1 – Surveys and datasets

Nikos Tzavidis, University of Southampton, UK - n.tzavidis@soton.ac.uk

Acknowledgements: Timo Schmid, Nicola Salvati, Ray Chambers, Stefano Marchetti & Natalia Rojas

Content of the session

- General remarks on surveys
- Introduction to selected surveys
  - EU-SILC Austria 2006
  - ENIGH Mexico 2013

Aim of sample surveys

Methodology for collecting information via samples on persons, households, or other units.

Survey designer:

- Design and selection of sample design.
  - Cost effectiveness of survey.
  - Frame effectiveness and practicability.
  - Efficiency of estimates (e.g. stratification and optimal allocation).
- Need of valid auxiliary information.

Researcher:

- ... is interested in estimation.
- Here we focus on estimation of population parameters at sub-national level.

Introduction of selected surveys

- EU-SILC Austria 2006
- ENIGH Mexico 2013

EU-SILC survey: Austria

- The *European Union Statistics on Income and Living Conditions* (EU-SILC) is one of the most well-known panel surveys and is conducted in EU member states and other European countries.
- It is mainly used as data basis for the *Laeken indicators*, a set of indicators for measuring risk-of-poverty in Europe. In particular,
  - Inequality: Quintile share ratio or Gini coefficient.
  - Poverty: At-risk-of-poverty-rate (head count ratio) or Poverty Gap.
- The survey serves as a starting point for the Europe 2020 strategy for smart, sustainable and inclusive growth.

Reference: Alfons et al. (2011); Alfons and Templ (2013)

Austrian EU-SILC dataset: Key facts

- The dataset contains 14,827 observations from 6000 households.
- Sample consists of 28 most important variables containing information on
  - Demographics
  - Income
  - Living conditions
- The data are synthetically generated from the original Austrian EU-SILC data from 2006.

Reference: Alfons et al. (2011); Alfons and Templ (2013)

Selected Austrian EU-SILC variables

Variable	Name
Equivalized household income	eqIncome
Region	db040
Household ID	db030
Household size	hsize
Age	age
Gender	rb090
Self-defined current economic status	pl030
Citizenship	pb220a
Employee cash or near cash income	py010n
Cash benefits or losses from self-employment	py050n
Unemployment benefits	py090n
Old-age benefits	py100n
Equivalized household size	eqSS

Reference: Alfons et al. (2011); Alfons and Templ (2013)

Equivalized household income

- Equivalized household income is the total income of a household that is available for spending or saving, divided by the number of household members converted into equivalized adults.
- Household members are equivalised or made equivalent by the following so-called modified OECD (Organisation for Economic Co-operation and Development) equivalence scale:
  - The first household member aged 14 years or more counts as 1 person
  - Each other household member aged 14 years or more counts as 0.5 person
  - Each household member aged 13 years or less counts as 0.3 person

Equivalized household income

The `head()`-command returns the first parts of a vector, matrix, table, data frame or function.

```
  # Loading libraries and the data
library(laeken)
data("eusilc")

# Additional information regarding
head(eusilc)

  db030 hsize db040 age  rb090 pb220a eqSS eqIncome
1     1     3 Tyrol  34 female     AT  1.8 16090.69
2     1     3 Tyrol  39   male  Other  1.8 16090.69
3     1     3 Tyrol   2   male   <NA>  1.8 16090.69
4     2     4 Tyrol  38 female     AT  2.1 27076.24
5     2     4 Tyrol  43   male     AT  2.1 27076.24
6     2     4 Tyrol  11   male   <NA>  2.1 27076.24
```

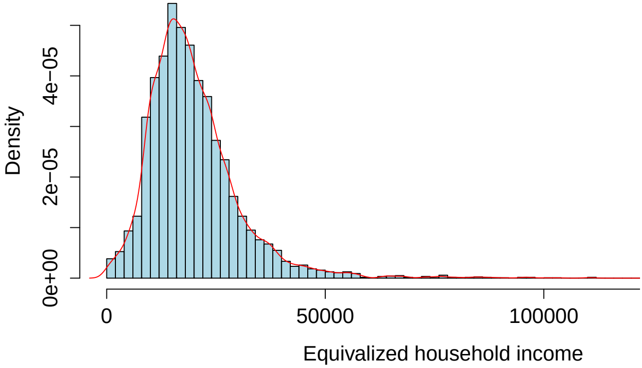
Equivalized household income

The `str()`-command compactly displays the internal structure of an R object.

```
# Additional information regarding
str(eusilc)

'data.frame': 14827 obs. of  8 variables:
 $ db030   : int  1 1 1 2 2 2 2 3 4 4 ...
 $ hsize   : int  3 3 3 4 4 4 4 1 5 5 ...
 $ db040   : Factor w/ 9 levels "Burgenland","Carinthia"
    ,..: 6 6 6 6 6 6 6 8 8 8 ...
 $ age     : int  34 39 2 38 43 11 9 26 47 28 ...
 $ rb090   : Factor w/ 2 levels "male","female": 2 1 1 2
    1 1 1 2 1 1 ...
 $ eqSS    : num  1.8 1.8 1.8 2.1 2.1 2.1 2.1 1 2.8 2.8
    ...
 $ eqIncome: num  16091 16091 16091 27076 27076 ...
```

Equivalized household income - Histogram

```
# Histogram
hist(eusilc_hh$eqIncome,main="Histogram",xlab="
    Equivalized household income",col = "lightblue",
      freq = F,breaks=100)
lines(density(eusilc_hh$eqIncome),col="red")
```



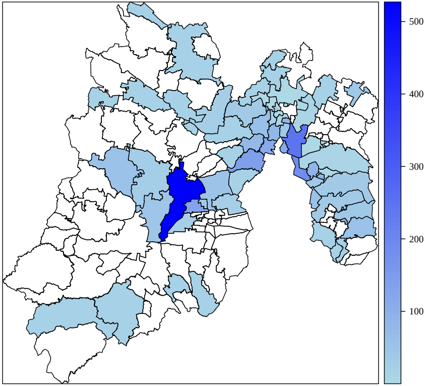
Mexican dataset: Key facts

- The data covers one of the 32 federal entities in Mexico; State of Mexico (EDOMEX).
- Household level survey data with income outcomes and potential covariates (ENIGH survey).
- Survey uses a stratified simple random cluster sample.
- The law requires access to estimates for each municipality.
- 125 municipalities in EDOMEX, 58 are part of the sample, 67 are out of sample.
- The survey includes 2748 households and 115 variables.

Reference: CONEVAL (2010)

Mexico and the State of Mexico

Mexican dataset: Sample Coverage



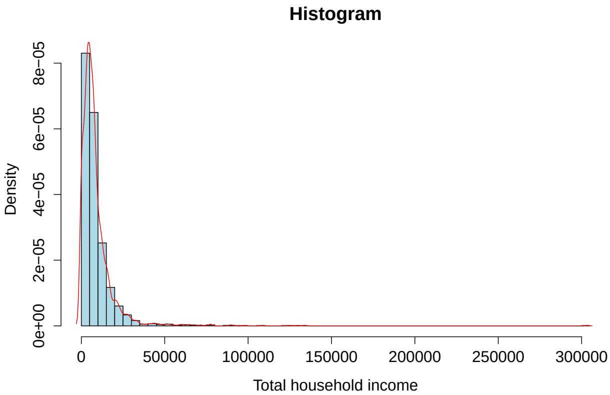
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sample sizes	0	0	0	21.98	20	527
Municipality sizes	931	4657	8494	29790	21170	411700

Selected variables of the Mexican dataset

Variable	Name
Total household income	inglab
Household income from work	inglabpc
Region	clusterid
Educational level of head of household	jnived
Total assets of goods in the household	bienes
Social class of the household	clase_hog
Percentage of employed people in the household	pcocup
Lack of access to health services	ic_asalud
Lack of access to food	ic_ali
Lack of access to education	ic_rezedu
Lack of access to basic housing space	ic_cv

Total household income - Histogram

```
# Histogram
  hist(survey_data$inglab,main="Histogram",xlab="Total
     household income",col = "lightblue",
    freq = F,breaks=100)
  lines(density(survey_data$inglab),col="red")
```



2 – Direct estimation

Acknowledgements: Thanks to Ralf Münnich (University of Trier) and Matthias Templ (TU Vienna) for providing useful materials.

Content of the session

- Direct estimation
- Variance estimation

Example: The sample mean (under simple random sampling)

$$\hat{\mu} = \overline{Y} = \frac{1}{n} \sum_{j=1}^{n} Y_j$$

as an estimator for the population mean μ_Y.

- $\hat{\mu}$ is the best linear unbiased estimator (BLUE) for μ.
- $\hat{\mu} \sim N(\mu, \frac{\sigma_Y^2}{n})$.

Example: EU-SILC Austria:

```
> library(laeken)
> data("eusilc")
> mean(eusilc$eqIncome)
[1] 19906.87
```

Is simple random sampling realistic?

Reference: Alfons and Templ (2013)

The need for sampling weights

Sampling weights are needed to correct for imperfections in the sample that might lead to bias and other departures between the sample and the reference population. In particular,

- To compensate for unequal probabilities of selection.
- To compensate for (unit) non-response.
- To adjust the weighted sample distribution for key variables of interest (for example, age, race, and sex) to make it conform to a known population distribution.

Horvitz-Thompson / Hajek estimator for means and totals

For estimating a total τ_Y of a variable of interest Y we take

$$\hat{\tau}_{HT} = \sum_{j \in s} \frac{y_j}{\pi_j} = \sum_{j \in s} y_j w_j,$$

where $w_j = 1/\pi_j$ denote the design weights (as reciprocal of the first order inclusion probabilities).
In order to estimate means, one can use the following estimator

$$\hat{\mu}_{HT} = \frac{\sum_{j \in s} w_j y_j}{\sum_{j \in s} w_j}$$

Reference: Horvitz and Thompson (1952)

Using R-package laeken

```
> # Loading libraries and the data
> library(laeken)
> data("eusilc")

> # Weighted mean vs. unweighted mean
> mean(eusilc$eqIncome)
[1] 19906.87
> weightedMean(eusilc$eqIncome,weights=NULL)
[1] 19906.87
> weightedMean(eusilc$eqIncome,weights=eusilc$rb050)
[1] 19890.81
```

Reference: Alfons and Templ (2013)

Poverty indicators: Head count ratio

- The Head Count ratio (HCR) also known as the at-risk-of-poverty-rate (ARPR).
- The HCR depends on a poverty threshold (at-risk-of-poverty threshold, ARPT), which is set at 60% of the national median income.

$$\widehat{ARPT} = 0.6 \cdot \hat{q}_{0.5},$$

where $\hat{q}_{0.5}$ is the median.

$$\widehat{HCR} := \frac{\sum_j I(y_j < \widehat{ARPT}) w_j}{\sum_{j=1}^{n} w_j} \cdot 100$$

Using R-package laeken: Head count ratio

```
> # Loading libraries and the data
> library(laeken)
> data("eusilc")
>
> # Weighted HCR vs. unweighted HCR
> arpr("eqIncome", weights = NULL, data = eusilc)
Value:
[1] 14.04869

Threshold:
[1] 10848.8
> arpr("eqIncome", weights = "rb050", data = eusilc)
Value:
[1] 14.44422

Threshold:
[1] 10859.24
```

Reference: Alfons and Templ (2013)

Inequality indicator: Quintile Share Ratio

For a given sample, let $\hat{q}_{0.2}$ and $\hat{q}_{0.8}$ denote the weighted 20% and 80% quantiles, respectively. Using index sets $I_{\leq \hat{q}_{0.2}}$ and $I_{>\hat{q}_{0.8}}$, the quintile share ratio is estimated by

$$\widehat{QSR} := \frac{\sum_{j \in I_{>\hat{q}_{0.8}}} w_j y_j}{\sum_{j \in I_{\leq \hat{q}_{0.2}}} w_j y_j}.$$

```
> # Loading libraries and the data
> library(laeken)
> data("eusilc")

> # Weighted QSR
> qsr("eqIncome", weights = "rb050", data = eusilc)
Value:
[1] 3.971415
```

Reference: Alfons and Templ (2013)

Inequality indicator: Gini Coefficient

The Gini coefficient is estimated from a sample by

```
> # Loading libraries and the data
> library(laeken)
> data("eusilc")

> # Weighted Gini
> gini("eqIncome", weights = "rb050", data = eusilc)
Value:
[1] 26.48962
```

Reference: Alfons and Templ (2013)

Direct estimation at domain level

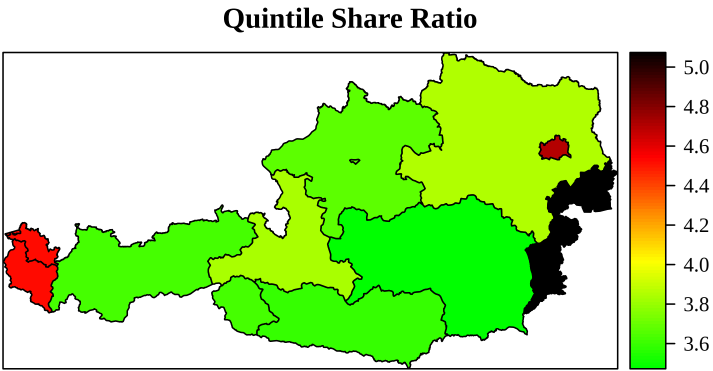
- One feature of `laeken` is that indicators can be computed for different subdomains (regions, age or gender).
- All the user needs to do is to specify such a categorical variable via the `breakdown` argument.
- Note that for the Head count ratio, the same overall at-risk-of-poverty threshold is used for all subdomains.

Using R-package laeken: QSR at domain level

```
> # Weighted QSR - breakdown by NUTS2
> qsr("eqIncome", weights = "rb050", data = eusilc,
   breakdown="db040")
Value:
[1] 3.971415

Value by domain:
         stratum    value
1     Burgenland 5.073746
2      Carinthia 3.590037
3  Lower Austria 3.845026
4       Salzburg 3.829411
5        Styria 3.472333
6         Tyrol 3.628731
7  Upper Austria 3.675467
8         Vienna 4.705347
9     Vorarlberg 4.525096
```

Reference: Alfons and Templ (2013)

Quintile share ratio breakdown by NUTS2



National Quintile share ratio: 3.97

Variance estimation

Measures of uncertainty

- Variance,
- Mean Squared Error (MSE)
- Coefficient of Variation

How can we estimate the variance of an estimator?

Resampling methods

- Jackknife
- Bootstrap

Analytical methods

- Taylor linerisation

Using R-package laeken: Variance estimation

```
> # Variance estimation
>
> # Weighted HCR
> hcr_national <- arpr("eqIncome", weights = "rb050",
    data = eusilc)
> variance("eqIncome", weights = "rb050", design = "db040
    ",data = eusilc, indicator = hcr_national, bootType =
     "naive", seed = 123,R=500)
Value:
[1] 14.44422

Variance:
[1] 0.08225841

Confidence interval:
   lower    upper
13.87129 15.00776
```

Reference: Alfons and Templ (2013)

Using R-package `laeken`: Variance estimation-subdomains

```
> hcr_nuts2<- arpr("eqIncome", weights = "rb050",
   breakdown = "db040", data = eusilc)
> variance("eqIncome", weights = "rb050", breakdown = "
   db040", design = "db040",
+           data = eusilc, indicator = hcr_nuts2, bootType
    = "naive", seed = 123,R=500)

Value by domain:
         stratum    value
1    Burgenland 19.53984
2     Carinthia 13.08627
3 Lower Austria 13.84362
...
6         Tyrol 15.30819
7 Upper Austria 10.88977
8        Vienna 17.23468
9    Vorarlberg 16.53731
```

Reference: Alfons and Templ (2013)

Using R-package `laeken`: Variance estimation-subdomains

```
Variance by domain:
         stratum       var
1     Burgenland 3.2426875
2      Carinthia 1.2348834
...
7 Upper Austria 0.3499630
8         Vienna 0.5600269
9     Vorarlberg 2.0032567

Confidence interval by domain:
         stratum     lower    upper
1     Burgenland 16.296501 23.13324
2      Carinthia 10.679302 15.24175
...
7 Upper Austria  9.720091 12.07298
8         Vienna 15.662437 18.62901
9     Vorarlberg 13.560864 19.14820
```

Reference: Alfons and Templ (2013)

Problems with direct estimation

- Often the sample not large enough for domain estimation
- Design of the survey does not account for competing interests regarding the targets of estimation
- Not all domains of interest include sampled units
- Small sample sizes $\rightarrow$ Large variance of direct estimates

Are the results reliable?

One way of measuring the reliability of estimates is by using the **coefficient of variation** (CV).
The CV is defined as the ratio of the standard deviation σ to the mean μ:

$$CV = 100 \cdot \frac{\sigma}{\mu}.$$

- Rule of thumb: CV up to 20% or 25% $\rightarrow$ reliable
- Cautious use of CV depending on the size of point estimates

3 – Small Area Estimation - Model-based methods

Content of the session

- Introduction to Small Area Estimation
- Model-based methods
- Focus on linear statistics e.g. small area averages

Introduction to Small Area Estimation

- Domain: sub-population of the population of interest planned or unplanned
  - Geographic areas (e.g. Regions, Provinces, Municipalities, Health Service Area)
  - Socio-demographic groups (e.g. Sex, Age, Race within a large geographic area)
  - Other sub-populations (e.g. the set of firms belonging to a industry subdivision)

Direct estimators may be unreliable due to small sample sizes.

Types of models & Data requirements

Unit level models

- Use unit-level data (e.g. from surveys) for model fit
- Area level covariates sufficient for small area prediction of averages
- Access to unit data $\rightarrow$ possible confidentiality issues

Area level models

- Use only area-level data for model fit and small area prediction

Reference: Jiang and Lahiri (2006)

Unit level models: Battese-Harter-Fuller model

Key Concept:
Include random area-specific effects to account for between area variation/ unexplained variability between the small areas.

Random effects model:
Notation: (i =domain, j =individual)

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}, j = 1, ..., n_i, i = 1, ..., m$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}.$$

- Random effects $u_i \sim N(0, \sigma_u^2)$
- Error term $e_{ij} \sim N(0, \sigma_e^2)$

Reference: Battese et al. (1988)

Unit level models: Battese-Harter-Fuller model

Empirical Best Linear Unbiased Predictor (EBLUP) of $\bar{y}_i$ is

$$\hat{\theta}_i^{BHF} = \hat{\bar{y}}_i = N_i^{-1}\Big\{\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}\Big\} = N_i^{-1}\Big\{\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} (\boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i)\Big\}$$

where

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \hat{\boldsymbol{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{y} \\
\hat{\boldsymbol{u}} &= \hat{\sigma}_u^2 \boldsymbol{Z}^T \hat{\boldsymbol{V}}^{-1} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
\hat{\boldsymbol{V}} &= \hat{\sigma}_u^2 \boldsymbol{Z}\boldsymbol{Z}^T + \hat{\sigma}_e^2 \boldsymbol{I}_n
\end{aligned}$$

The variance components are estimated by ML or REML theory.

Reference: Battese et al. (1988)

Analytic MSE estimation: The Battese-Harter-Fuller model

An MSE estimator of the small area estimator of the mean under BHF is

$$MSE(\hat{\theta}_i^{BHF}) = g_{1i}(\sigma_u^2, \sigma_e^2) + g_{2i}(\sigma_u^2, \sigma_e^2) + g_{3i}(\sigma_u^2, \sigma_e^2)$$

- $g_{1i}(\sigma_u^2, \sigma_e^2)$ is due to random effects
- $g_{2i}(\sigma_u^2, \sigma_e^2)$ is due to $\boldsymbol{\beta}$ estimate
- $g_{3i}(\sigma_u^2, \sigma_e^2)$ is due to the variance components

An approximately correct estimator of the MSE is

$$\widehat{MSE}(\hat{\theta}_i^{BHF}) = g_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + 2g_{3i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$$

Remark: Alternative (for more complex models) use bootstrap (parametric or non-parametric) or jackknife.

Reference: Prasad and Rao (1990)

Using R-package sae: The Battese-Harter-Fuller model

Based on a synthetic population

```
> # Direct estimation of mean using sae-package
> fit_direct<-direct(y=eqIncome,dom=region,data=eusilcS_
   HH,replace=T)
>
> # Estimation of the Unit-level model (Battese-Harter-
   Fuller)
> fit_EBLUP<-eblupBHF(formula=as.numeric(eqIncome)~py010n
    + py050n+hy090n,dom=region,data=eusilcS_HH,meanxpop=
   Xmean,popnsize=Popsize)
>
> # MSE estimation of the Unit-level model
> MSE_EBLUP<-pbmseBHF(formula=as.numeric(eqIncome)~py010n
    + py050n+hy090n,dom=region,data=eusilcS_HH,meanxpop=
   Xmean,popnsize=Popsize)
```

Reference: Molina and Marhuenda (2015)

Using R-package sae: The Battese-Harter-Fuller model

```
> # Comparison of direct and EBLUP
        Domains   Direct EBLUP_est    CV EBLUP_CV
     Burgenland 15781.61  20954.84 18.45     5.47
  Lower Austria 20476.21  20727.56  6.45     5.21
         Vienna 18996.19  21022.50  5.09     5.39
      Carinthia 20345.62  20526.51  9.01     5.74
         Styria 21184.01  20839.66  6.64     5.42
  Upper Austria 21074.00  21433.11  5.36     5.75
       Salzburg 18716.99  20841.91  7.41     5.74
          Tyrol 18060.43  20805.72 10.38     5.32
     Vorarlberg 18922.28  22028.77 10.69     5.93
```

Reference: Molina and Marhuenda (2015)

Outlier robust projective SAE: Robust EBLUP

Idea Replace $\hat{\boldsymbol{\beta}}$, $\hat{u}_i$ in EBLUP with outlier robust alternatives $\hat{\boldsymbol{\beta}}^{\psi}$, $\hat{u}_i^{\psi}$ leading to outlier robust predictor $y_{ij}^{\psi} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}^{\psi} + \boldsymbol{z}_{ij}^T \hat{u}_i^{\psi}$

$$\hat{\bar{y}}_i = N_i^{-1} \Big\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}^{\psi} \Big\}$$

Reference: Sinha and Rao (2009)

Outlier robust projective SAE: M-quantile estimation

Idea Model between area heterogeneity by fitting a different linear M-quantile models to each area (domain), leading to the outlier robust within area predictor $y_{ij}^{\psi} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{q(i)}^{\psi}$

$$\hat{\bar{y}}_i = N_i^{-1} \Big\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}^{\psi} \Big\}$$

Reference: Chambers and Tzavidis (2006)

Outlier robust predictive SAE: Bias corrected robust projective SAE

Idea Correct the possible bias of the robust projective estimator

$$\hat{\hat{y}}_i^{\psi\phi}(t) = \int t d\hat{F}_i^{WR}(t) = N_i^{-1}\Big[\underbrace{\sum_{j\in s_i} + \sum_{j\in r_i} \hat{y}_{ij}^{\psi}}_{\text{robust projective}} + \underbrace{\frac{N_i - n_i}{n_i}\sum_{j\in s_i}\hat{\omega}_{ij}^{\psi}\phi\Big\{\frac{y_{ij}-\hat{y}_{ij}^{\psi}}{\hat{\omega}_{ij}^{\psi}}\Big\}}_{\text{robust bias correction}}\Big]$$

- In session 4 we will explore the use of transformations under the linear mixed model when we are concerned about the validity of the model assumptions.

Reference: Chambers et al. (2014)

Area level models: The Fay-Herriot model

Sampling model

$$\hat{\theta}_i^{direct} = \theta_i + e_i$$

- $\hat{\theta}_i^{direct}$ is a direct design-unbiased estimator, for instance the Horvitz-Thompson estimator.
- e_i is the sampling error of the direct estimator.

Linking model

$$\hat{\theta}_i^{direct} = \mathbf{x}_i\boldsymbol{\beta} + u_i + e_i, \quad i = 1, \ldots, m,$$

where $u_i \sim N(0, \sigma_u^2)$ and $e_i \sim N(0, \sigma_{e_i}^2)$, with $\sigma_{e_i}^2$ assumed known.

Reference: Fay and Herriot (1979)

Area level models: The Fay-Herriot estimator

The EBLUP under the **Fay-Herriot** (FH) model is obtained by

$$\hat{\theta}_i^{FH} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \hat{u}_i$$

$$= \gamma_i \hat{\theta}_i^{direct} + (1 - \gamma_i) \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}},$$

where $\gamma_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\sigma_{e_i}^2}{n_i}}$ denotes the shrinkage factor for area i.

Analytic MSE estimation: The Fay-Herriot model

The MSE of the Fay-Herriot small area estimator is

$$MSE(\hat{\theta}_i^{FH}) = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2)$$

- $g_{1i}(\sigma_u^2)$ is due to random errors
- $g_{2i}(\sigma_u^2)$ is due to $\boldsymbol{\beta}$ estimate
- $g_{3i}(\sigma_u^2)$ is due to the estimate of σ_u^2

An approximately correct estimator of the MSE is

$$\widehat{MSE}(\hat{\theta}_i^{FH}) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2)$$

An alternative is to use bootstrap (e.g. parametric under the FH model) or jackknife techniques for MSE estimation.

Reference: Prasad and Rao (1990)

Using R-package sae: Fay-Herriot

Based on a synthetic population

```
> # Direct estimation of mean using sae-package
> fit_direct<-direct(y=eqIncome,dom=region,data=eusilcS_
   HH,replace=T)
>
> # Aggregation of the covariates on region level
> eusilcP_HH_agg<-tbl_df((eusilcP_HH))%>%group_by((region
   ))%>%summarise(hy090n=mean(hy090n))%>%
+ ungroup()%>%mutate(Domain=fit_direct$Domain)
>
> # Merging the datasets
> data_frame<-left_join(eusilcP_HH_agg,fit_direct,by="
   Domain")%>%mutate(var=SD^2)
>
> # Estimation of the FH-model
> fit_FH<-mseFH(formula=Direct ~ hy090n,vardir=var,data=
   as.data.frame(data_frame))
```

Reference: Molina and Marhuenda (2015)

Using R-package sae: Fay-Herriot

```
> # Comparison of direct and FH
         Domains SampSize   Direct   FH_est    CV FH_CV
     Burgenland       14 15781.61 16595.25 18.45 12.29
  Lower Austria       71 20476.21 19912.64  6.45  5.14
         Vienna       95 18996.19 20135.40  5.09  6.65
      Carinthia       34 20345.62 20260.46  9.01  4.30
         Styria       46 21184.01 20541.93  6.64  5.33
  Upper Austria       67 21074.00 19702.94  5.36  5.84
       Salzburg       26 18716.99 18908.88  7.41  5.82
          Tyrol       32 18060.43 19729.34 10.38  4.01
     Vorarlberg       15 18922.28 18342.81 10.69  6.22
```

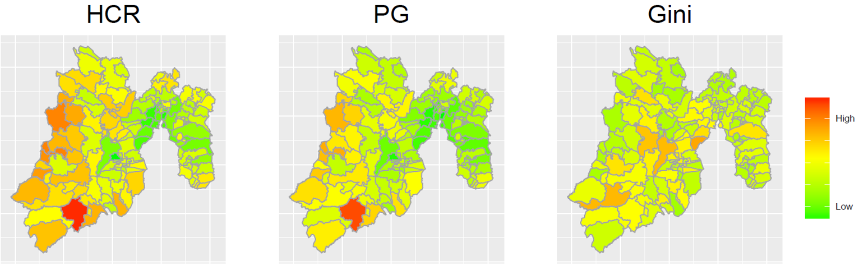
Reference: Molina and Marhuenda (2015)

4 – Small Area Estimation of non-linear indicators

Content of the session

- Empirical Best Prediction (EBP)
- Transformations in small area estimation
- Simulations studies

Typical results of poverty mapping



Non-linear income-based indicators

- Small area estimation methods mainly focus on estimating means and proportions
- New developments in SAE methodologies focus on estimating non-linear statistics e.g poverty/inequality indicators
- Methodology is general and covers linear and non-linear indicators

Data Requirements

Estimation of non-linear statistics require access to unit-level population covariates (e.g. Census microdata) $\rightarrow$ Access to such data is challenging

Recent methodologies

- The World Bank method (ELL)
  (Elbers et al., 2003)
- The Empirical Best Predictor (EBP) method
  (Molina and Rao, 2010)
- EBP based on normal mixtures
  (Elbers and van der Weide, 2014)
- Methods based on M-Quantiles
  (Tzavidis et al., 2010)

Empirical Best Prediction (EBP)

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, D,$$

1. Use the sample data to estimate $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{u}_i$ and $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.
2. For $l = 1, ..., L$
   - Compute $E(y_r|y_s)$ under the assumption of normal errors
   - Generate $e_{ij}^* \sim N(0, \hat{\sigma}_e^2)$ and $u_i^* \sim N(0, \hat{\sigma}_u^2 \cdot (1 - \hat{\gamma}_i))$, simulate a pseudo-population

$$y_{ij}{}^{*(l)} = \mathbf{x}_{ij}^T \hat{\beta} + \hat{u}_i + u_i^* + e_{ij}^*$$

   - Calculate the measures of interest, e.g. poverty indicator, $\theta_i^{(l)}$.
3. Obtain $\hat{\theta}_i^{EBP} = 1/L \sum_{l=1}^{L} \hat{\theta}_i^{(l)}$ for each area i.

Reference: Molina and Rao (2010).

Parametric bootstrap: MSE estimation

- Fit the random effects model to the original sample
- Generate $u_i^* \sim N(0, \hat{\sigma}_u^2)$, $e_{ij}^* \sim N(0, \hat{\sigma}_e^2)$
- Construct B bootstrap populations

$$y_{ij}^* = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^* + e_{ij}^*$$

- For each b population compute the population value θ_i^{*b}
- From each bootstrap population select a bootstrap sample
- Implement the EBP with the bootstrap sample, get $\hat{\theta}_i^{*b}$

$$\widehat{MSE}(\hat{\theta}_i) = B^{-1} \sum_{b=1}^{B} (\hat{\theta}_i^{*b} - \theta_i^{*b})^2$$

Reference: González-Manteiga et al. (2008).

Using R-package `emdi`: EBP method

- The R package **emdi** inlcudes two synthetic data sets
  - `eusilcS_HH`: sample data from Austrian regions about household income and demographics
  - `eusilcP_HH`: population micro-data for the Austrian regions

  → Both data sets contain the same covariates, measured in the same way

- Build a model for equivalized income in Austria

Reference: Kreutzmann et al. (2019).

Using R-package emdi: EBP method

Implemented in the R package **emdi** via function `ebp()`.

```
# EBP estimation function
ebp_au <- ebp(fixed = eqIncome ~ gender + eqsize +
                      py010n + py050n + py090n +
                      py100n + py110n + py120n +
                      py130n + hy040n + hy050n +
                      hy070n + hy090n + hy145n,
              pop_data = eusilcP_HH,
              pop_domains = "region",
              smp_data = eusilcS_HH,
              smp_domains = "region",
              pov_line = 0.6*median(eusilcS_HH$eqIncome
                 ),
              transformation = "no",
              L=50,
              MSE = T,
              B = 50)
```

Reference: Kreutzmann et al. (2019).

Using R-package emdi: EBP method - Summary output

```
# Summary for the EBP method
  > summary(ebp_au)
```

```
  Out-of-sample domains:  0
  In-sample domains:  9

  Sample sizes:
  Units in sample:  503
  Units in population:  25000
                 Min. 1st Qu. Median  Mean  3rd Qu. Max.
  Sample_domains   16   26       43   55.9    94    101
  Pop_domains     799 1671     1889   2778  4071   5857
```

Reference: Kreutzmann et al. (2019).

Using R-package emdi: EBP method - Summary output

```
Explanatory measures:
Marginal_R2 Conditional_R2
0.5198029      0.5198029

Residual diagnostics:
             Skewness Kurtosis Shapiro_W    Shapiro_p
Error         2.17646 12.5925  0.8551573   4.0933e-21
Random_effect 0.64311  2.6048  0.8870226   1.8589e-01

ICC:  2.610126e-08
```

Reference: Kreutzmann et al. (2019).

Motivating alternative methods

- EBP relies on Gaussian assumptions :
  - ✓ $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, the random area-specific effects
  - ✓ $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$

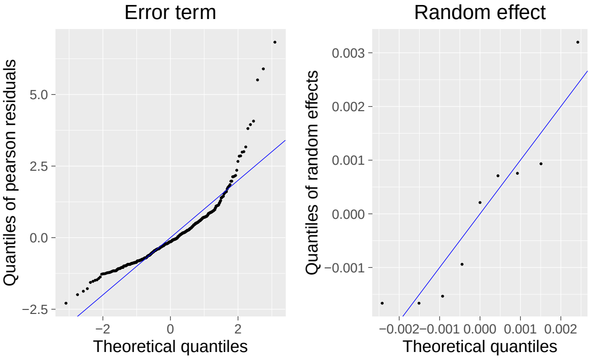
Model Checking (Residual diagnostics)

- Q-Q plots of residuals at different levels
- Influence diagnostics
- Plot standardised residuals vs fitted values - Heteroscedasticity

Graphical investigation of normality

Q-Q plots can help to assess the normality assumptions and it belongs to one of the plots that are automatically provided when applying the function `plot` to an `emdi` object.

```
# Residual diagnostics
    > plot(ebp_au)
```



Model adaptations

- Use an EBP formulation under an alternative distribution (Graf et al., 2015) - Model under generalised Beta distribution of the second kind
- Use robust methods as an alternative to transformations (Chambers and Tzavidis, 2006; Ghosh, 2008; Sinha and Rao, 2009; Chambers et al., 2014; Schmid et al., 2016).
- Use non-parametric models (Opsomer et al., 2008; Ugarte et al., 2009).
- Elaborate the random effects structure e.g. include spatial structures (Pratesi and Salvati, 2008; Schmid and Münnich, 2014).
- Use of transformations

Why transformations might help?

- Attempt to satisfy the model assumptions:
  - Normality: Reducing skewness and controlling kurtosis
  - Homoscedasticity: Variance-stabilization
  - Linearity: linearizing relation between variables

Use of transformations in SAE income applications

- Highly positive unimodal skewed and leptokurtic data sets
- Requires extensions of the transformations to the mixed model
- Appropriate for handling with zero and negative values
- Target parameters
  - Poverty gap, head count ratio
  - Gini coefficient, quantile share ratio

Transformations

- Shifted transformations
  - Log-shift
- Power transformations
  - Box-Cox
  - Exponential
  - Sign power
  - Modulus
  - Dual power
  - Convex-to-concave
- Multi-parameter transformations
  - Johnson
  - Sinh-arcsinh

Scaled transformations

- Using scaled transformations allows use of standard ML theory

Scaled Log-Shift Transformation (λ)

$$T_\lambda(y_{ij}) = \alpha \log(y_{ij} + \lambda),$$

Scaled Box-Cox Transformation (λ)

$$T_\lambda(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\alpha^{\lambda-1}\lambda}, & \lambda \neq 0 \\ \alpha \log(y_{ij} + s), & \lambda = 0 \end{cases},$$

Scaled Dual Power Transformation (λ)

$$T_\lambda(y_{ij}) = \begin{cases} \frac{2}{\alpha} \frac{(y_{ij}+s)^\lambda - (y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda > 0; \\ \alpha \log(y_{ij} + s) & \text{if } \lambda = 0. \end{cases}$$

with α chosen in such a way that the Jacobian of the transformation is 1.

Reference: Rojas-Perilla et al. (2019).

Estimation methods of (λ) for linear mixed models

- Skewness minimization
- Divergence minimization
- ML/REML

Estimation algorithm (λ)

REML Algorithm for the EBP Method:

1. Choose a transformation type
2. Define a parameter interval for λ
3. Set λ to a value inside the interval
4. Maximize the residual log-likelihood function conditional on fixed λ
5. Repeat 3 and 4 until maximum until $\hat{\lambda}$ is found
6. Apply the EBP method using $\hat{\lambda}$

Reference: Rojas-Perilla et al. (2019)

Parametric bootstrap for MSE estimation

1. For $b = 1, ..., B$
   - Using the already estimated $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}$ from the transformed data $T(y_{ij}) = \tilde{y}_{ij}$, simulate a bootstrap superpopulation $\tilde{y}_{ij}^{*(b)} = \mathbf{x}_{ij}^T \hat{\beta} + u_i^* + e_{ij}^*$
   - Transform $\tilde{y}_{ij}^{*(b)}$ to original scale resulting in ${y_{ij}}^{*(b)}$
   - For each b population compute the population value θ_i^{*b}
   - Extract the bootstrap sample in ${y_{ij}}^{*(b)}$ and use the EBP method.
   - Estimate λ with the bootstrap sample.
   - Obtain $\hat{\theta}_i^{*b}$.
2. $\widehat{MSE}(\hat{\theta}_i) = B^{-1} \sum_{b=1}^{B} (\hat{\theta}_i^{*b} - \theta_i^{*b})^2$

Reference: Rojas-Perilla et al. (2019)

Using emdi

Currently function ebp() includes a logarithmic or Box-Cox transformation and applies the EBP method.

```
# EBP estimation function under a Box-Cox
    transformation
ebp_au <- ebp(fixed = eqIncome ~ gender + eqsize +
                       py010n + py050n + py090n +
                       py100n + py110n + py120n +
                       py130n + hy040n + hy050n +
                       hy070n + hy090n + hy145n,
               pop_data = eusilcP_HH,
               pop_domains = "region",
               smp_data = eusilcS_HH,
               smp_domains = "region",
               pov_line = 0.6*median(eusilcS_HH$eqIncome
                  ),transformation = "box.cox",L=50,
                 MSE = T,B = 50)
```

Reference: Kreutzmann et al. (2019)

Using emdi - Summary output

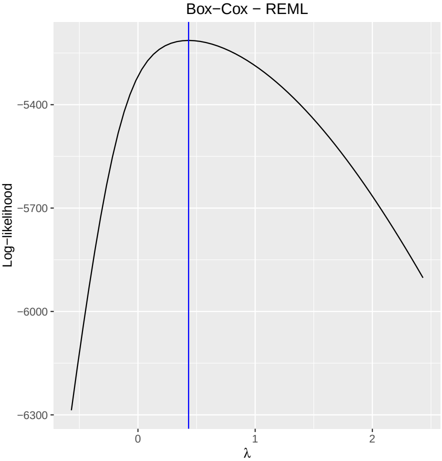
```
# Summary for the EBP method
  > summary(ebp_au)
```

```
Transformation:
 Transformation Method Optimal_lambda Shift_parameter
        box.cox   reml      0.4317972               0

Explanatory measures:
 Marginal_R2 Conditional_R2
   0.4543301      0.4543301

Residual diagnostics:
               Skewness Kurtosis Shapiro_W   Shapiro_p
Error           0.76051   6.3646   0.95643  4.9497e-11
Random_effect   0.58501   2.5533   0.95227  7.1501e-01
```

Reference: Kreutzmann et al. (2019)

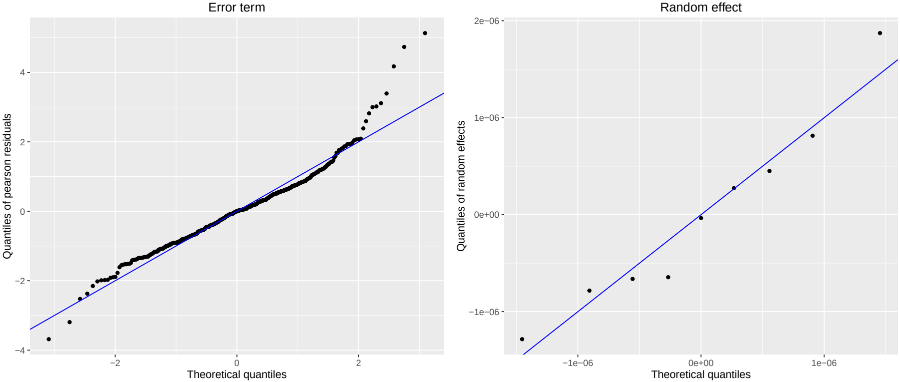
Finding $\hat{\lambda}$

Graphical representation of the optimal $\hat{\lambda}$ is made using the function `plot`.



Graphics

Q-Q plots of model residuals under the Box-Cox transformation.
Automatically provided when using function `plot`.



Model and Design-based simulations

Complementary Evaluations:

- **Model-based evaluation**
  - Uses synthetic data generated under a model
  - Sampling is performed repeatedly from the population generated in each Monte-Carlo round
  - Useful for evaluating performance and sensitivity of new methods under different assumptions
- **Design-based evaluation**
  - Uses frame data (e.g. census data) or synthetic data (not generated under a model) that preserve the survey characteristics
  - Sampling is performed repeatedly by keeping the population fixed
  - Useful for comparing competing methods in more realistic settings

Quality measures - R simulations

Root mean square error:

$$RMSE_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{i,r} - \theta_{i,r}\right)^2}$$

Relative bias [%]:

$$RB_i = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\theta}_{i,r} - \theta_{i,r}}{\theta_{k,r}} \cdot 100$$

Absolute bias:

$$Bias_i = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{i,r} - \theta_{i,r}$$

Model-based evaluation

Population data: is generated for $m = 50$ areas with $N = 200$ via

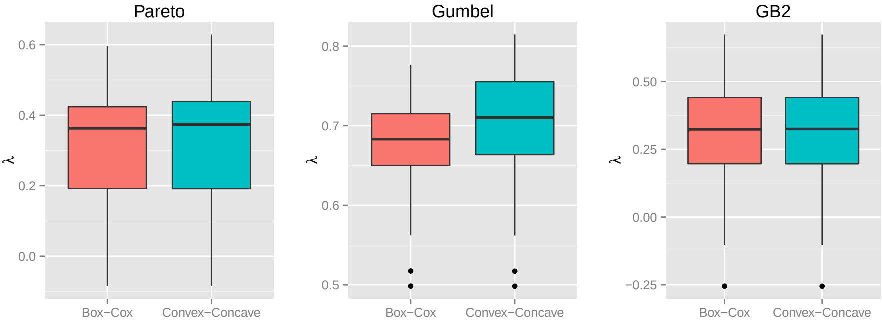
$$y_{ij} = 4500 - 400x_{ij} + u_i + e_{ij}$$

- Covariates $x_{ij} \sim N(\mu_i, 3^2)$ with $\mu_i \sim U(-3, 3)$
- Random effects $u_i \sim N(0, 500^2)$
- Unbalanced design leading to a sample size of $n = 921$ ($min = 8$, $mean = 18.4$, $max = 29$)
- 100 Monte Carlo replicates with L=50 bootstraps

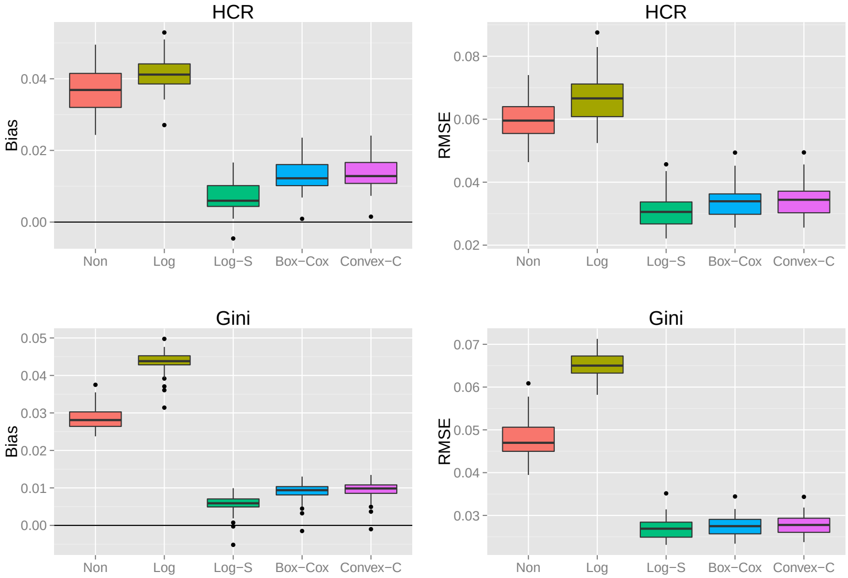
Scenarios: Three different income distribution are investigated:

$$e_{ij} \sim \text{Pareto}(2.5, 100)$$
$$e_{ij} \sim GB2(3, 700, 1, 0.8)$$
$$e_{ij} \sim Gumbel(1, 1000)$$

Estimated transformation parameters



Performance under the Pareto scenario using REML



Design-based evaluation: State of Mexico (EDOMEX)

- **Target geography**: State of Mexico is made up of 125 administrative divisions
- **Survey**: 58 are in-sample and 67 out-of-sample
- **Census**: From the 219514 households, there are 2748 in the sample
- **Sample sizes**:

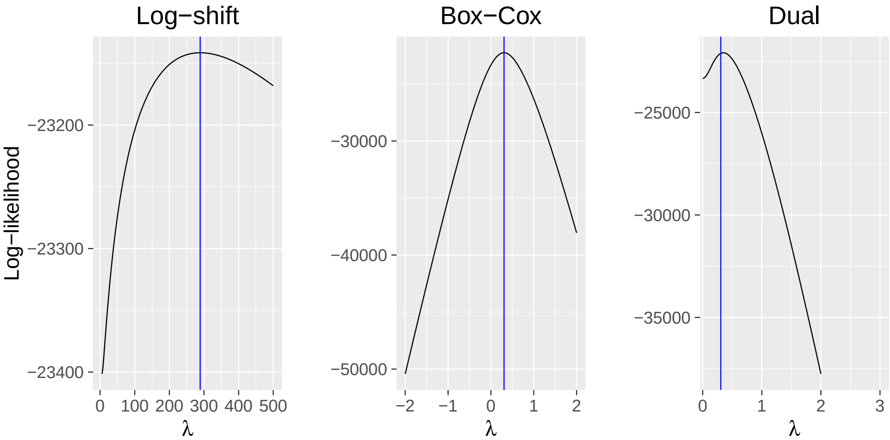
	Min.	Q1.	Median	Mean	Q3	Max.
Survey	3	17	21	47	42	527
Census	650	923	1161	1756	1447	13580

Outcome: Two income variables are available in the survey.
The target variable is available only on the survey. Earned per capita income from work is also available on the Census micro data

Design-based evaluation: Setup

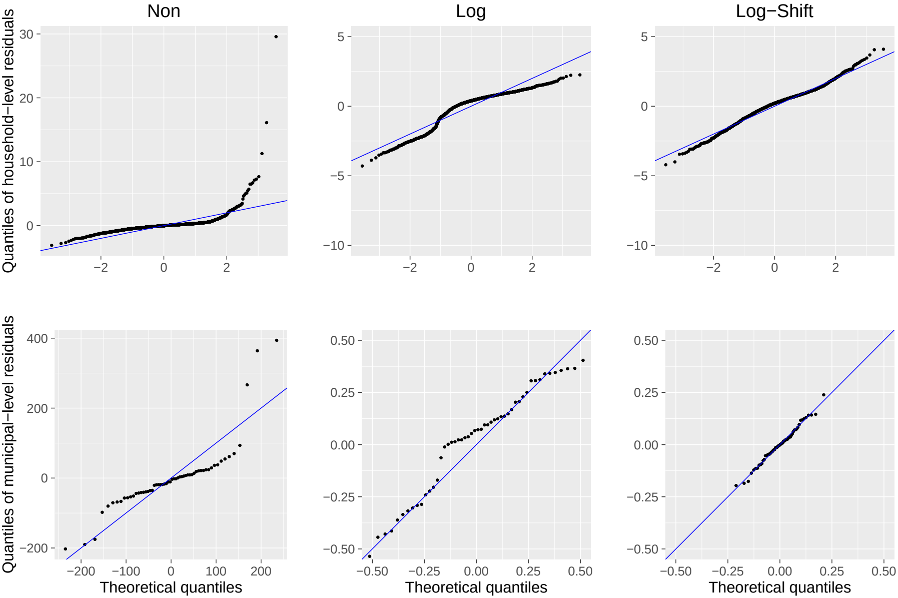
- Design-based simulation with 500 MC-replications repeatedly drawn from EDOMEX Census
- Unbalanced design leading to a sample size of $n = 2195$ ($\min = 8$, $\text{mean} = 17.6$, $\max = 50$)
- Sampling from each municipality

Transformation parameters - Estimation



	Log-shift	Box-Cox	Dual
λ	289.46	0.31	0.35

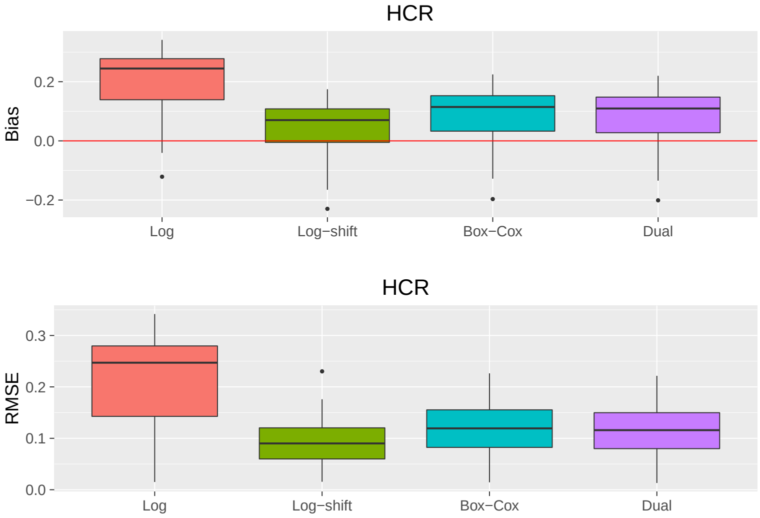
Residual diagnostics



Model diagnostics

Transformation	No	Log	Log-Shift	Box-Cox	Dual
R^2	0.30	0.40	0.52	0.48	0.48
ICC	0.004	0.046	0.032	0.029	0.027

Estimated HCR under alternative transformations



References I

Alfons, A., S. Kraft, M. Templ, and P. Filzmoser (2011). Simulation of close-to-reality population data for household surveys with application to eu-silc. *Statistical Methods & Applications 20*, 383–407.

Alfons, A. and M. Templ (2013). Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software 54*, 1–25.

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association 83*, 28–36.

Chambers, R., H. Chandra, N. Salvati, and N. Tzavidis (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B 76*, 47–69.

Chambers, R. and N. Tzavidis (2006). M-quantile models for small area estimation. *Biometrika 93*, 255–268.

CONEVAL (2010). Methodology for multidimensional poverty measurement in Mexico. Report.

Elbers, C., J. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. *Econometrica 71*, 355–364.

Elbers, C. and R. van der Weide (2014). Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. *World Bank Policy Research Working Paper No. 6962.*.

Fay, R. E. and R. A. Herriot (1979). Estimation of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association 74*, 269–277.

Ghosh, M. (2008). Robust estimation in finite population sampling. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, pp. 116–122.

González-Manteiga, W., M. Lombardía, I. Molina, D. Morales, and L. Santamaría (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation 78*, 443–462.

Graf, M., J. Marin, and I. Molina (2015). Estimation of poverty indicators in small areas under skewed distributions. In *Proceedings of the 60th World Statistics Congress of the International Statistical Institute*, The Hague, Netherlands.

Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–685.

References II

Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation. *TEST 15*, 1–96.

Kreutzmann, A.-K., S. Pannier, N. Rojas, T. Schmid, N. Tzavidis, and M. Templ (2019). emdi: An r package for estimating and mapping regional disaggregated indicators. *To appear, Journal of Statistical Software*.

Molina, I. and Y. Marhuenda (2015). sae: An R package for small area estimation. *The R Journal 7*, 81–98.

Molina, I. and J. N. K. Rao (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics 38*, 369–385.

Opsomer, J., G. Claeskens, M. Ranalli, G. Kauermann, and F. Breidt (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society Series B 70*, 265–283.

Prasad, N. G. N. and J. N. K. Rao (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association 85*, 163–171.

Pratesi, M. and N. Salvati (2008). Small area estimation: the eblup estimator based on spatially correlated random area effects. *Statistical Methods & Applications 17*, 113–141.

Rojas-Perilla, R., T. Schmid, N. Tzavidis, and S. Pannier (2019). Transformations of small area estimation methods for poverty mapping. Working paper.

Schmid, T. and R. Münnich (2014). Spatial robust small area estimation. *Statistical Papers 55*, 653–670.

Schmid, T., N. Tzavidis, R. Münnich, and R. Chambers (2016). Outlier robust small area estimation under spatial correlation. *Scandinavian Journal of Statistics 43*, 806–826.

Sinha, S. K. and J. N. K. Rao (2009). Robust small area estimation. *The Canadian Journal of Statistics 37*, 381–399.

Tzavidis, N., S. Marchetti, and R. Chambers (2010). Robust estimation of small area means and quantiles. *Australian and New Zealand Journal of Statistics 52*, 167–186.

Ugarte, M., T. Goicoa, A. Militino, and M. Durban (2009). Spline smoothing in small area trend estimation and forecasting. *Computational Statistics & Data Analysis 53*, 3616–3629.