Study on the Limits of Data Disaggregation in Household Surveys for Population Subgroups and Geographical Areas and the Requirements to Overcome them

Application to poverty mapping in Palestine

Isabel Molina and Eduardo García
Dept. of Statistics, Univ. Carlos III de Madrid

BASIC TERMINOLOGY

- **Areas/domains:** Subpopulations.
- **Direct estimator:** Based on survey data for the target area/domain.
- **Small area:** Area/domain for which the considered direct estimator of the target indicator is inefficient (too large sampling error).

POVERTY AND INEQ. INDICATORS

- E_{dj} **welfare** measure for indiv. j in domain d.
- $z =$ poverty line.
- **FGT poverty indicator of order α for domain d:**

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

- When $\alpha = 0 \Rightarrow$ **Poverty incidence** (or at-risk-of-poverty rate)
- When $\alpha = 1 \Rightarrow$ **Poverty gap**
- **Other:** Quintile share ratio, Gini coef., Sen index, Theil index, Generalized entropy, Fuzzy monetary/supplementary index.

✓ *Foster, Greer & Thornbecke (1984), Econom.*
✓ *Neri, Ballini & Betti (2005), Stat. in Transition*

DIRECT ESTIMATORS

- FGT pov. indicator as a mean:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} F_{\alpha dj}, \quad F_{\alpha dj} = \left(\frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z)$$

- Direct estimators:

$$\hat{F}^{DIR}_{\alpha d} = \frac{1}{\sum_{j \in s_d} w_{dj}} \sum_{j \in s_d} w_{dj} F_{\alpha dj}, \quad \hat{F}^{S}_{\alpha d} = \frac{1}{n_d} \sum_{j \in s_d} F_{\alpha dj}.$$

- Very **inefficient** (unstable) for small n_d. Estimates can be **zero** for several areas!

INDIRECT ESTIMATION

- **Indirect estimator:** It **borrows strength** from other areas by making some kind of **homogeneity** assumption across areas (model with **common** parameters) that uses **auxiliary information**.
- **Synthetic estimators:** Do not allow for between-area heterogeneity beyond that explained by covariates.
- **Non-Synthetic estimators:** Do incorporate between-area heterogeneity beyond that explained by covariates.

NESTED ERROR MODEL

- The distribution of expenditures E_{dj} is highly right skewed.
- Select a transformation $T()$ such that the distribution of $y_{dj} = T(E_{dj})$ is approximately Normal.
- **Assumption:** $y_{dj} = T(E_{dj})$ satisfies the **nested error model**:

$$y_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \ldots, N_d, \; d = 1, \ldots, D$$

$$u_d \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2)$$

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

EB METHOD FOR POVERTY ESTIMATION

- Poverty indicators in terms of $\mathbf{y}_d = (y_{d1}, \ldots, y_{dN_d})'$:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left\{ \frac{z - T^{-1}(y_{dj})}{z} \right\}^{\alpha} I\left\{ T^{-1}(y_{dj}) < z \right\} = h_{\alpha}(\mathbf{y}_d).$$

- Partition $\mathbf{y}_d$ into sample and out-of-sample: $\mathbf{y}_d = (\mathbf{y}'_{ds}, \mathbf{y}'_{dr})'$
- **Best predictor:** Minimizes the MSE

$$\tilde{F}^{B}_{\alpha d} = E_{\mathbf{y}_{dr}} \left[F_{\alpha d} | \mathbf{y}_{ds}; \boldsymbol{\beta}, \sigma^2_u, \sigma^2_e \right].$$

- **Empirical best (EB) predictor:** $\hat{F}^{EB}_{\alpha d} = \tilde{F}^{B}_{\alpha d}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_u, \hat{\sigma}^2_e)$.

✓ *Molina and Rao (2010), CJS*

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.

SAMPLE SIZES

- Sample size: $n = 18,363$ out of $N = 4,266,953$ (43 out of 10,000).

	Women	Men	Total
Gaza	2569	2578	5147
West Bank	6550	6666	13216
Total	9119	9244	18363

- Sample sizes of localities by gender:

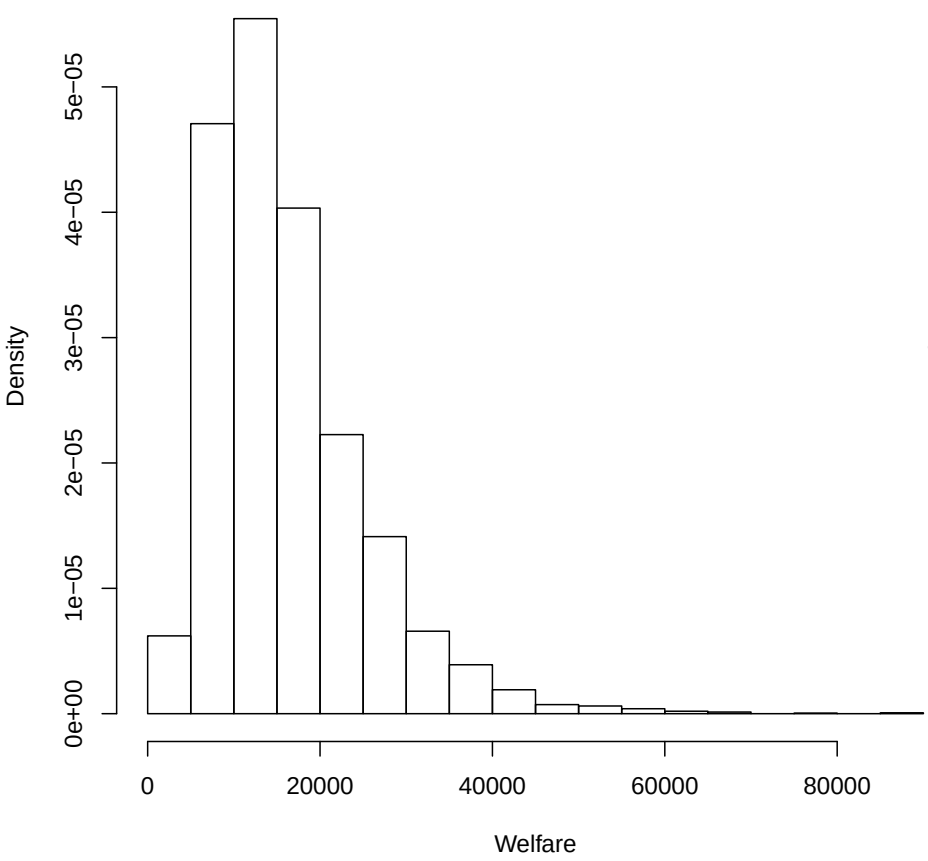
	Min	1st Qu	Median	Mean	3rd Qu.	Max
Women	14.00	26.00	35.00	56.29	61.50	405.00
Men	13.00	28.00	36.00	57.06	63.00	464.00

FITTED MODEL

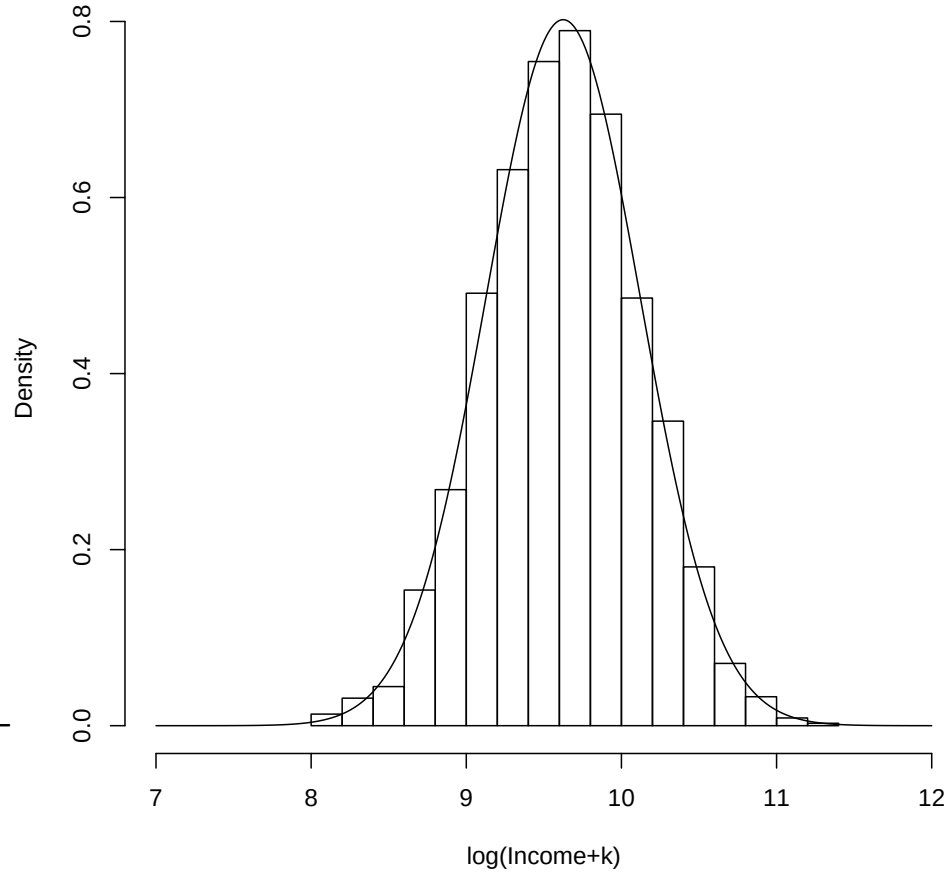
- We fit a separate model for each gender.
- **Transformation:** We consider the nested-error model for the log of expenditure: $y_{dj} = T(E_{dj}) = \log(E_{dj} + k)$.
- **Explanatory variables:**
  - ✓ Indicators of region (Gaza, West Bank), type of locality (rural/urban, camp).
  - ✓ Household characteristics (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.)
- **Fitting results:** All covariates with significant categories for both genders.
- **Explanatory power:** $R^2 = 53.6\,\%$, both genders.

NORMALITY EXPENDITURE

Original scale

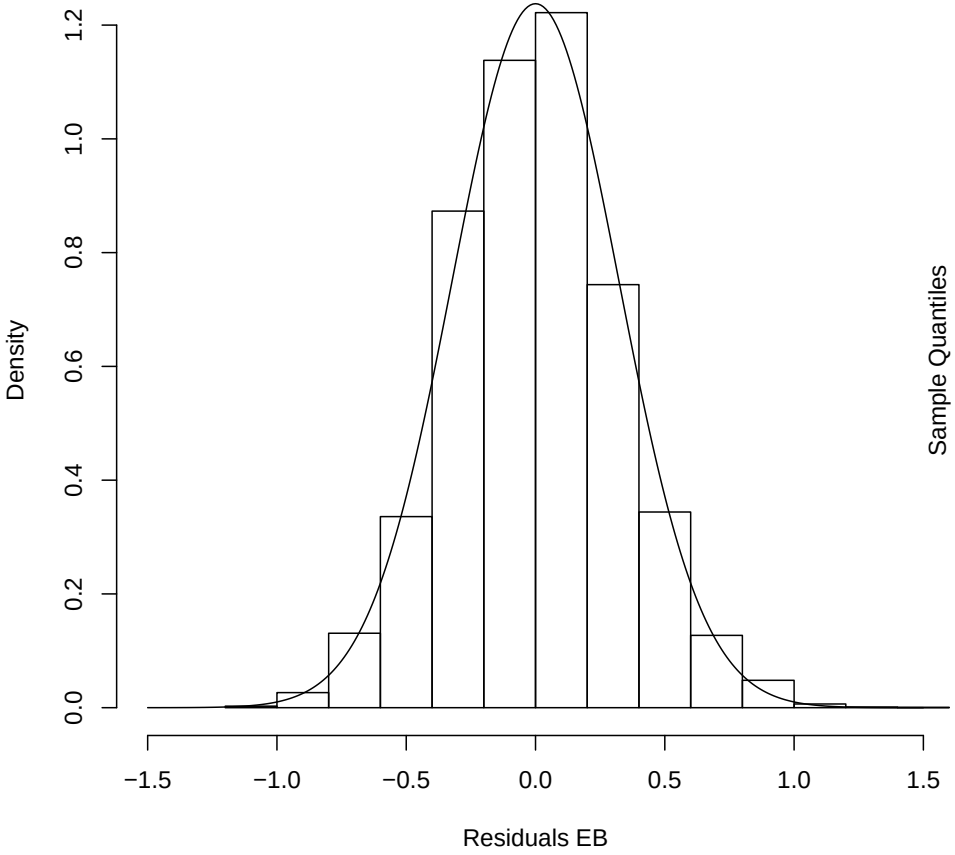


Log-scale after shift

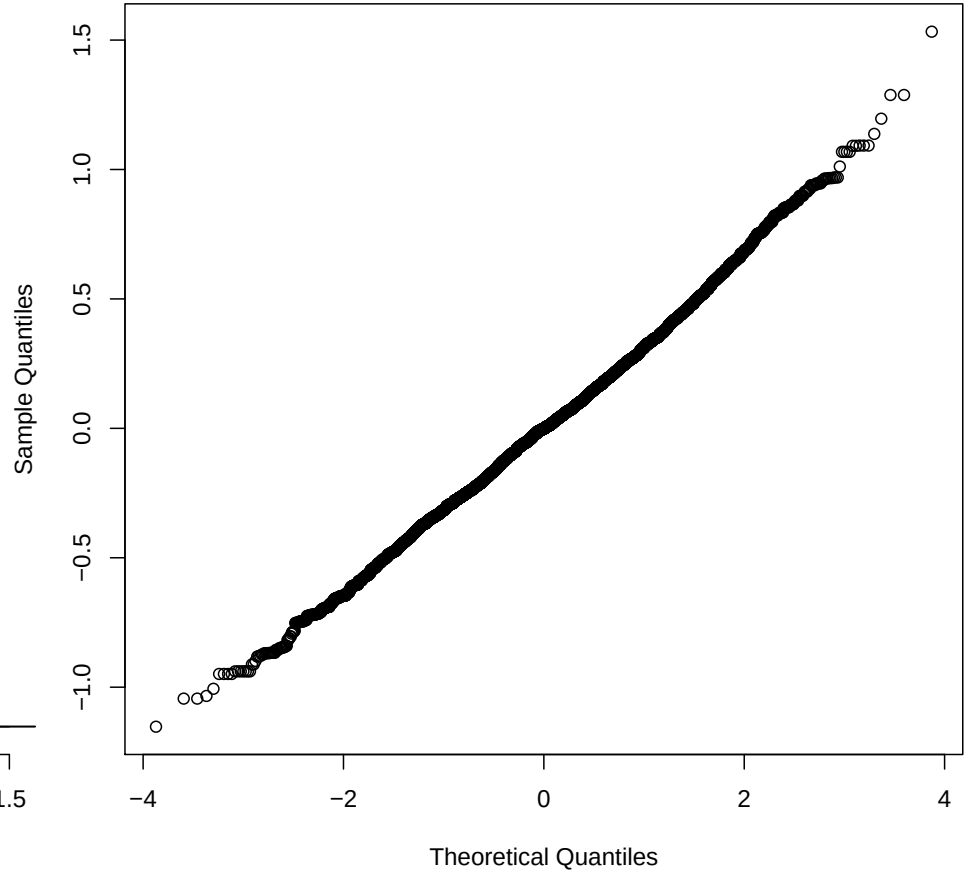


NORMALITY UNIT LEVEL RESIDUALS

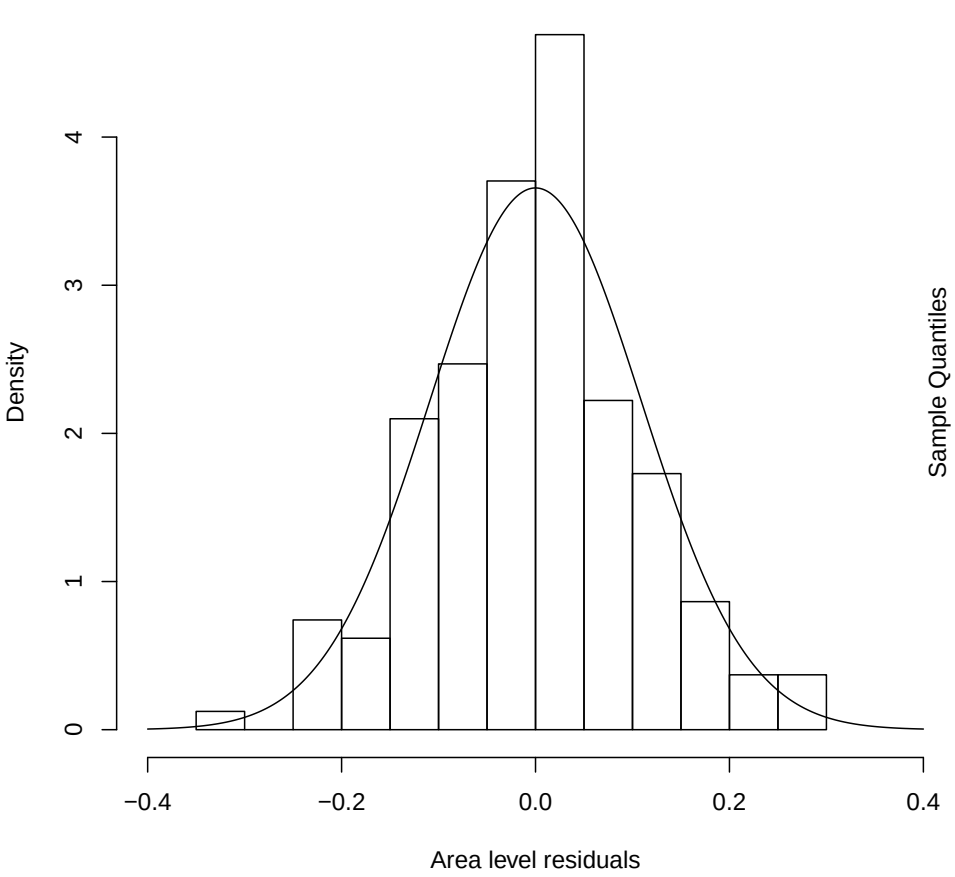
Histogram

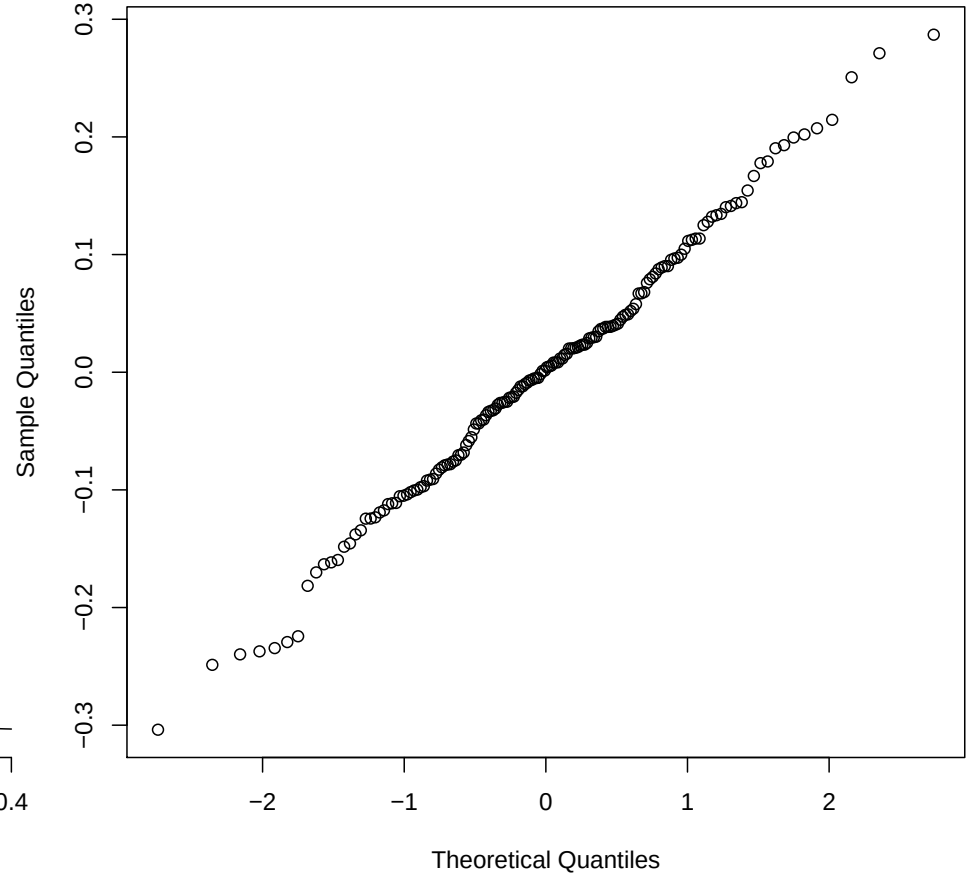


QQ normality plot

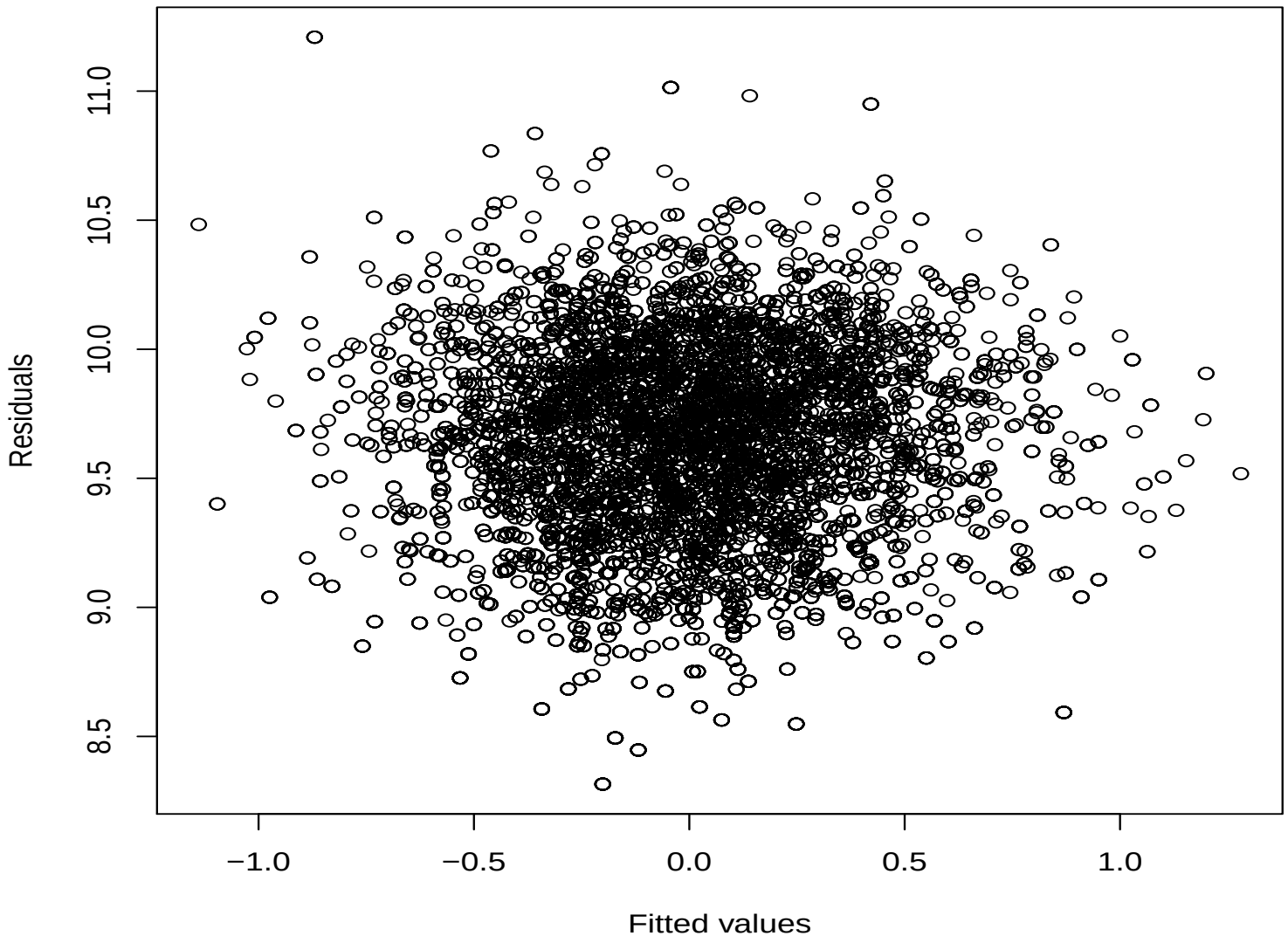


NORMALITY AREA EFFECTS





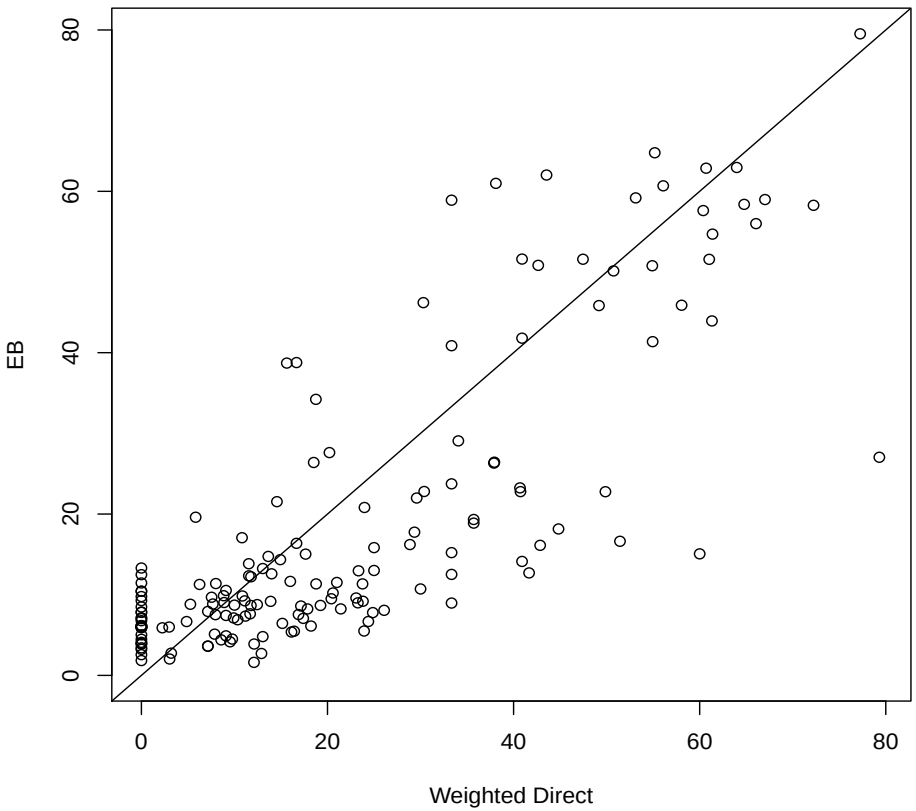
RESIDUALS vs. FITTED VALUES



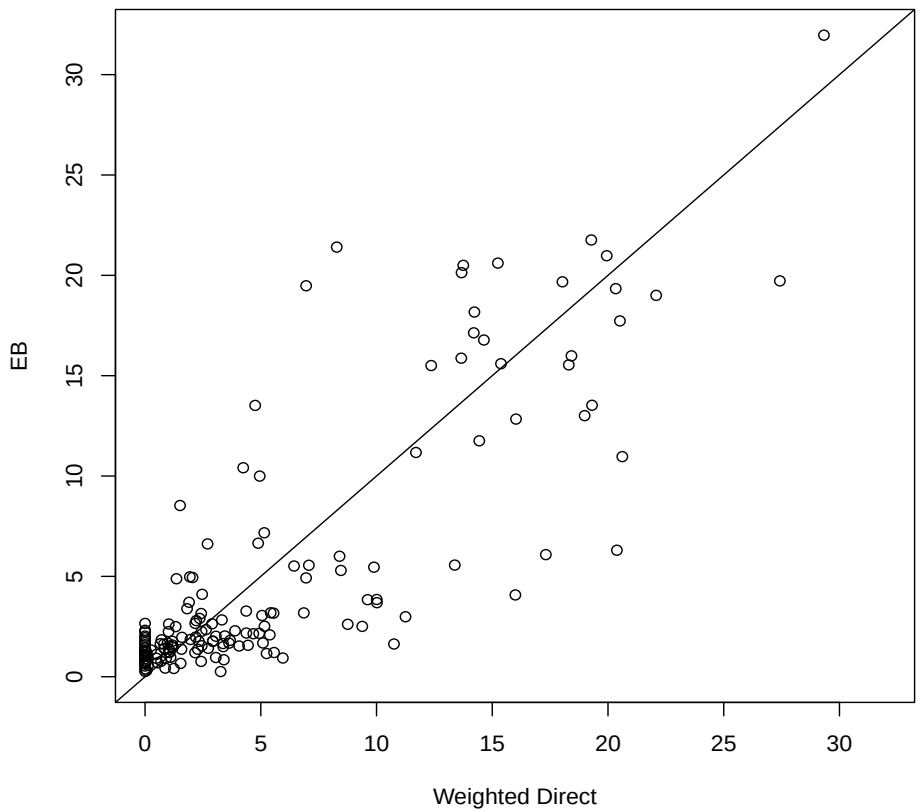
EB vs. DIRECT

✓ **No** visible **systematic bias** for EB!

Pov. Rate

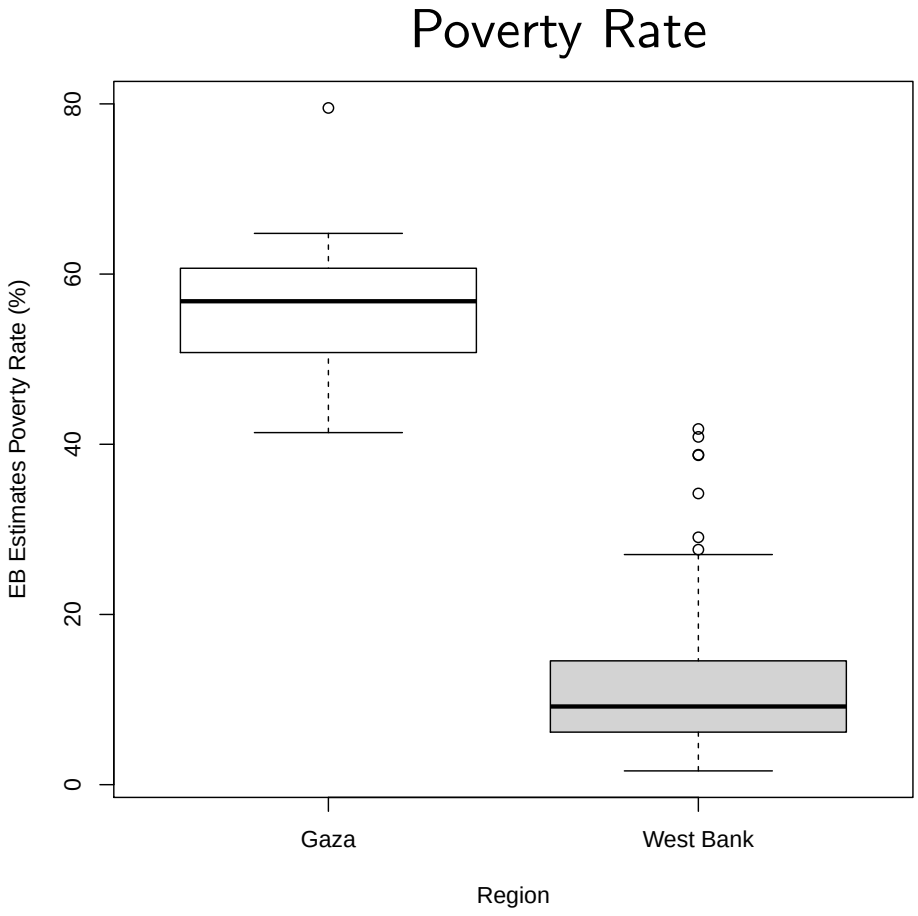


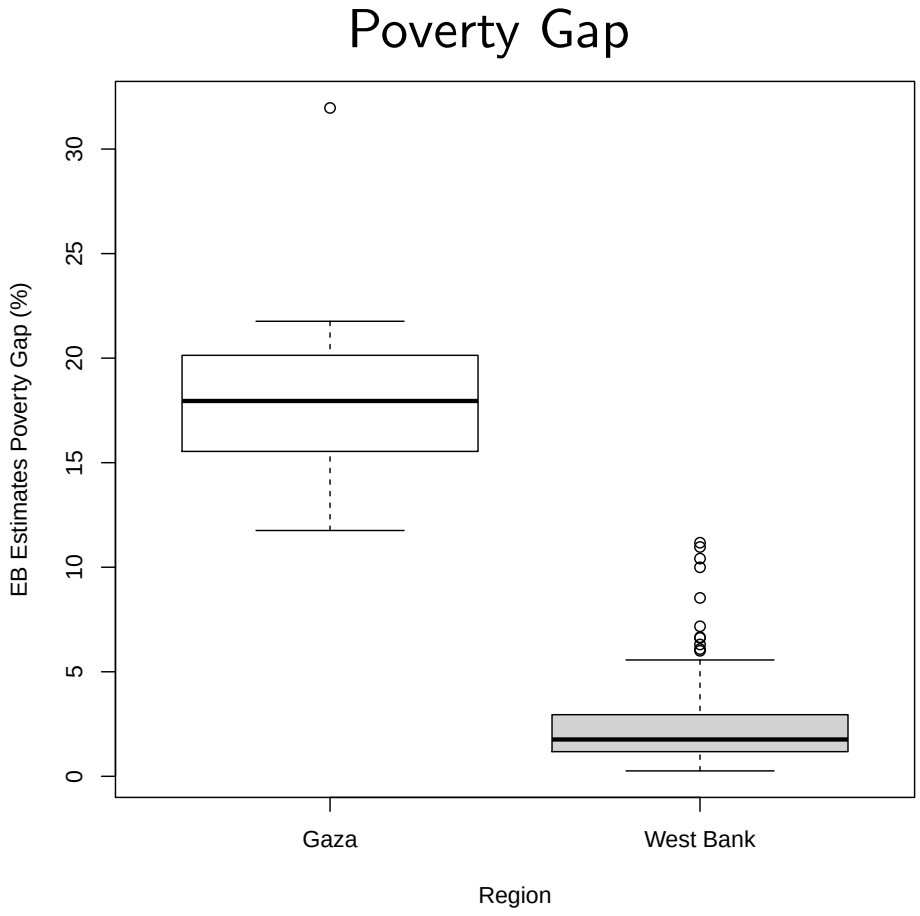
Pov. Gap



COMPARISON BY REGION

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

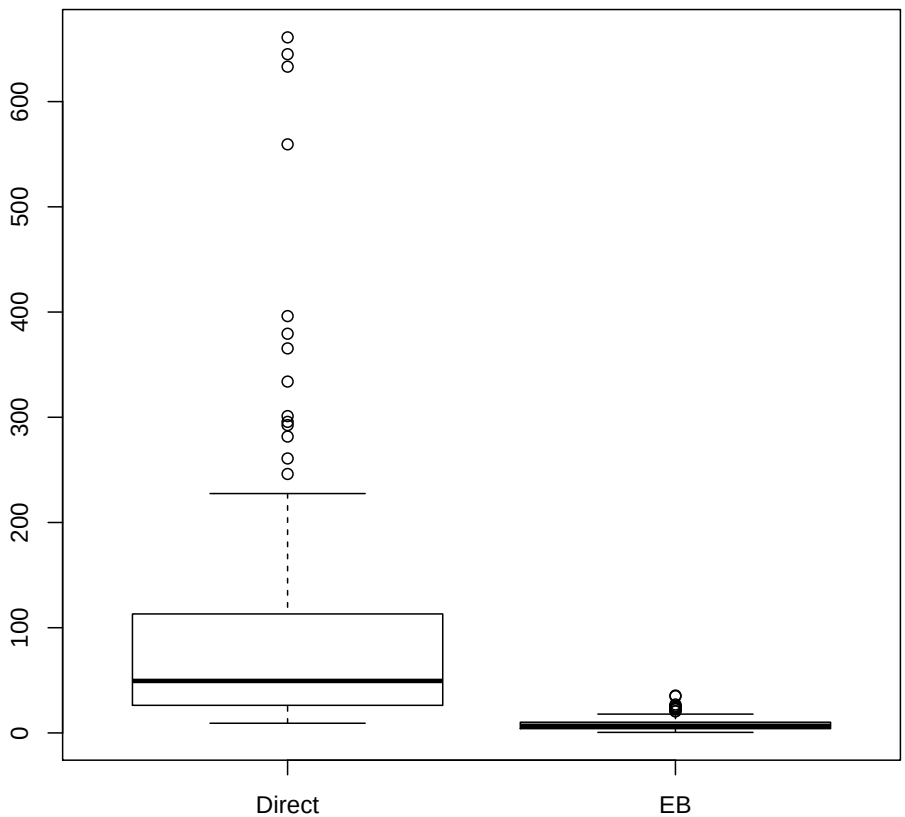




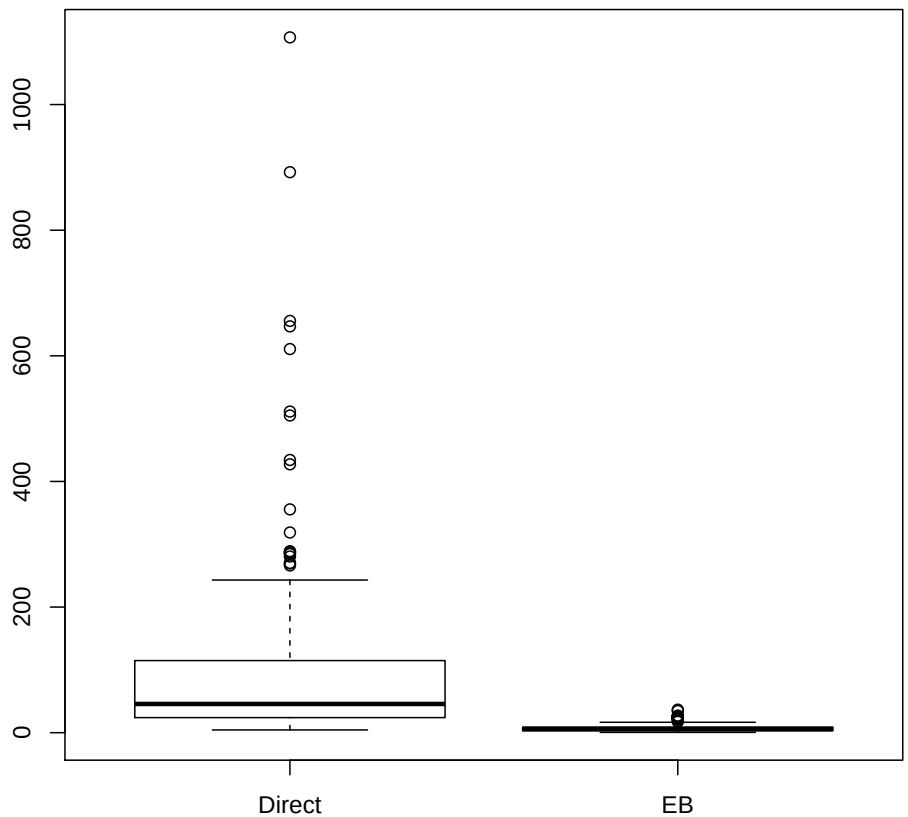
QUALITY EB vs. DIRECT: POV. RATE

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

MSE: Women

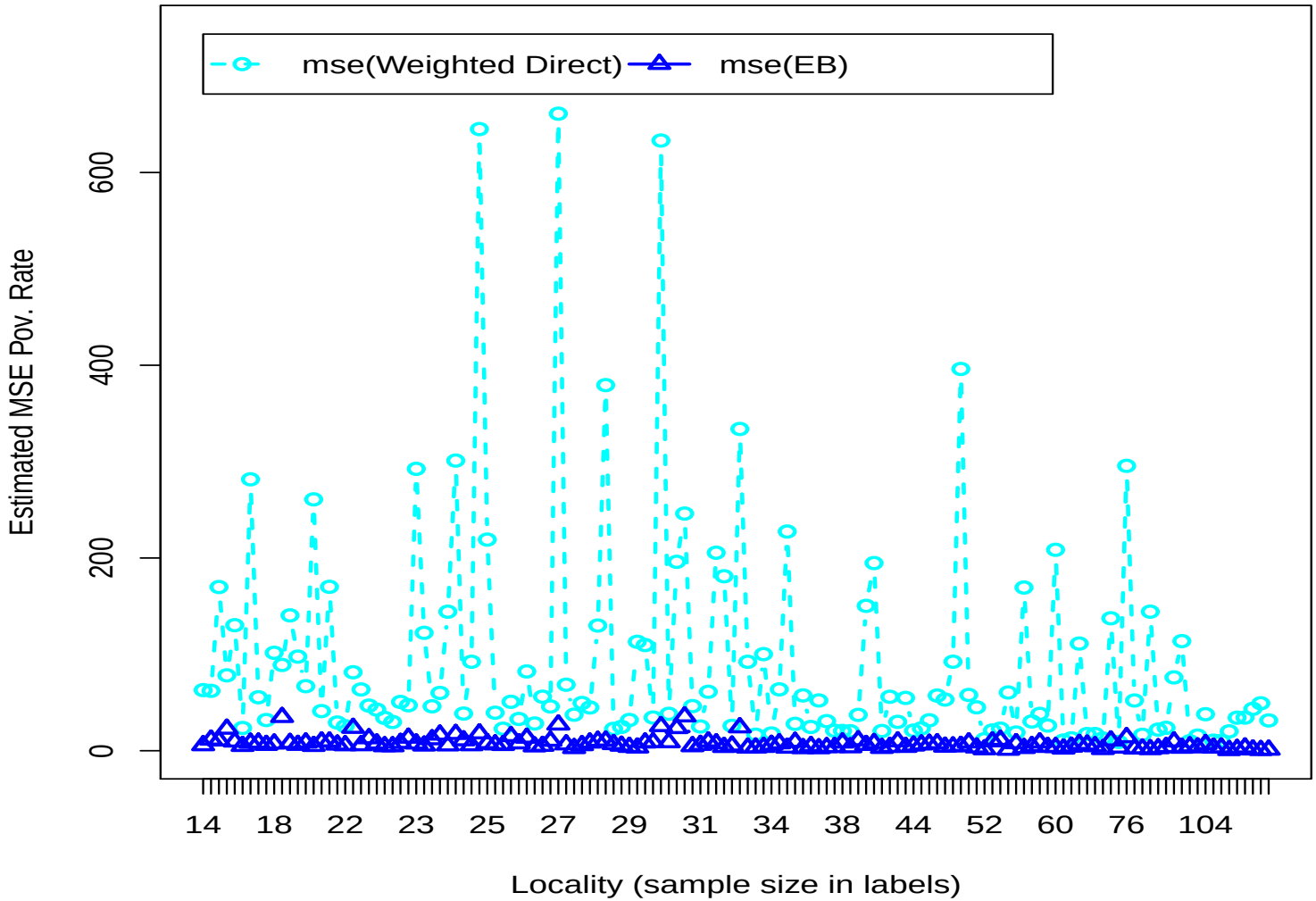


MSE: Men

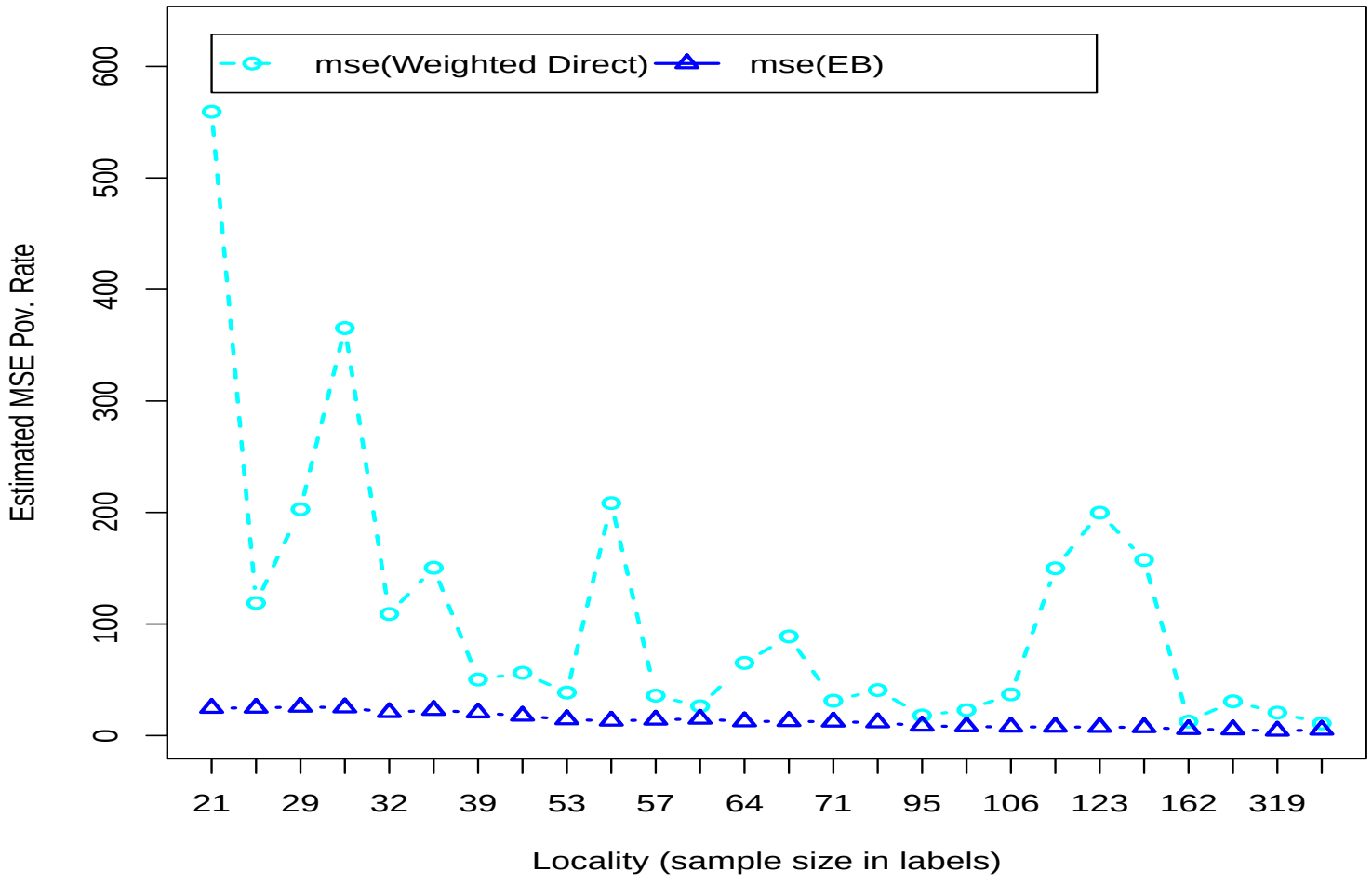


MSE EB vs. DIRECT POV. RATE: WEST BANK

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



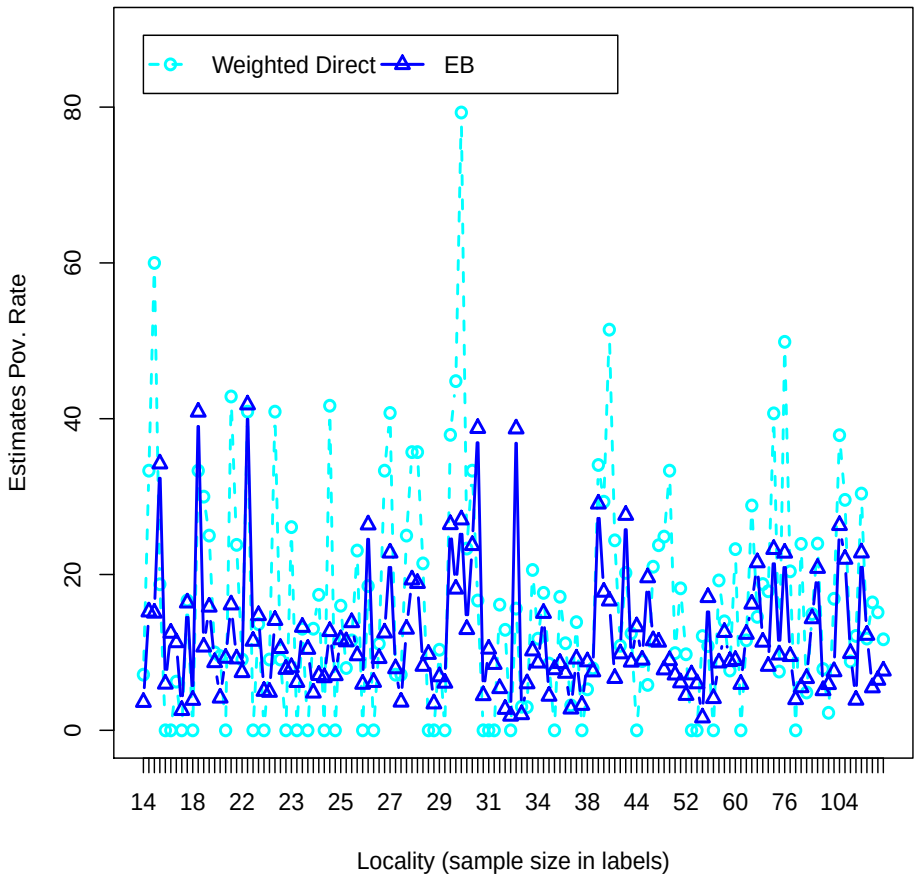
MSE EB vs. DIRECT POV. RATE: GAZA



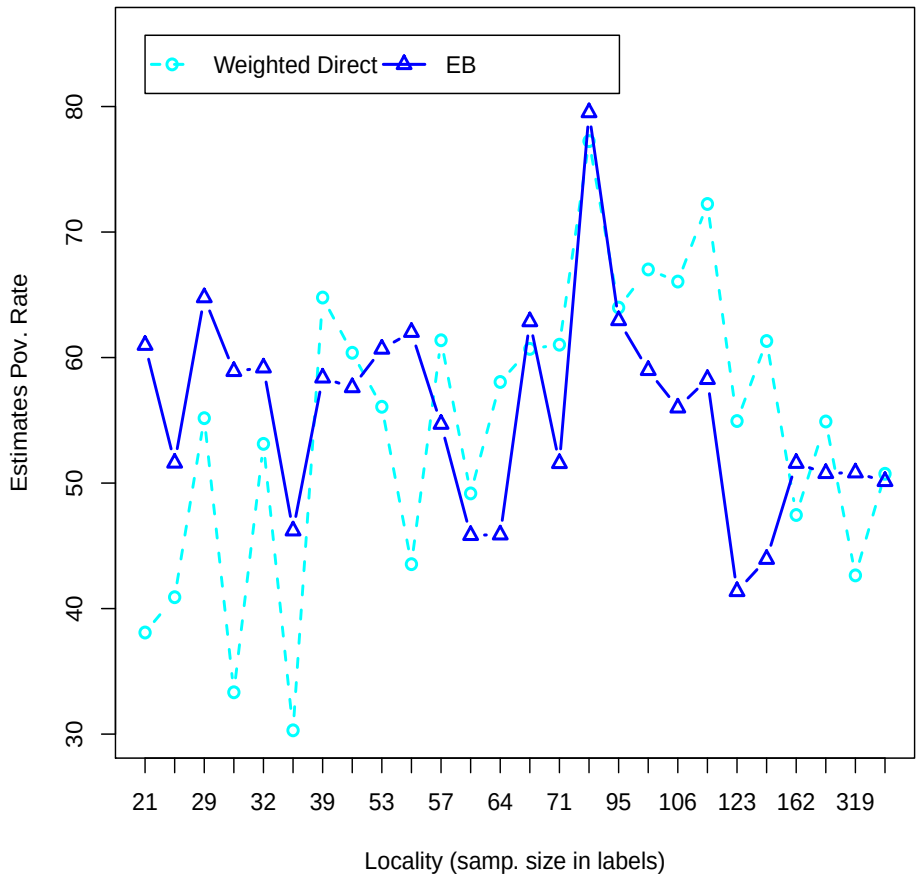
✓ **Great gains** also for Pov. **Gap** (not shown)!

ESTIMATED POV. RATE: WOMEN

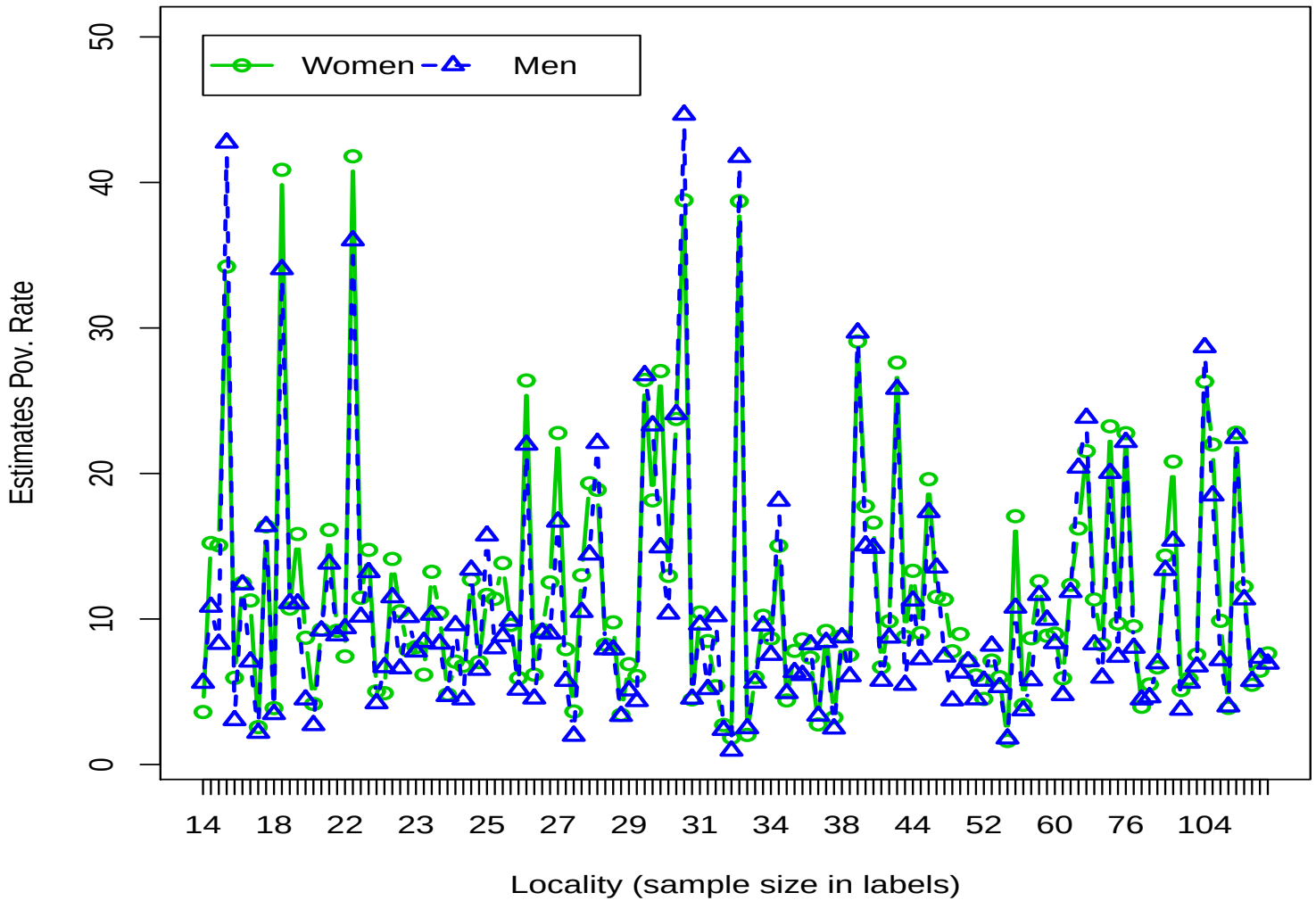
West Bank



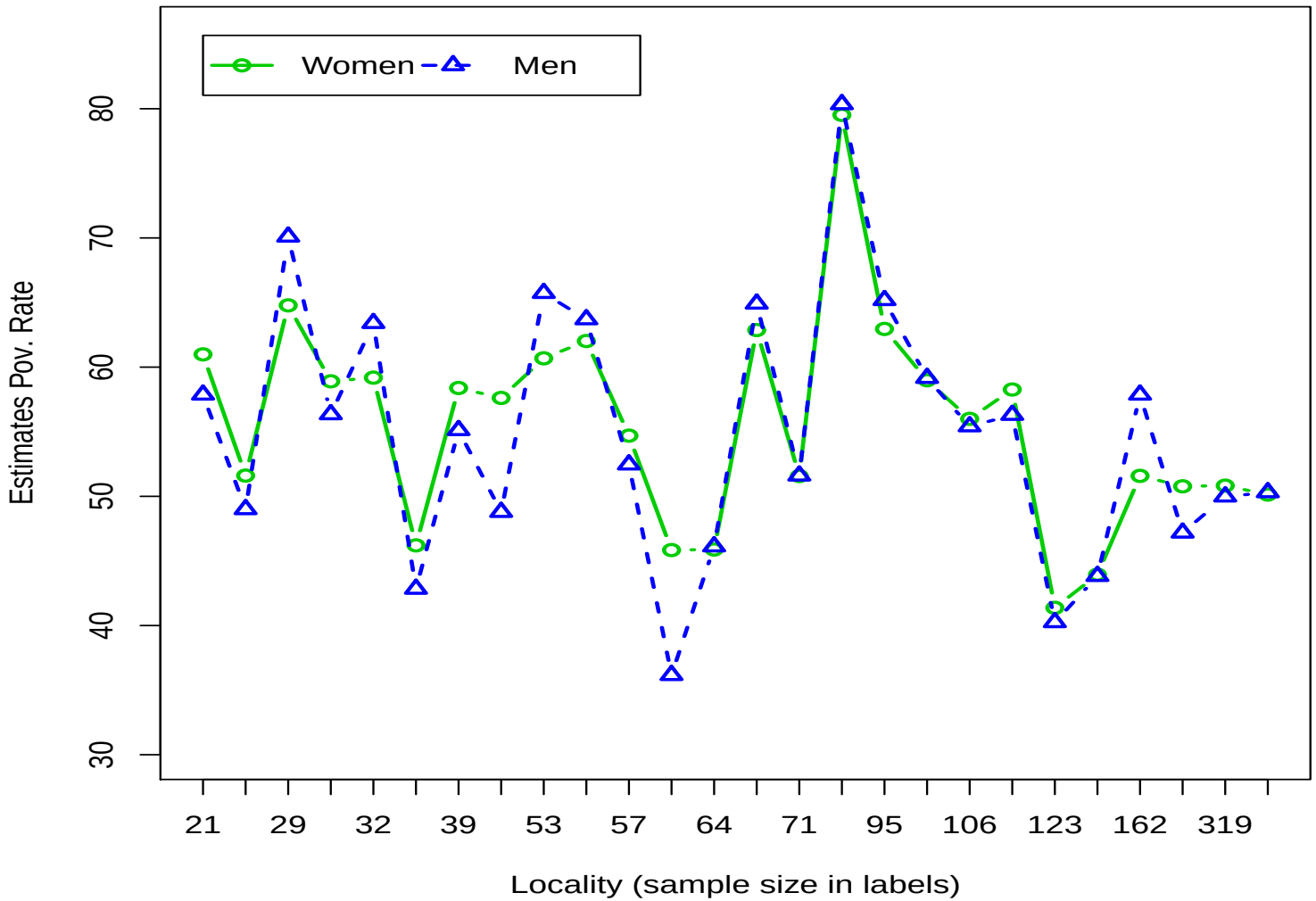
Gaza



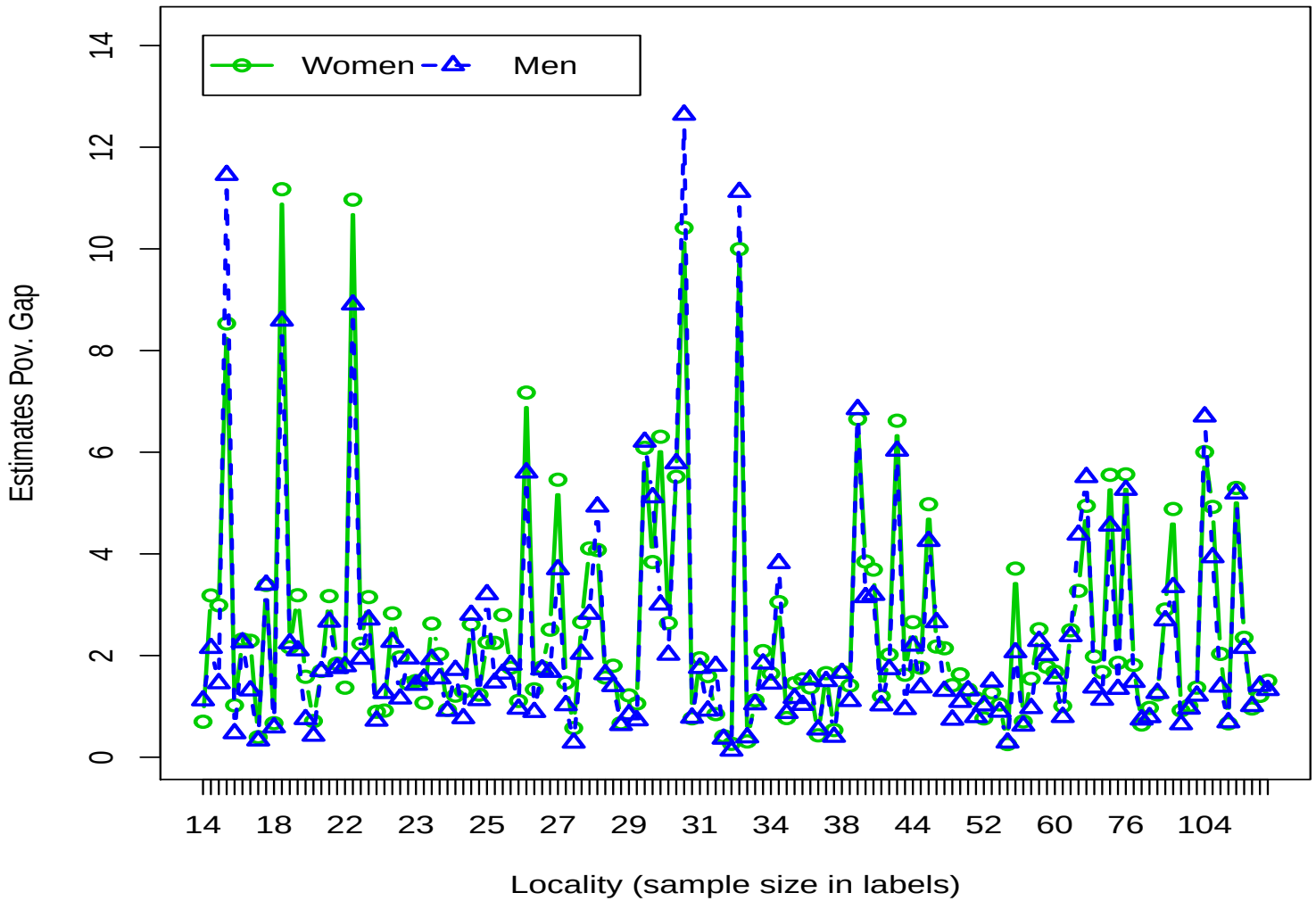
EB POV. RATE BY GENDER: WEST BANK



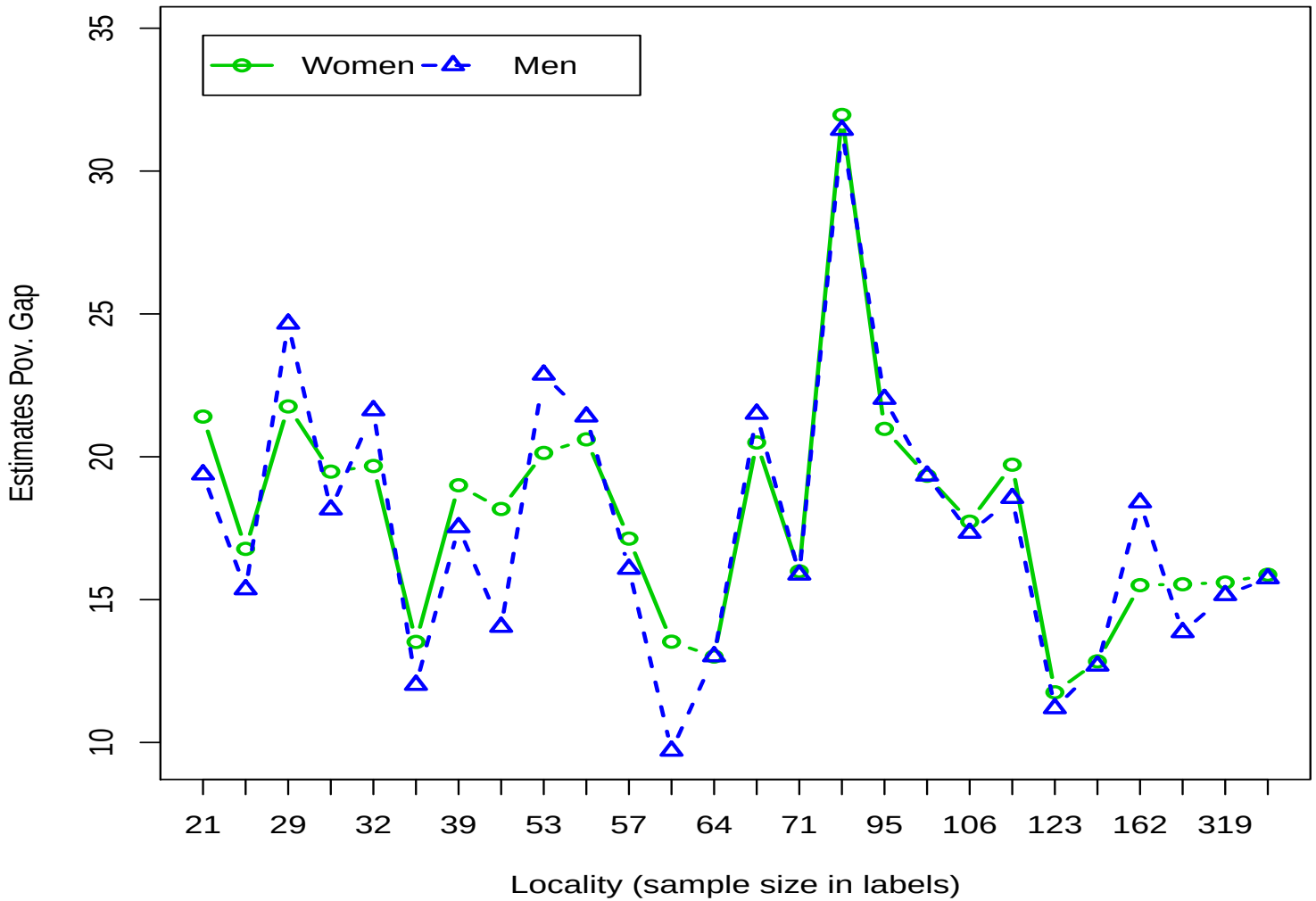
EB POV. RATE BY GENDER: GAZA



EB POV. GAP BY GENDER: WEST BANK



EB POV. GAP BY GENDER: GAZA



CONCLUSIONS

- The model **fits rather well** these data.
- We consider a model for the **individuals** instead of **households**. This allows estimation by gender.
- The unit level model allows to disaggregate estimates **at any desired level**. It allows to estimate **whatever indicator** that is function of expenditure.
- **Great efficiency gains** of EB with respect to direct estimators (over 82 % reduction in MSE for pov. rates and gaps).

CONCLUSIONS

- **Direct** estimates equal to **zero** for many localities (32 for Men, 29 for Women) and **highly unstable**.
- **EB** estimates **never zero** and much more stable. Perhaps some underestimation in few localities, model variations can be further explored.
- Gaza has **much larger** pov. rates and gaps. Perhaps using a different pov. line.
- No great differences between men and women, although women with slightly greater estimates for about 70 % of localities in West Bank.

- ✓ MANY THANKS TO UN-ESCWA AND PCBS FOR GREAT DATA PREPARATION!
- ✓ THANK YOU ALL FOR YOUR ATTENTION!