest $h^{*(r)}$.
- Take the resampling distribution of $h^{*(r)}$ for the computation of variances.

Design information in scientific use files

- Data producers do not always provide survey design information
  (like e.g. stratum identifiers) with scientific use files of micro data
- Some reasons for withholding such information:
  - Concerns about confidentiality
    (e.g. re-identification risk when first level of stratification is geographic)
  - Majority of data users *may* (unfortunately) not care about survey design
  - Correct variance estimation for complex survey designs *may* be too difficult for *typical* data users

Provision of replicate weights

- Resampling techniques, like the different variants of bootstrapping, are quite flexible and do not require explicit formulas
- Without full design information data users cannot use bootstrapping *per se*
- Solution is a *simulation* of bootstrapping using sets of so-called replicate weights mimicking the repeated drawing of sub samples
  (see above)
- Typically a large number of such replicate weights is provided in scientific use files

Replicate weighting in the HFCS

- The Eurosystem Household Finance and Consumption Network (HFCN) used the rescaling bootstrap of Rao and Wu (1988) and Rao et al. (1992) in the context of the HFCS
- For each unit (household) the HFCN provides a set of 1,000 replicate weights in the micro data set
- Replicate weights were generated using SAS and Stata routines
- Additional calibration step (equal to treatment of final weights)

Variance estimation in the HFCS

The multiple stochastic imputation routine used in dealing with item non-response in core survey items (relating to wealth etc.) has to be taken into account as well.

Exemplary variance estimation for the total of y (θ) using

- n final household weights $(w_i,\ i = 1, \ldots, n)$,
- $n \cdot R$ replicate weights $(\omega_{ir},\ i = 1, \ldots, n,\ r = 1, \ldots, R)$ and
- M imputed data sets $(m = 1, \ldots, M)$

following Rubin (1987) and HFCN (2013):

$$\widehat{\theta}_m = \sum_{i=1}^{n} w_i \cdot y_{im} \qquad \forall \quad m\,(m = 1, \ldots, M)$$

$$\widehat{\theta}_{mr} = \sum_{i=1}^{n} \omega_{ir} \cdot y_{im} \qquad \forall \quad m\,(m = 1, \ldots, M), r\,(r = 1, \ldots, R)$$

Variance estimation in the HFCS (2)

$$\overline{\widehat{\theta}}_m = \frac{1}{R} \cdot \sum_{r=1}^{R} \widehat{\theta}_{mr} \qquad \forall \quad m\,(m = 1, \ldots, M)$$

$$U_m = \frac{1}{R-1} \cdot \sum_{r=1}^{R} (\widehat{\theta}_{mr} - \overline{\widehat{\theta}}_m)^2 \qquad \forall \quad m\,(m = 1, \ldots, M)$$

$$W = \frac{1}{M} \cdot \sum_{m=1}^{M} U_m$$

$$\overline{\widehat{\theta}} = \frac{1}{M} \cdot \sum_{m=1}^{M} \widehat{\theta}_m$$

Variance estimation in the HFCS (3)

$$Q = \frac{1}{M-1} \cdot \sum_{m=1}^{M} \left(\widehat{\theta}_m - \overline{\widehat{\theta}} \right)^2$$

Finally, using the within-imputation variance (W) and the between-imputation variance (Q), the total variance (T) may be calculated as:

$$T = W + \left(1 + M^{-1}\right) \cdot Q$$

A combination of the R packages `survey` and `mitools` (cf. Lumley, 2010 and 2014) can deal with such a setup, but is quite cumbersome to use. Custom-made functions can be faster.

Unavoidable trade-off and user discretion

- Small number of replications ⇒ (potentially) no convergence
- Large number of replications ⇒ long computation time
- Data user faces inherent trade-off
- Data user has discretion over choice of number of replicate weights to be used
- Data user has leeway for manipulative use of replicate weights
- Finding appropriate critical values for significance tests in such a setting is a non-trivial task, allowing additional user discretion.

Discuss:

What should a data user do in case of seemingly randomly missing replicate weights throughout a whole country data set?

Illustrating the problems

The associated problems can be illustrated by estimating a logit model to explain a household's probability to hold at least one mortgage using HFCS data (cf. Bover et al., 2014). The exogenous variables used cover socio-demographic characteristics of the household's core members as well as the number of adult members and the household's total gross income.

The model is estimated for 11 countries in the following variants:

1. Naive: No weights used
2. Weighted: Final weights used
3. Design-based 100: First 100 replicate weights used
4. Significant: 100 replicate weights resulting in lowest estimated variance used
5. Non-significant: 100 replicate weights resulting in highest estimated variance used
6. Design-based 1,000: All 1,000 replicate weights used

Naive

```
             Estimate Std. error      z       Signif.
(Intercept)   -6.2787   1.447906 -4.336           ***
age_16_34     -0.4983   0.230202 -2.165            **
age_45_54     -0.6053   0.224987 -2.691           ***
age_55_64     -0.5830   0.245051 -2.379            **
age_above_65  -1.4861   0.385968 -3.850           ***
age_diff       0.0379   0.024843  1.525
edu_low       -0.0685   0.218294 -0.314
edu_high       0.1008   0.185546  0.543
edu_diff       0.1176   0.191418  0.614
emp_self      -0.0915   0.237936 -0.384
emp_ret       -1.0581   0.323465 -3.271           ***
emp_inact_un  -0.7540   0.392399 -1.921             *
partner_emp    0.3620   0.200540  1.805             *
log_adult      0.1768   0.207858  0.851
log_inc        0.5575   0.133216  4.185           ***
```

Weighted

```
             Estimate Std. error     z       Signif.
(Intercept)   -7.7075   1.530110 -5.037          ***
age_16_34     -0.3223   0.217942 -1.479
age_45_54     -0.6323   0.226648 -2.790          ***
age_55_64     -0.4067   0.264630 -1.537
age_above_65  -1.5237   0.447095 -3.408          ***
age_diff       0.0475   0.027408  1.734            *
edu_low       -0.1230   0.198272 -0.620
edu_high       0.0107   0.195933  0.054
edu_diff       0.1565   0.201387  0.777
emp_self      -0.1688   0.312497 -0.540
emp_ret       -0.9612   0.370500 -2.594          ***
emp_inact_un  -0.8635   0.391238 -2.207           **
partner_emp    0.4560   0.207779  2.195           **
log_adult      0.1027   0.212504  0.483
log_inc        0.6802   0.142425  4.776          ***
```

Design-based 100

```
             Estimate Std. error      z       Signif.
(Intercept)   -7.7075   1.873515 -4.114           ***
age_16_34     -0.3223   0.241787 -1.333
age_45_54     -0.6323   0.254407 -2.485            **
age_55_64     -0.4067   0.285327 -1.425
age_above_65  -1.5237   0.492477 -3.094           ***
age_diff       0.0475   0.027592  1.722             *
edu_low       -0.1230   0.244550 -0.503
edu_high       0.0107   0.212018  0.050
edu_diff       0.1565   0.233819  0.669
emp_self      -0.1688   0.338755 -0.498
emp_ret       -0.9612   0.421281 -2.282            **
emp_inact_un  -0.8635   0.468787 -1.842             *
partner_emp    0.4560   0.238829  1.909             *
log_adult      0.1027   0.245242  0.419
log_inc        0.6802   0.178875  3.803           ***
```

Significant

```
             Estimate Std. error       z       Signif.
(Intercept)   -7.7075   0.802204 -9.608            ***
age_16_34     -0.3223   0.212764 -1.515
age_45_54     -0.6323   0.207368 -3.049            ***
age_55_64     -0.4067   0.239040 -1.701              *
age_above_65  -1.5237   0.280574 -5.431            ***
age_diff       0.0475   0.030142  1.576
edu_low       -0.1230   0.195971 -0.628
edu_high       0.0107   0.182384  0.059
edu_diff       0.1565   0.187934  0.832
emp_self      -0.1688   0.261167 -0.646
emp_ret       -0.9612   0.271325 -3.543            ***
emp_inact_un  -0.8635   0.329870 -2.618            ***
partner_emp    0.4560   0.187901  2.427             **
log_adult      0.1027   0.217782  0.472
log_inc        0.6802   0.076118  8.936            ***
```

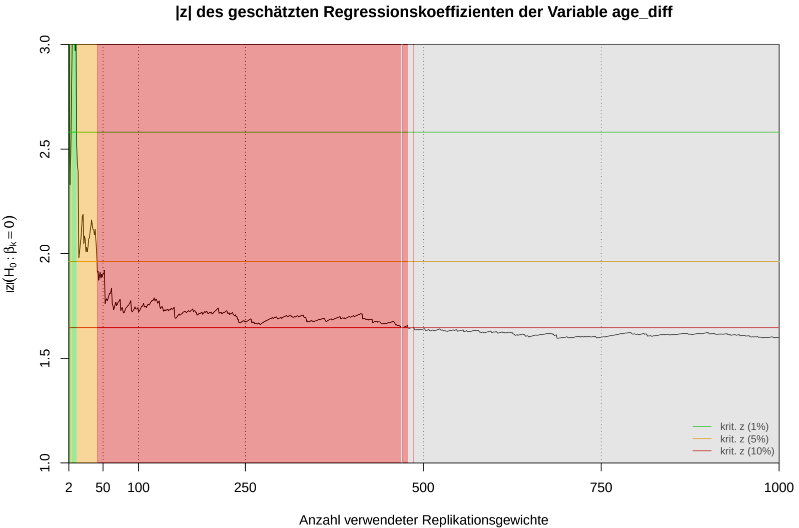
Non-significant

```
             Estimate Std. error      z       Signif.
(Intercept)   -7.7075   3.863880 -1.995            **
age_16_34     -0.3223   0.264625 -1.218
age_45_54     -0.6323   0.254261 -2.487            **
age_55_64     -0.4067   0.343336 -1.185
age_above_65  -1.5237   0.543656 -2.803           ***
age_diff       0.0475   0.028588  1.662             *
edu_low       -0.1230   0.308104 -0.399
edu_high       0.0107   0.257177  0.042
edu_diff       0.1565   0.294790  0.531
emp_self      -0.1688   0.349501 -0.483
emp_ret       -0.9612   0.453414 -2.120            **
emp_inact_un  -0.8635   0.606371 -1.424
partner_emp    0.4560   0.202479  2.252            **
log_adult      0.1027   0.295666  0.347
log_inc        0.6802   0.356598  1.907             *
```

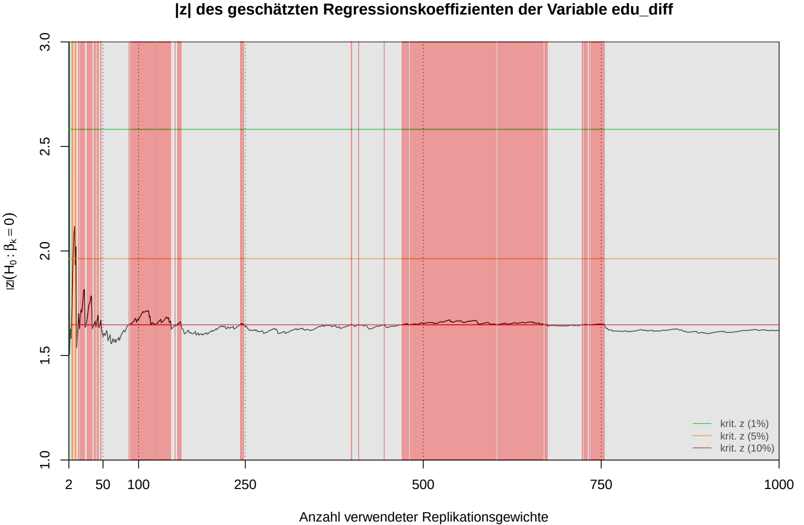
Design-based 1,000

```
             Estimate Std. error       z       Signif.
(Intercept)   -7.7075   1.947833 -3.957           ***
age_16_34     -0.3223   0.262167 -1.229
age_45_54     -0.6323   0.258428 -2.447            **
age_55_64     -0.4067   0.311386 -1.306
age_above_65  -1.5237   0.511940 -2.976           ***
age_diff       0.0475   0.029689  1.600
edu_low       -0.1230   0.238568 -0.516
edu_high       0.0107   0.232261  0.046
edu_diff       0.1565   0.246025  0.636
emp_self      -0.1688   0.322827 -0.523
emp_ret       -0.9612   0.425495 -2.259            **
emp_inact_un  -0.8635   0.503979 -1.713             *
partner_emp    0.4560   0.229529  1.987            **
log_adult      0.1027   0.255610  0.402
log_inc        0.6802   0.180461  3.769           ***
```

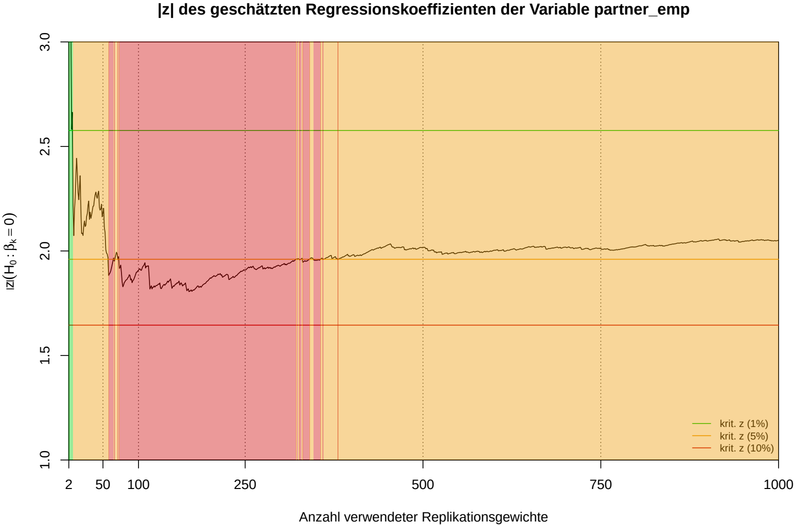
z values of mortgage model parameter for LU



z values of mortgage model parameter for NL



z values of mortgage model parameter for FI



Preliminary conclusions

- Variance estimation using replicate weights not stable and prone to manipulation
- (Typical) ignorance of (replicate) weights seems negligent
- In the case of the HFCS data:
  Ideally use 1,000 replication weights, at least 350 replication weights
- Data producers should consider provision of design information (possibly finding other ways to deal with re-identification risk)

How large is the sample space?

- Design optimization may lead to *very small sample spaces*, i.e. the number of all possible samples is very low
- In practice, this may be rejected as sampling procedure due to missing randomness
- How large should a sample space be? How to we *measure* this space?
- Example: balanced sampling with (too) many constraints
- Possible solution: relaxing the hard constraints
- But this may lead to biased estimators!

Some more issues ...

Does modeling spoil the design world?

- Sample selection biases may occur
- Example: oversampling high incomes (HFCS) in linear regression may lead to biases of even wrong methods
- Can we easily use *non-informative* sampling methods?

How may synthetic data generation be influenced?

- Constructing synthetic household data relies on *appropriate* household and address structures. However, many marginal distributions are known on individual level.
- How can we (re-)sample from these distributions while considering household interactions under marginal constraints?
- Rejective sampling with changing probabilities may be one solution. How should we measure the outcome?

Can we think about BIG DATA subsampling?

Big data analytics always suffers from *modeling on huge and unbalanced datasets*!

- Interesting models do not work on that data unless new algorithms are developed
- The data streams, are in general not balanced, i.e. biased results are very likely
- Sampling from big data may help to reduce the computation burden considerably while reducing the selection bias
- Sampling using satellite data is already known but does not consider *all peculiarities of interest*

But this makes new sampling ideas necessary!