Small area estimation using R, with application to poverty mapping[1]

Isabel Molina

Department of Statistics
Universidad Carlos III de Madrid

[1] Thanks to Swiss FSO, UN-ESCWA and PCBS.

INTRODUCTION TO SAE

- **Finite population:** U of size N.
- **Areas/domains:** $U_1, \ldots, U_D$, of sizes $N_1, \ldots, N_D$, which form a partition of U.
- **Sample:** s sample of size n obtained from U.
- **Sub-sample:** $s_d = s \cap U_d$ sub-sample from area/domain d, of size $n_d \leq N_d$.
- **Sample complement:** $r_d = U_d - s_d$ set of out-of-sample units from area/domain d.

INTRODUCTION TO SAE

- **Variable of interest:** Y_{di} for unit i within area/domain d.
- **Target indicators:** $\delta_d = \delta_d(Y_{d1}, \ldots, Y_{dN_d})$, $d = 1, \ldots, D$.

  Example:

$$\delta_d = \bar{Y}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} Y_{di}, \ d = 1, \ldots, D.$$

- **Direct estimator:** $\hat{\delta}_d$ based on the n_d sample survey observations from the area/domain: $\{Y_{di}, \ i \in s_d\}$.

  Example: HT or GREG/CAL for $\bar{Y}_d$:

$$\hat{\bar{Y}}_d^{DIR} = \frac{1}{N_d} \sum_{i \in s_d} w_{dj} Y_{di}.$$

INDIRECT ESTIMATION

- **Small area:** Area/domain U_d for which the **direct** estimator of the target indicator has **unacceptable** sampling error.
- **Indirect estimators:** They "borrow strength" from other areas by means of **homogeneity** assumptions that link all the areas, using information from auxiliary data sources.

MODEL-BASED ESTIMATORS

- They assume a regression model, typically including **random area effects** that account for the (unexplained) area heterogeneity.
- **Area-level:** Model assumed for the **area aggregates**.
- **Unit-level:** Model assumed for the variable of interest at the **unit level**.

R package `sae` (I. Molina and Y. Marhuenda)

It contains functions for estimation in small areas, including MSE estimation:

- Basic **area-level** (FH) model: `eblupFH()`, `mseFH()`.
- **Spatial** FH: `eblupSFH()`, `mseSFH()`, `pbmseSFH()`, `npbmseSFH()`.
- **Spatio-temporal** FH model: `eblupSTFH()`, `pbmseSTFH()`.
- Basic **unit-level** model (BHF): `eblupBHF()`, `pbmseBHF()`.
- EB method for estimation of **non-linear indicators** with unit-level model: `ebBHF()`, `pbmseebBHF()`.
- Basic **direct** and **indirect** estimators: `direct()`, `pssynt()`, `ssd()`

R package `sae` (I. Molina and Y. Marhuenda)

It also includes **data sets** and examples:

- `milk`
- `cornsoybean`
- `grapes`
- `incomedata`

Other R packages for SAE

- `Josae` (Johannes Breidenbach). Unit and area-level EBLUP estimators including also heteroscedasticity.
- `hbsae` (Harm Jan Boonstra): Hierarchical Bayes methods for basic area-level and unit-level models (also REML fitting).
- `BayesSAE` (Chengchun Shi) Bayesian methods for a variety of models, including unmatched models and spatial models.
- `saeeb` (Rizki Ananda Fauziah, Ika Yuni Wulansari): EB estimators under small area models for counts.
- `rsae` (Tobias Schoch). Robust methods for area and unit-level models.
- `saeRobust` (Sebastian Warnhol): Robust EBLUP under area-level models, including models with spatial and temporal correlation.

Other R packages for SAE

- `saeSim` (Sebastian Warnholz, Timo Schmid): Simulation and model fitting in SAE.
- `saery` (M.D. Esteban, D. Morales, A. Pérez): EBLUP for temporal area-level Rao-Yu model.
- `mme` (E. Lopez-Vizcaino, M.J. Lombardia and D. Morales). Multinomial area-level models for small area estimation of proportions, including models with temporal correlation.
- `saeME` (Muhammad Rifqi Mubarak, Azka Ubaidillah): SAE under measurement error of covariates.
- `emdi` (Sylvia Harmening, Ann-Kristin Kreutzmann, Soeren Pannier, Natalia Rojas-Perilla, Nicola Salvati, Timo Schmid, Matthias Templ, Nikos Tzavidis, Nora Würz): Functions that support estimating, assessing and mapping regional disaggregated indicators.

BASIC UNIT-LEVEL MODEL

- **Nested-error** model:

$$Y_{di} = \mathrm{x}'_{di}\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \ldots, N_d, \ d = 1, \ldots, D$$
$$u_d \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{di} \overset{iid}{\sim} N(0, \sigma_e^2)$$

- $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$ vector of unknown model parameters **(nuisance)**.
- $\boldsymbol{\beta}$ **common** for all areas, allowing to "borrow strength".
- u_d **area-specific** effect, modelling unexplained heterogeneity.

✓ *Battese, Harter & Fuller (1988), JASA*

BLUP UNDER NESTED-ERROR MODEL

- δ_d **linear** in $\mathrm{y}_d = (Y_{d1}, \ldots, Y_{dN_d})'$: Find the **linear** function of $\{Y_{di}; i \in s_d\}$ that is **unbiased** and has **minimum** MSE (BLUP).
- For an area mean:

$$\bar{Y}_d = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} Y_{di}\right).$$

- BLUP of $\bar{Y}_d$ under nested-error model:

$$\tilde{\bar{Y}}_d^{BLUP} = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} \tilde{Y}_{di}^{BLUP}\right).$$

- $\tilde{Y}_{di}^{BLUP} = \mathrm{x}_{di}'\tilde{\boldsymbol{\beta}} + \tilde{u}_d$ predicted values (BLUPs of Y_{di}, $i \in r_d$)
- $\tilde{\boldsymbol{\beta}}$ WLS estimator of $\boldsymbol{\beta}$, $\tilde{u}_d = \hat{E}(u_d|\mathrm{y}_s)$ BLUP of u_d.

BLUP UNDER NESTED-ERROR MODEL

- BLUP for areas with $n_d/N_d \approx 0$:

$$\tilde{\bar{Y}}_d^{BLUP} \approx \gamma_d \left\{ \bar{y}_{ds} + (\bar{\mathrm{X}}_d - \bar{\mathrm{x}}_{ds})'\tilde{\boldsymbol{\beta}} \right\} + (1-\gamma_d)\bar{\mathrm{X}}_d'\tilde{\boldsymbol{\beta}}.$$

- Composition between "survey-regression estimator", $\bar{y}_{ds} + (\bar{\mathrm{X}}_d - \bar{\mathrm{x}}_{ds})'\tilde{\boldsymbol{\beta}}$, and regression-synthetic estimator, $\bar{\mathrm{X}}_d'\tilde{\boldsymbol{\beta}}$, with weight:

$$\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/n_d}.$$

- It **automatically** "borrows strength" when it is actually needed.
- **Empirical** BLUP (EBLUP): Replace consistent estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in the BLUP.

EB METHOD: GENERAL INDICATORS

- δ_d **non linear** in $\mathrm{y}_d = (Y_{d1}, \ldots, Y_{dN_d})'$
- $\mathrm{y}_d = (Y_{d1}, \ldots, Y_{dN_d})' = (\mathrm{y}_{ds}', \mathrm{y}_{dr}')'$
- y_{ds} sample part (survey), y_{dr} non-sample part (unknown) .
- **Best predictor** of $\delta_d = \delta_d(\mathrm{y}_d)$: Minimizes the MSE,

$$\tilde{\delta}_d^B(\boldsymbol{\theta}) = E_{\mathrm{y}_{dr}}[\delta_d(\mathrm{y}_d)|\mathrm{y}_{ds}; \boldsymbol{\theta}],$$

- **Empirical** best (EB): Replace consistent estimator $\hat{\boldsymbol{\theta}}$.
- EB requires **linking** the survey and census/register data sets.
- **Census EB:** Variation of EB when survey and census/register data sets cannot be linked (updated WB method).

✓ *Molina & Rao (2010), CJS*

FGT POVERTY INDICATORS

- E_{di} welfare measure for individual i ($i = 1, \ldots, N_d$) from area/domain d ($d = 1, \ldots, D$).
- z poverty line.
- FGT indicator of order $\alpha \geq 0$ for domain d:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{z - E_{di}}{z} \right)^{\alpha} I(E_{di} < z), \quad \alpha \geq 0.$$

- $\alpha = 0 \Rightarrow$ at risk of poverty rate (frequency)
- $\alpha = 1 \Rightarrow$ poverty gap (depth)

✓ *Foster, Greer & Thornbecke (1984), Econom.*

BP: FGT INDICATORS

- Welfares have a markedly right-skewed distribution.
- We assume the nested error model for a one-to-one transformation $Y_{di} = T(E_{di})$; e.g. $Y_{di} = \log(E_{di} + c)$, $c > 0$.
- FGT poverty indicator in terms of model responses Y_{di}:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left\{ \frac{z - T^{-1}(Y_{di})}{z} \right\}^{\alpha} I\left\{ T^{-1}(Y_{di}) < z \right\} = \delta_d(\mathrm{y}_d).$$

- Best predictor of $F_{\alpha d}$:

$$\tilde{F}^{B}_{\alpha d}(\boldsymbol{\theta}) = E_{\mathrm{y}_{dr}}(F_{\alpha d}|\mathrm{y}_{ds}; \boldsymbol{\theta}).$$

✓ *Molina & Rao (2010), CJS*

EMPIRICAL BEST PREDICTOR

- $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ consistent estimator of $\boldsymbol{\theta}$.
- **Empirical best predictor (EB):**

$$\hat{F}_{\alpha d}^{EB} = \tilde{F}_{\alpha d}^{B}(\hat{\boldsymbol{\theta}}).$$

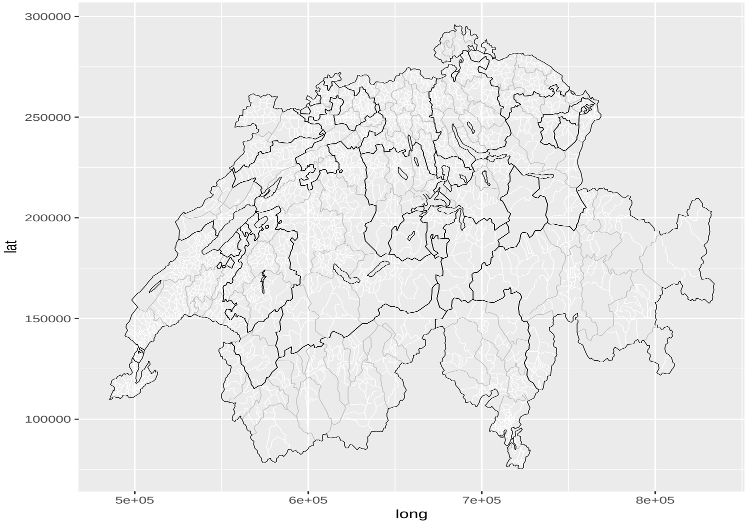
- They can be calculated **analytically** for $\alpha = 0, 1$ and for $Y_{di} = \log(E_{di} + c)$, $c > 0$.
- In general, they can be approximated by **Monte Carlo**.
- MSE under the model can be approximated by **bootstrap**.

✓ *Molina & Rao (2010), CJS*

DATA DESCRIPTION

- **Data:** 2012 Structural Survey (employment, housing) and two administrative registers: the 2012 population and household statistic and the 2011 old age survival insurance (OASI) data (social insurance system).
- **Target variable:** $Y_{di} \in \{0, 1\}$, 1=active, 0=non active.
- **Areas:** In register/s, 2485 **communes**→ $D = 2475$ in survey.
- **Target:** Estimate activity rates for the sampled Swiss communes,

$$\bar{Y}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} Y_{di}, \ d = 1, \dots, D.$$



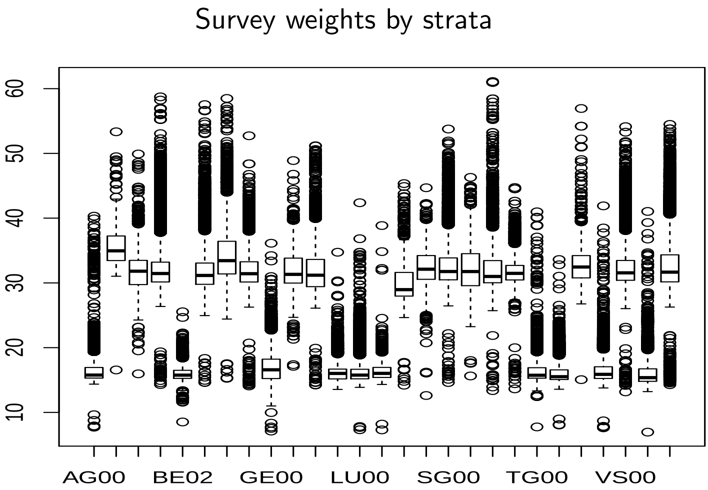
Switzerland: Cantons (black line), districts (grey line) and communes (white line) in 2017. ©OFS, ThemaKart.

DATA DESCRIPTION

- **Sample size:** $n = 286,015$ out of $N = 6,662,333$ Swiss permanent residents aged 15 years or older who live in private households.
- Num. communes with sample sizes below selected levels:

$n_d \leq$	10	20	30	40	50	100
# communes	356	697	964	1173	1337	1792

SAMPLING DESIGN

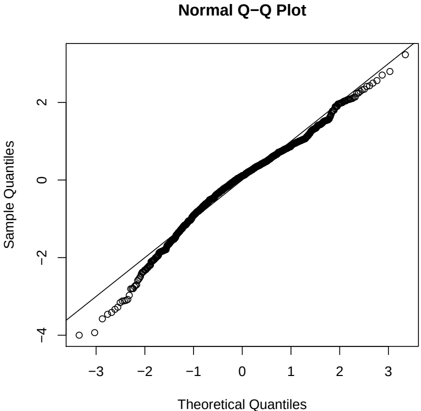


FITTED MODEL

- ✓ Dummy indicator of strata group.
- ✓ Fixed effects for outlying district and commune.
- ✓ Age group, gender, civil status, Swiss nationality, secondary residence, household size, income group, contrib. to OASI only 1st quarter.
- ✓ Interactions: age group$\times$gender, civil status$\times$gender.

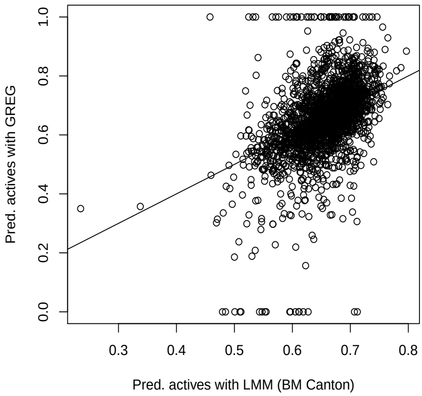
MODEL CHECKING

- All covariates with **significant** categories and regression coef. with **reasonable** signs.
- **Failure classification rate:** 11.2 %
- Approximate normality of predicted commune effects:

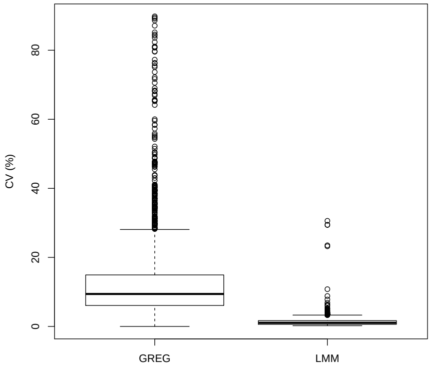


ACTIVITY RATES IN SWISS COMMUNES

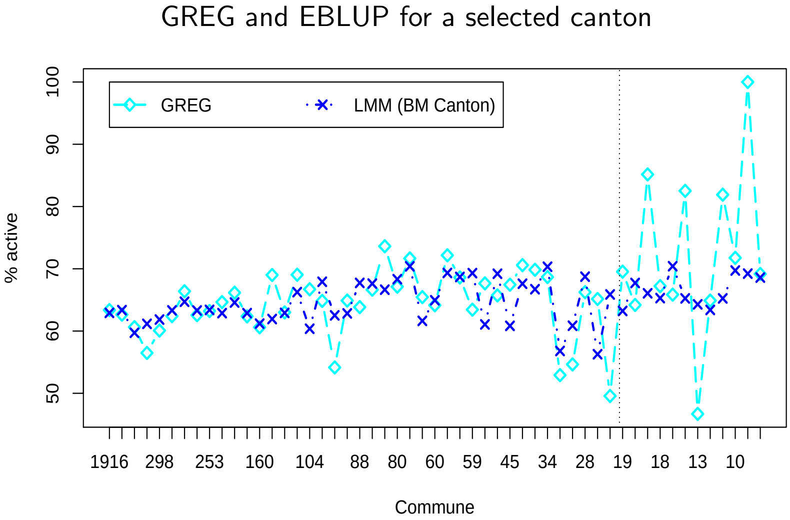
GREG vs. benchmarked EBLUP



Estimated CVs

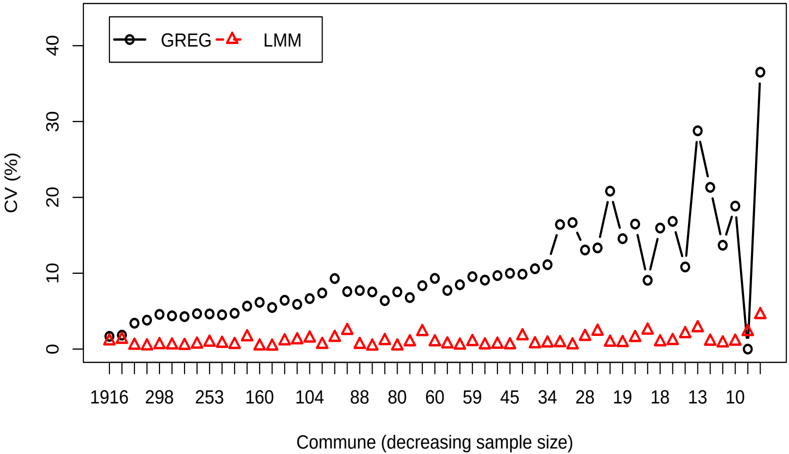


ACTIVITY RATES IN SWISS COMMUNES



ACTIVITY RATES IN SWISS COMMUNES

Estimated CV(GREG) and RRMSE(EBLUP)



DATA DESCRIPTION

- **Data:** Palestinian Expenditure Consumption Survey (PECS) from 2016/2017 and Population Census from 2017.
- **Target:** Estimate poverty rates and gaps for Palestinian localities by gender.
- **Areas:** In census, 319 **localities**$\rightarrow$ $D = 162$ in survey.
We compute estimates for each **sampled** locality by gender.
- **Welfare measure:** E_{dj} monthly expenditure per adult equivalent (ILS).
- **Poverty line:** $z = 10,027$ ILS $\rightarrow$ approx. **26 %** popn. below pov. line.

FITTED MODEL

- We fit a separate model for each gender.
- **Explanatory variables:**
  - ✓ Indicators of region (Gaza, West Bank), type of locality (rural/urban, camp).
  - ✓ Household characteristics (size, prop. females, employed ratio).
  - ✓ Household head characteristics (unemployed, employisrasett, employnatgov, refugstat, diff, neverschool, secondabove).
  - ✓ Dwelling characteristics (type, tenure, num. rooms).
  - ✓ Supplies (water, waste, heating systems, freezer, etc.)

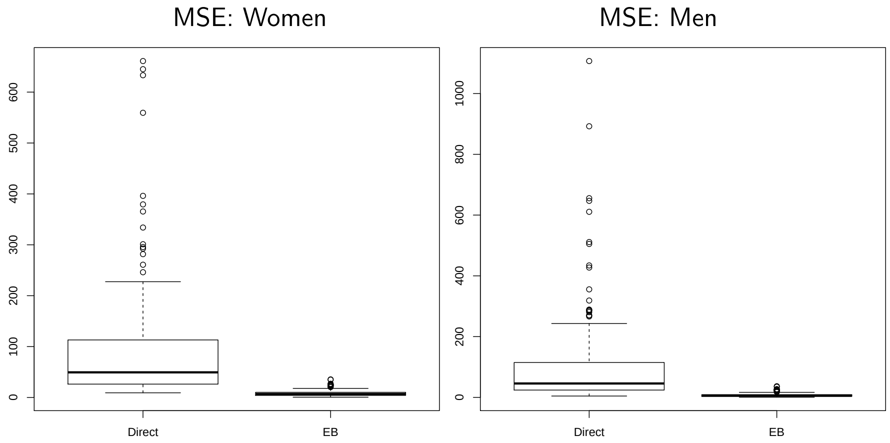
✓ *García-Portugués & Molina (2020), ESCWA.*

MODEL CHECKING

- Model coefficients take reasonable signs.
- All covariates with significant categories for both genders.
- **Explanatory power:** $R^2 = 53.6\,\%$, both genders.
- Data indicates nothing against normality of model residuals, linearity, heteroscedasticity. Model seems to fit well.

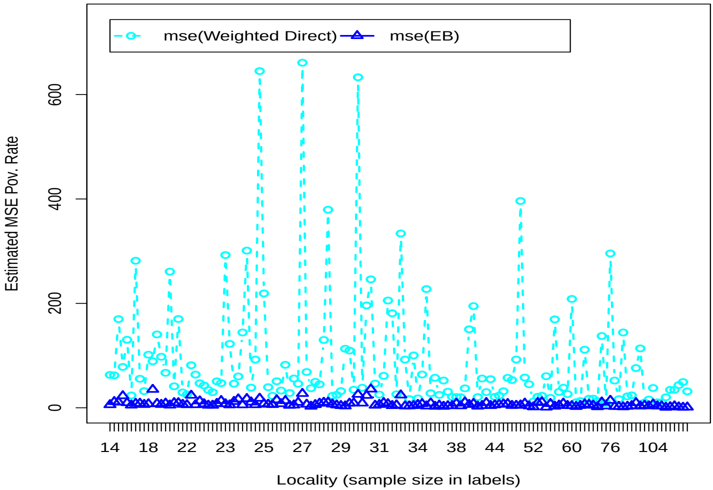
QUALITY EB vs. DIRECT: POV. RATE

- ✓ Median MSE Women: Direct **47**, EB: **6.7**
- ✓ Median MSE Men: Direct **45.8**, EB: **5.5**

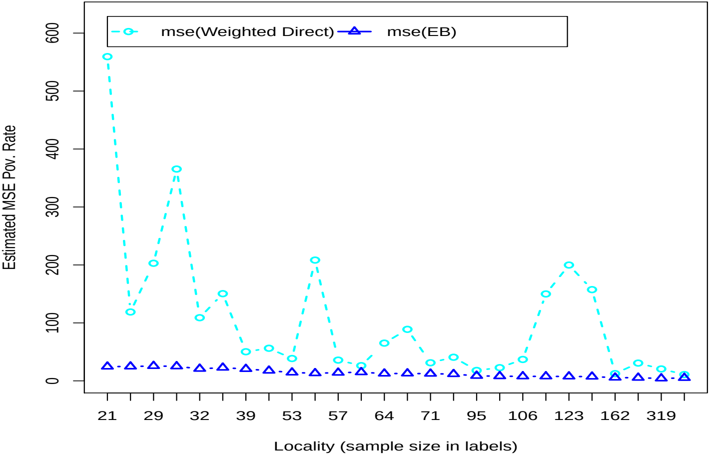


EB vs. DIRECT: WOMEN, WEST BANK

✓ Reduction in **all** but one locality, **84 %** average MSE reduction!



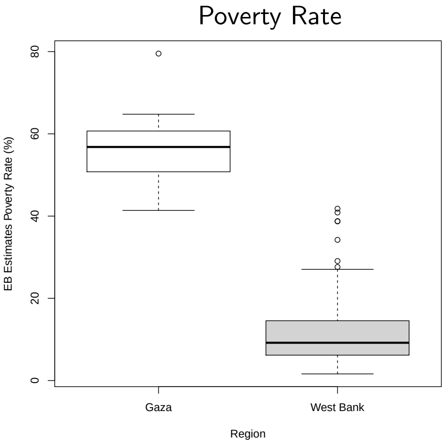
EB vs. DIRECT: WOMEN, GAZA

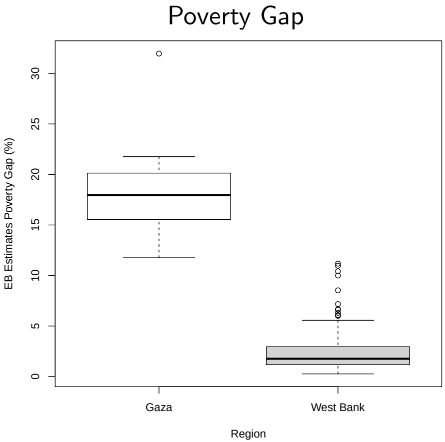


✓ **Great gains** also for Pov. **Gap** (not shown)!

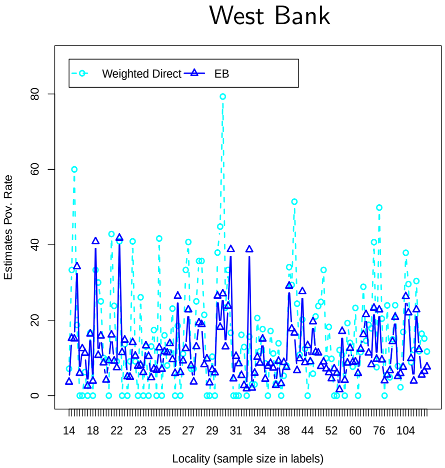
ESTIMATES BY REGION

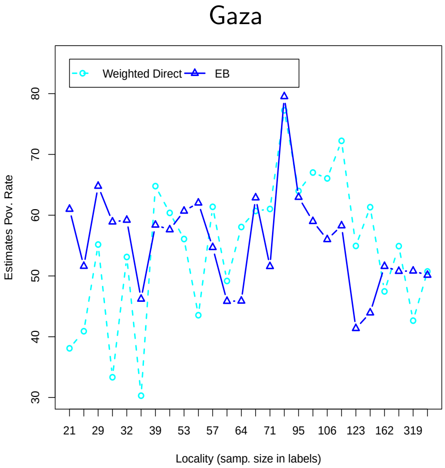
✓ Median Pov. Rate: Gaza **55 %**, West Bank: **8.3 %**
✓ Median Pov. Gap: Gaza **17.4 %**, West Bank: **1.5 %**





ESTIMATED POV. RATE: WOMEN





Union is strength!!