Small Area Estimation for Poverty Mapping

Nov. 10 – 14 2025

Training prepared for UBOS

Haoyu Wu (hwu4@worldbank.org)
Kristina Noelle Vaughan (kvaughan@worldbank.org)

Course Outline

Day 1 – Introduction of SAE and Unit-level Model

- Introduction of SAE
- Unit-level model: ELL, EB, CensusEB

Day 2 – Area-level Model

- Fay–Herriot model

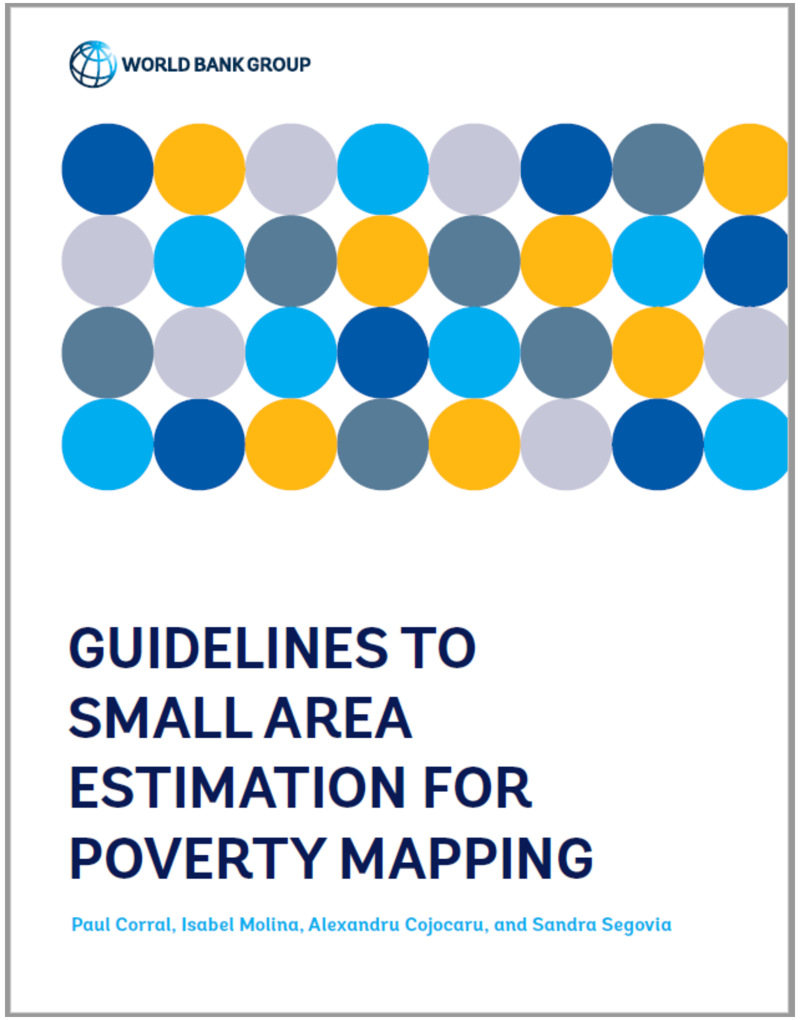
Day 3 – Poverty Mapping in Off-Census Years

- Unit-context model
- Machine learning model

Day 4 – Practical Application

- Stata demo with sample data: MEX and BGD
- Uganda case: new survey + 10% census

The following materials have been produced under guidance and quality control from world bank poverty team



Reference material and training

1. Guidelines online: https://ssegoviajuarez.github.io/sae/00_welcome.html
2. Guidelines pdf: https://openknowledge.worldbank.org/handle/10986/37728

Statistical Software

1. Stata
2. R

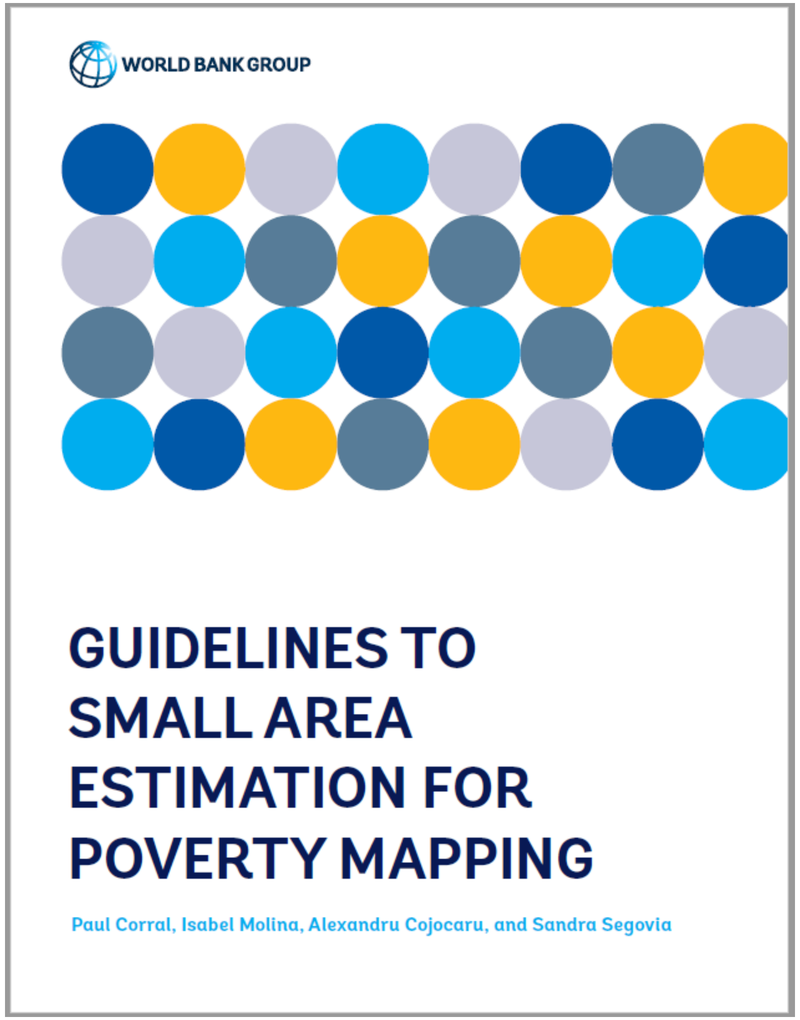
Stata package for SAE

1. Unit-level (EB and ELL): https://github.com/pcorralrodas/SAE-Stata-Package
2. Area-level (Fay-Herriot): https://github.com/jpazvd/fhsae

The guidelines also assist with the suggested quality control process for poverty maps

1. While doing the maps we are always available for consultations and advice
2. Once done, it's recommended teams submit their report and codes for review
    - Since methods have changed, it's important we get the methodology sections right, and that models and assumptions are fully checked
3. There is a set of people within the Poverty GP who can serve as reviewers

4. Deviations from recommendations require proper justification and validation
    - All caveats noted for the methodology should be stated



Small Area Estimation

Motivation and Basic Concepts

WORLD BANK GROUP

What are Small Area Estimation?

- Household surveys are not designed to produce reliable poverty estimates for small geographic domains (e.g., districts, sub-counties).

- Direct survey estimates suffer from:
  - small sample sizes,
  - large standard errors,
  - unstable poverty rates across small areas.

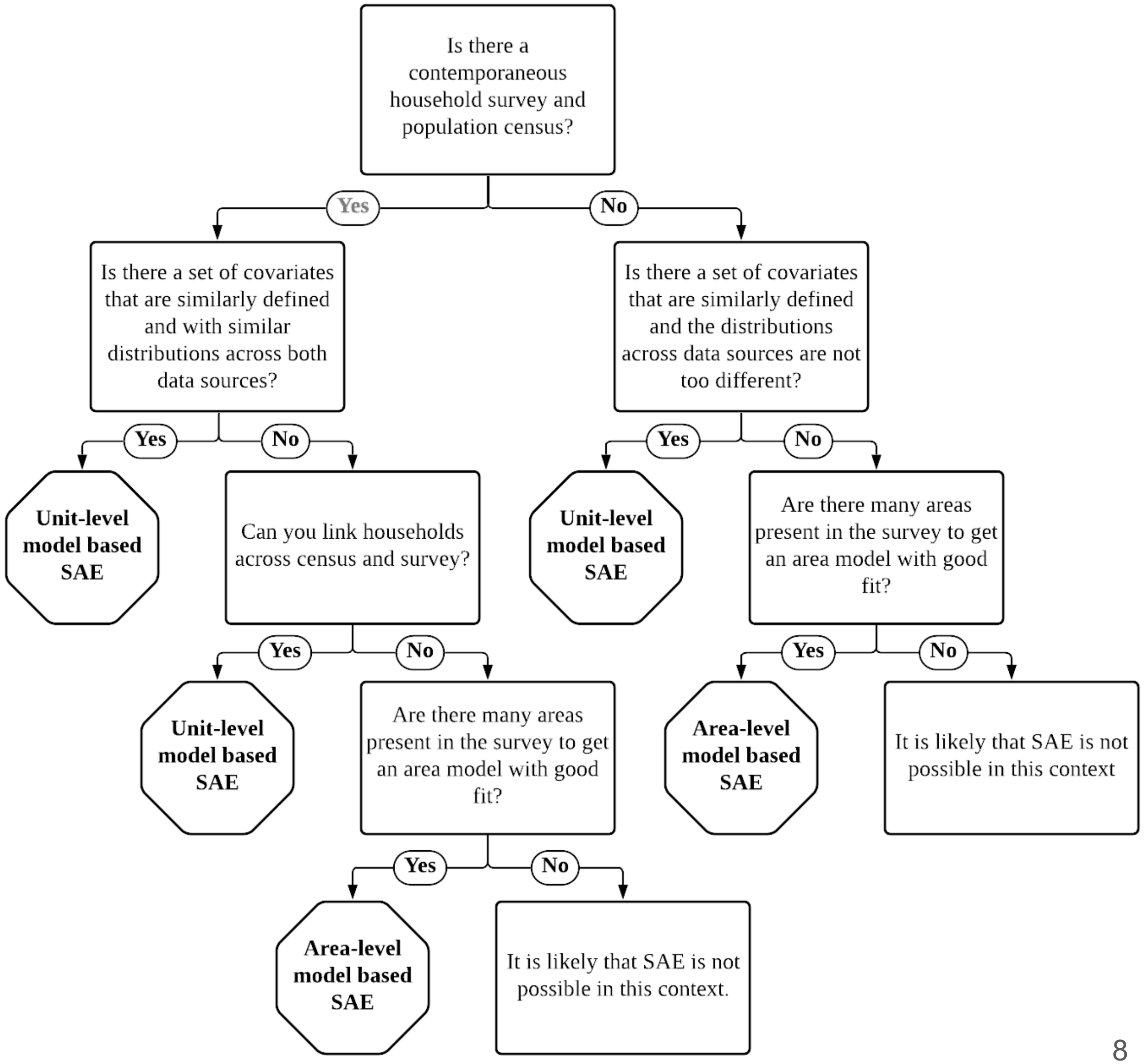
- Small Area Estimation (SAE) provides model-based estimates that "borrow strength":
  - across areas,
  - across households,
  - across predictors available in both survey and census.

Two Families of Small Area Estimation Methods

1. **Unit-Level Models (used in poverty mapping)**
    - Model built at the household level
    - Requires survey–census common variables
    - Estimator: ELL, EB, CensusEB
    - Output: welfare distribution → FGT poverty rates

2. **Area-Level Models (used when microdata in censsu are unavailable)**
    - Model at area level (means, rates, aggregates)
    - Requires published direct estimates + area covariates
    - Estimator: Fay–Herriot (FH)
    - Output: smoothed area means, proportions

The guidelines present a simple decision tree on what may be the recommended approach for your data environment…as we learn more about other methods, we hope to expand it



Small Area Estimation

Unit level models

WORLD BANK GROUP

What You Will Learn for Day 1?

- ✅ Understand the rationale for small area estimation (why SAE)
- ✅ Explain the logic and assumptions behind **unit-level models**
- ✅ Distinguish between **ELL, EB, and CensusEB**
- ✅ Perform model estimation and poverty simulation in Stata
- ✅ Apply the method to **real word data**

Unit-level Model Introduction

1. Monetary poverty (FGT0) is a non-linear parameter
2. Initial work at the WB was done by Hentschel, Lanjouw, Lanjouw, and Poggi (1998) but ignored random location effects – we'll se later why that matters
   - The welfare distribution is simulated in the census data based on parameters estimated using household survey data which has detailed information on welfare
   - This yields a welfare value for every single household in the census
   - Simulating the full welfare distribution allows for estimates beyond just poverty and mean welfare for a given area.
3. Elbers, Lanjouw and Lanjouw (2003) improved on the method
4. It was the go-to method for poverty maps done by the World Bank or with the World Bank's assistance until ~2015

Criticism and the publication from Molina and Rao's EB method led the institution to update methods

- In a review of World Bank research, the ELL method was scrutinized by Banerjee et al. (2008)
  - The reviewers were concerned that the method may not accurately present the true precision of the method's estimates – among other things
- Haslett (2010) recommended adjusting the fitting method described in ELL (2002)
  - Indicated it leads to an asymmetric variance-covariance matrix
- Molina and Rao (2010) showed in simulation studies that ELL is noisy and underperforms under the model's assumptions
  - Showed that EB methods are considerably superior and yield substantially less noisy estimates
- A shift to Molina and Rao's EB methods by the World Bank and its toolkit is detailed in Corral, Molina and Nguyen (2020)

The assumed model used for unit-level SAE

The nested error model used for unit level small area estimation comes from Battese, Harter and Fuller (1988):

$$y_{ch} = x_{ch}\beta + \eta_c + e_{ch}; \quad h = 1, \dots, N_c; c = 1, \dots, C$$

- C is the number of locations, N_c is the number of observations in location c
- β show how each variable relates to welfare.
- η_c is the random area effect
- e_{ch} is the idiosyncratic error
- We assume $\eta_c \sim N(0, \sigma_\eta^2)$ and $e_{ch} \sim N(0, \sigma_e^2)$

What is Necessary in order to do a Unit-level Poverty Map?

1. Household survey and Census must have variables in common between them
   - Questions should be defined in a similar manner in both data sources
   - Variables should have similar distributions

2. Common variables should be sufficiently correlated with the welfare measure of interest (income or consumption)

3. Additionally, we need a location variable in order to link the census and survey at that level

The main steps

Step 1: Prepare the data in the household survey and the census

- Check definitions, harmonize variables, reweight if needed
- Ensure that data is comparable (compare distributions, remove outliers)

Step 2: From the survey, we estimate a model for household consumption:

$$\boldsymbol{y_{ch}^{survey} = x_{ch}^{survey}\beta + \eta_c + e_{ch}}$$

Predict $\boldsymbol{\hat{\beta}}$ and the variance for the two error terms: $\boldsymbol{\hat{\sigma}_\eta^2}$ and $\boldsymbol{\hat{\sigma}_e^2}$

Step 3: Apply the model to the census with $\boldsymbol{x_{ch}^{census}}$(the observed characteristics) to predict y ($\boldsymbol{\hat{y}_{ch}^{census}}$).

$$\boldsymbol{\hat{y}_{ch}^{census} = x_{ch}^{census}\hat{\beta} + \hat{\eta}_c^{census} + \hat{e}_{ch}^{census}}$$

However, we don't know the exact values of $\boldsymbol{\hat{\eta}_c}$ and $\boldsymbol{\hat{e}_{ch}}$ as they are random and different from survey model

The main steps (cont.)

Step 3: To account for randomness in the model (i.e. not all the variation in y will be explained by the x variables), we simulate the errors by drawing random values:

$$\hat{\boldsymbol{\eta}}_{\boldsymbol{c}} \sim \boldsymbol{N}\left(\mathbf{0}, \hat{\boldsymbol{\sigma}}_{\boldsymbol{\eta}}^{\mathbf{2}}\right) \qquad \hat{\boldsymbol{e}}_{\boldsymbol{ch}} \sim \boldsymbol{N}\left(\mathbf{0}, \hat{\boldsymbol{\sigma}}_{\boldsymbol{e}}^{\mathbf{2}}\right)$$

Step 4: Because $\hat{\boldsymbol{\eta}}_{\boldsymbol{c}}$ and $\hat{\boldsymbol{e}}_{\boldsymbol{ch}}$ are random and unknown, we repeat the simulation many times because the census lacks the random components of the model, and Monte Carlo draws are necessary to recover the welfare distribution and quantify uncertainty.

Step 5: averaging across simulations, we get:

- A **mean estimate** of welfare
- A **standard error** (measuring uncertainty).

Understanding the Model Outputs Before Evaluation

- After estimating the GLS (Henderson III) model, three key components are produced:
  - $\boldsymbol{\beta}$ (beta coefficients): show how each variable relates to welfare.
  - $\boldsymbol{\sigma_\eta^2}$ (area-level variance): variation explained by clusters or communities.
  - $\boldsymbol{\sigma_e^2}$ (household-level variance): variation across households within clusters.

- These components help explain how much of the welfare differences are due to location effects versus household characteristics.

- Intraclass Correlation Coefficient (ICC): $\boldsymbol{\sigma_\eta^2/(\sigma_\eta^2+\sigma_e^2)}$. The share of total variation in welfare that comes from differences across areas (clusters, communities), rather than across households within areas.

WORLD BANK GROUP

How to judge my model?

1. **Check Model Fit**
   - Are β-coefficients plausible?
   - Signs and magnitudes consistent with economic intuition?
   - No obviously spurious or unstable estimates.

2. **Residual Diagnostics**
   - **Linearity:**
     - Residuals vs. fitted values → no strong patterns.
   - **Normality:**
     - Q-Q plots for household residuals ($\boldsymbol{e_{ch}}$) and area effects ($\boldsymbol{\eta_c}$).
     - Normal approximation is required for MC simulation.
   - **Homoscedasticity:**
     - Constant variance across fitted values and key covariates.
     - If not → consider heteroskedasticity modeling.
     - ICC: Too low → model not exploiting area effects.
       Too high → area effects dominate; check model specification

What to watch out for?

Multicollinearity

- A common perception is that multicollinearity is not a problem when all we care about is prediction, but we are not predicting
  - Produces imprecise coefficients
  - We are not predicting on to the same data, and thus out of sample predictions will be imprecise
    - Big problem if the patterns of the multicollinearity in the target data changes; this may lead to considerable problems in our results

Overfitting

- May occur when too many interactions or polynomials are included
  - This produces artificially low mean squared errors, which may lead to overly confident poverty estimates

Outliers

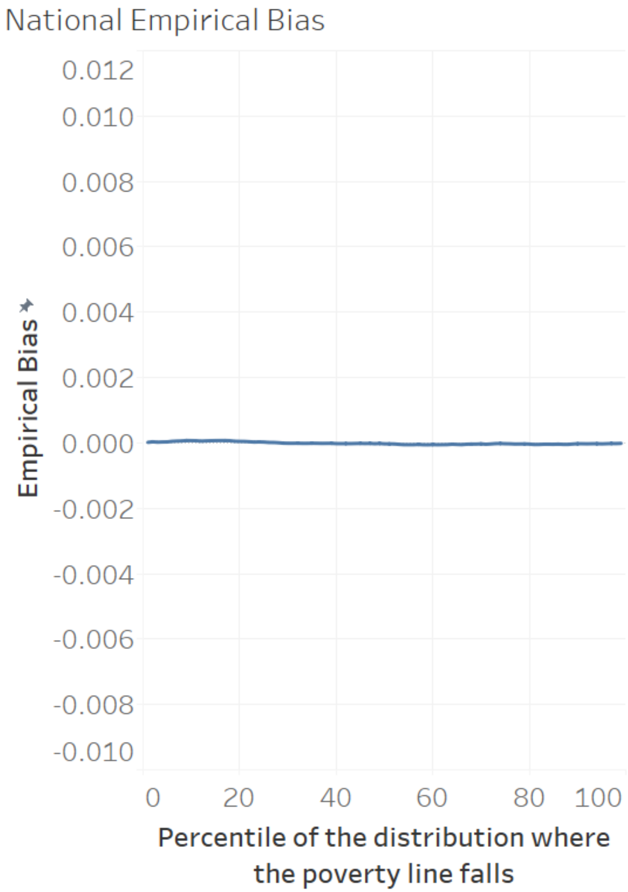
- Large residuals Dependent variable unusual given predictors
- Outliers in our covariates (points which have leverage on our coefficients)
  - Have impact on size of coefficients

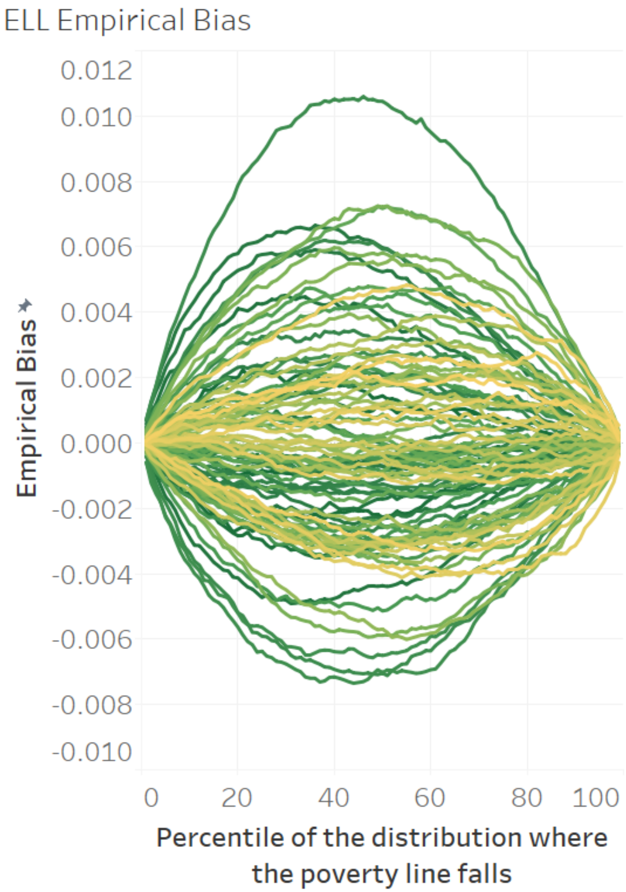
WORLD BANK GROUP

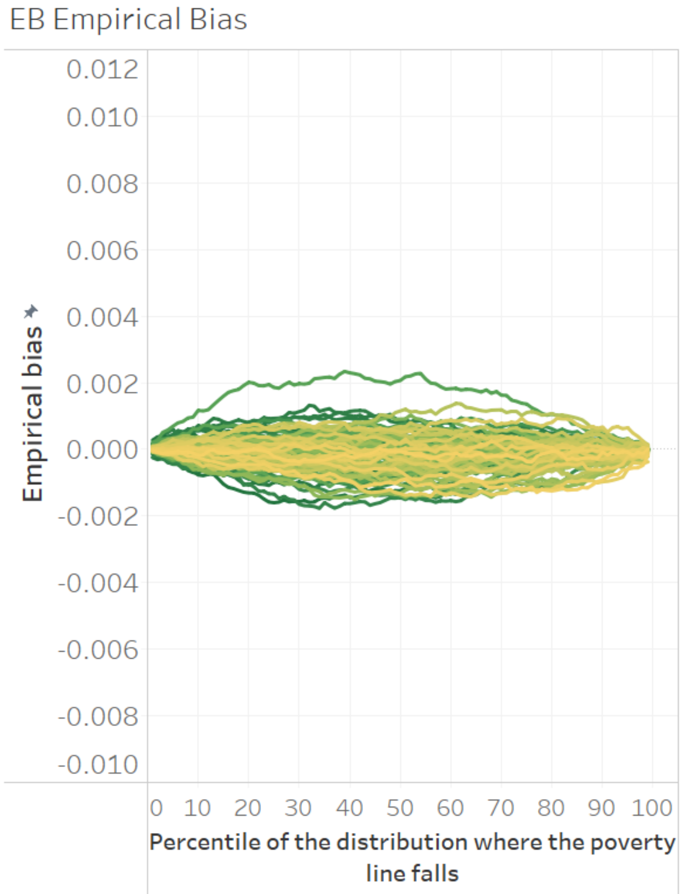
ELL vs EB vs Census EB

	ELL (2003)	EB (2010)	Census EB (2019)
Main Idea	Simulate entire census using model parameters repeatedly	Condition on observed survey data to predict census	Simulate census as in EB but without linking to survey
Predictor	$E\left(\tau_c\,;\hat{\theta}\right)$	$E\left(\tau_c \vert y_{c,S}\,;\hat{\theta}\right)$	$E\left(\tau_c\,;\hat{\theta}\right)$
Census welfare vector	$y_c = \left(y_{c,1}^T, \dots, y_{c,N_c}^T\right)^T$	$y_c = \left(y_{c,S}^T, y_{c,r}^T\right)^T$	$y_c = \left(y_{c,1}^T, \dots, y_{c,N_c}^T\right)^T$
Simulation	Bootstrap (resample parameters each time)	Monte Carlo (fixed parameters) + Parametric Bootstrap	Monte Carlo (fixed parameters) + Parametric Bootstrap
Sampling weight	No	No	Yes
Model fitting approaches	Generalized Least Squares (GLS)	restricted maximum likelihood (REML)	Generalized Least Squares (GLS)
Limitations	More simulation noise; ignores survey info	Needs matching between survey & census	Slightly less efficient but nearly identical results to EB when sampling small

How do EB methods improve upon ELL methods?







- **EB ensures predictions of the dependent variable at the area level are aligned to the survey** – best information at hand
- At the national level (left) the ELL model works well and predicts poverty accurately across the welfare distribution
- At the area level, because of the lack of EB, the method is less reliable and predictions for a given area show more bias than EB which leads to larger MSE

ELL and EB results from simulation where headcount poverty estimates are obtained under 99 different poverty lines covering the 99 percentiles of the simulated welfare distribution

Stata recap

Let's now go to Stata to get ready for SAE

Open following do-file:

~01.dofiles/1 - Stata_Refresher.do

STATA package introduction

WORLD BANK GROUP

SAE package – Model options

1. **One-fold Nested Error Model with GLS (ELL)** (Incorporate weights or heteroskedasticity)
- Used to be the standard method, but replaced by EB.

2. **One-fold Nested Error Model with REML (EB)** (Does not incorporate weights or heteroskedasticity)
- Survey weights not crucial
- Homoskedasticity assumed
- Estimates needed at the single level

3. Two-fold Nested Error Model (EB two-fold)
- Point estimates needed at two levels:

$v_{ch} \sim N(0, \sigma_v^2)$ =cluster effect

$$\boldsymbol{y_{chi} = x_{chi}\beta + \eta_c + v_{ch} + e_{chi}}$$

4. Henderson's Method III (H3) – CensusEB
- Survey weights are important
- Heteroskedasticity present
- Only one-fold nesting needed

WORLD BANK GROUP

Best Practices 1

Model Selection

1. Consider data transformation to achieve normality (transformation)
2. Remove non-significant covariates sequentially (LASSO)
3. Check for and remove high multicollinearity (VIF > 5)
4. Validate model assumptions and diagnostics

Random Effects

- Specify random effects at the same level as the desired estimates
- Avoid estimating at higher levels than random effects
- Use two-fold nested models when estimates are needed at multiple levels

Best Practices 2

Data Requirements

1. Contemporaneous survey and census recommended
2. Variables must be measured consistently between sources
3. Compare variable distributions across sources
4. Verify survey weights can be replicated in census

Alpha Model for Heteroskedasticity

- Consider when heteroskedasticity present
- Can improve estimates despite typically low R^2
- Available with H3 estimation method
- Test residuals to determine if needed

Best Practices 3

Validation Steps

1. Check model diagnostics and assumptions
2. Validate normality of residuals and random effects
3. Examine outliers and influential observations
4. Compare estimates to direct estimators where possible
5. Assess precision through MSE estimates

MSE Estimation

- Use parametric bootstrap for Census EB
- Bootstrap procedure more computationally intensive but more accurate
- Consider minimum 100 Monte Carlo simulations
- Recommended 200+ bootstrap replications for MSE

Let's now go to Stata to see the assumed model…
Open following do-file:
~01.dofiles/1.assumed_model.do

WORLD BANK GROUP

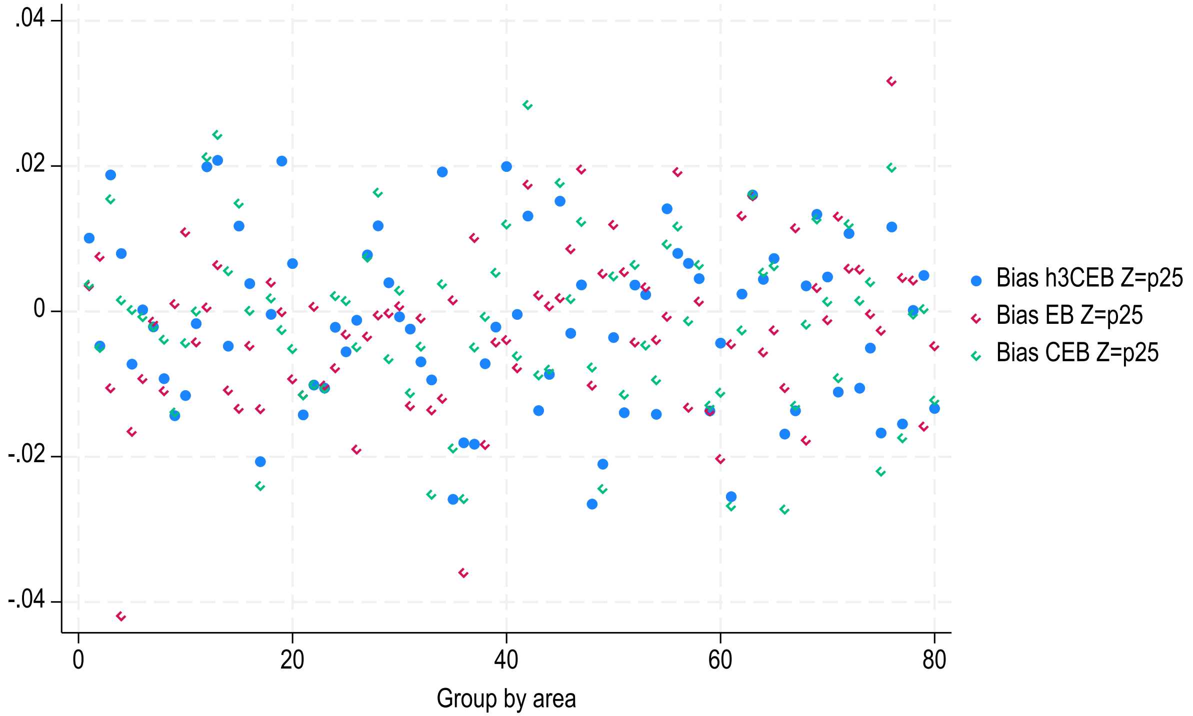
Let's now go to Stata and see how to validate a method through model-based simulations

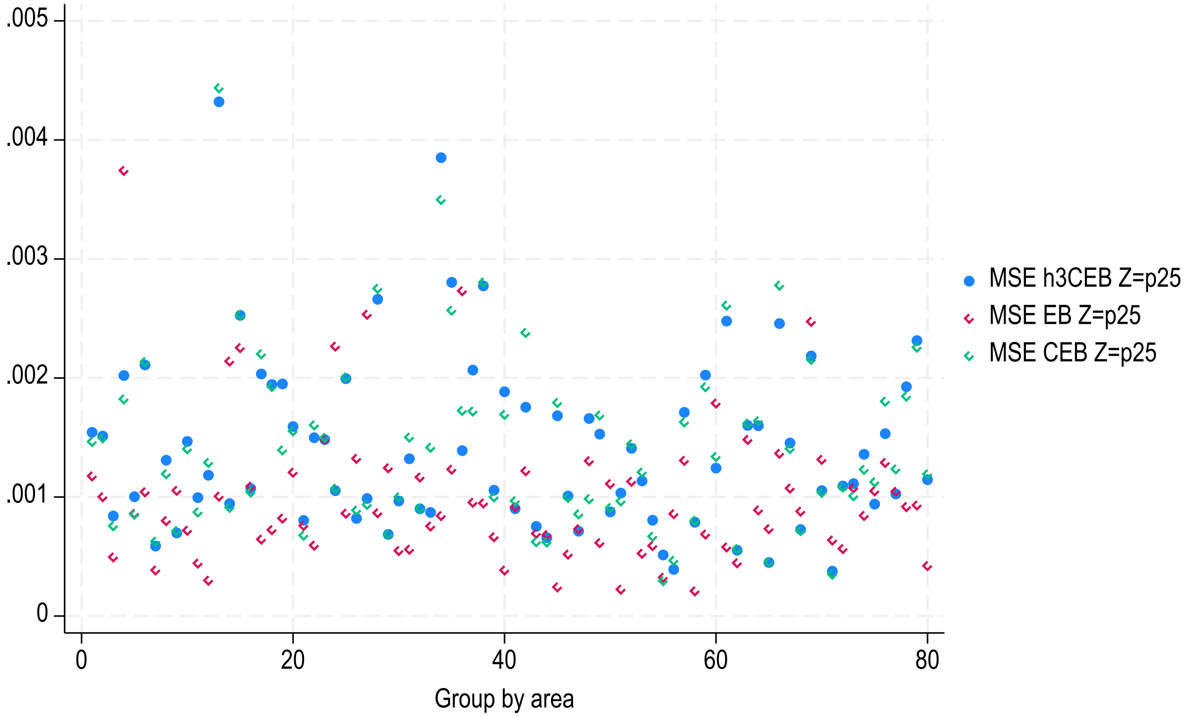
Open the following do-file:

~01.dofiles/3.ModelBasedSim.do

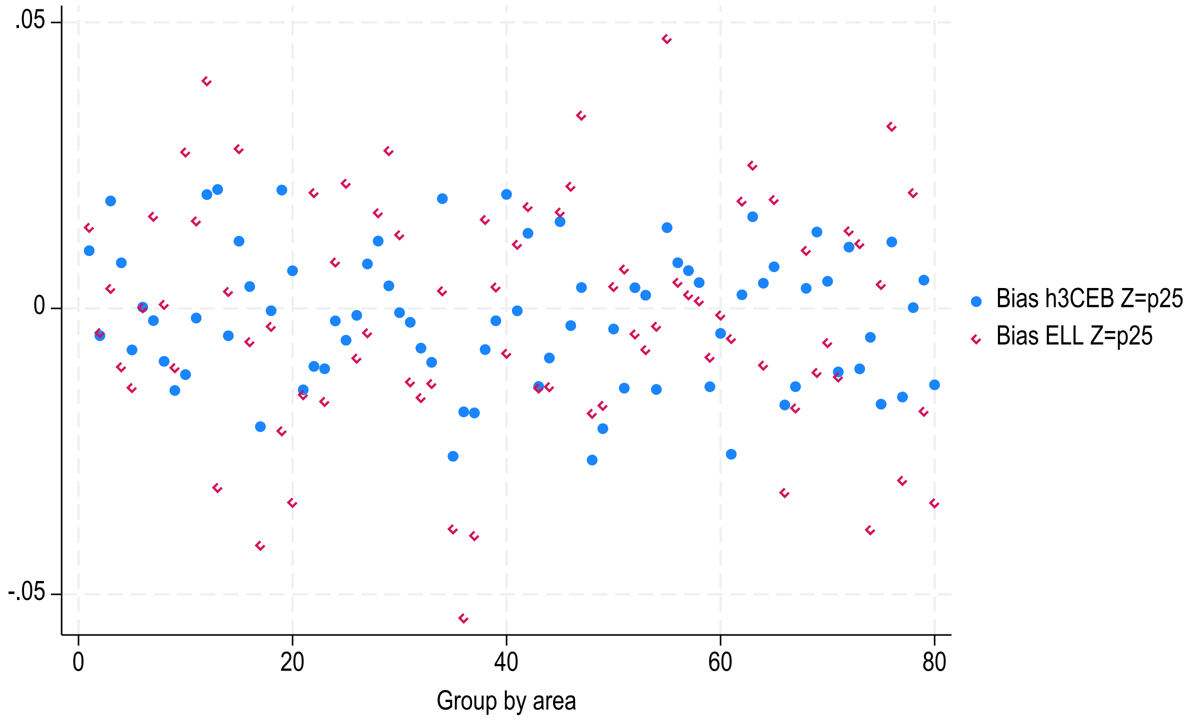
WORLD BANK GROUP

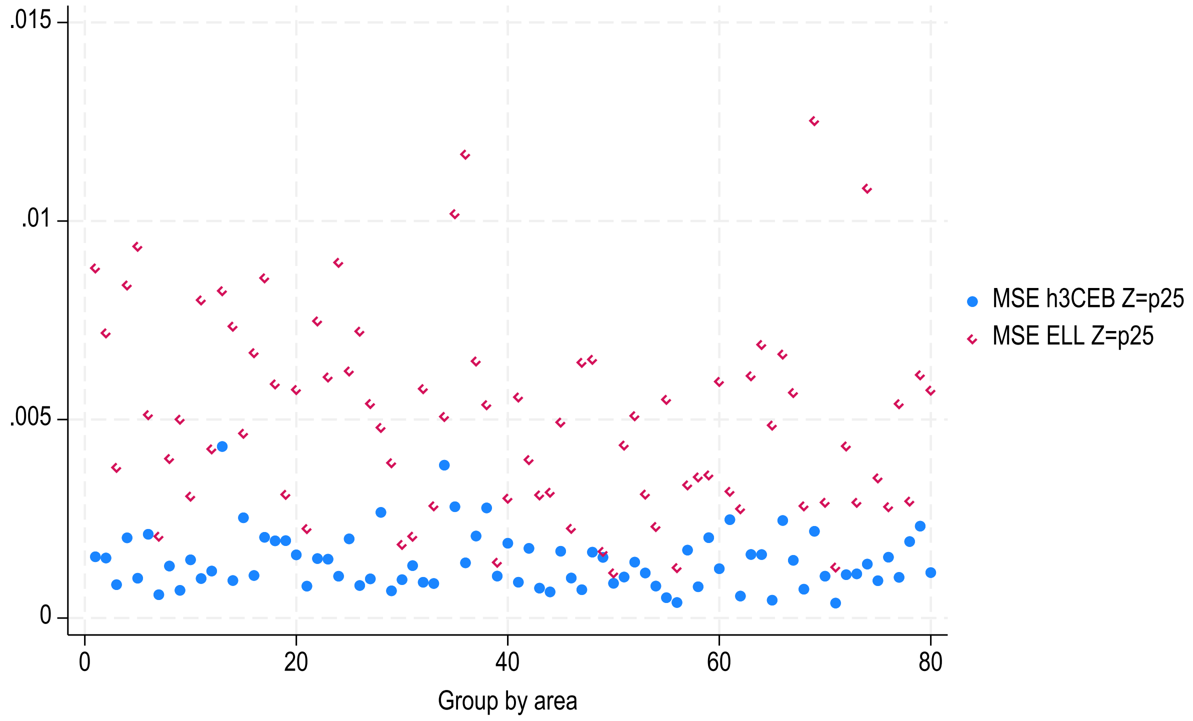
H3CEB vs EB vs CEB





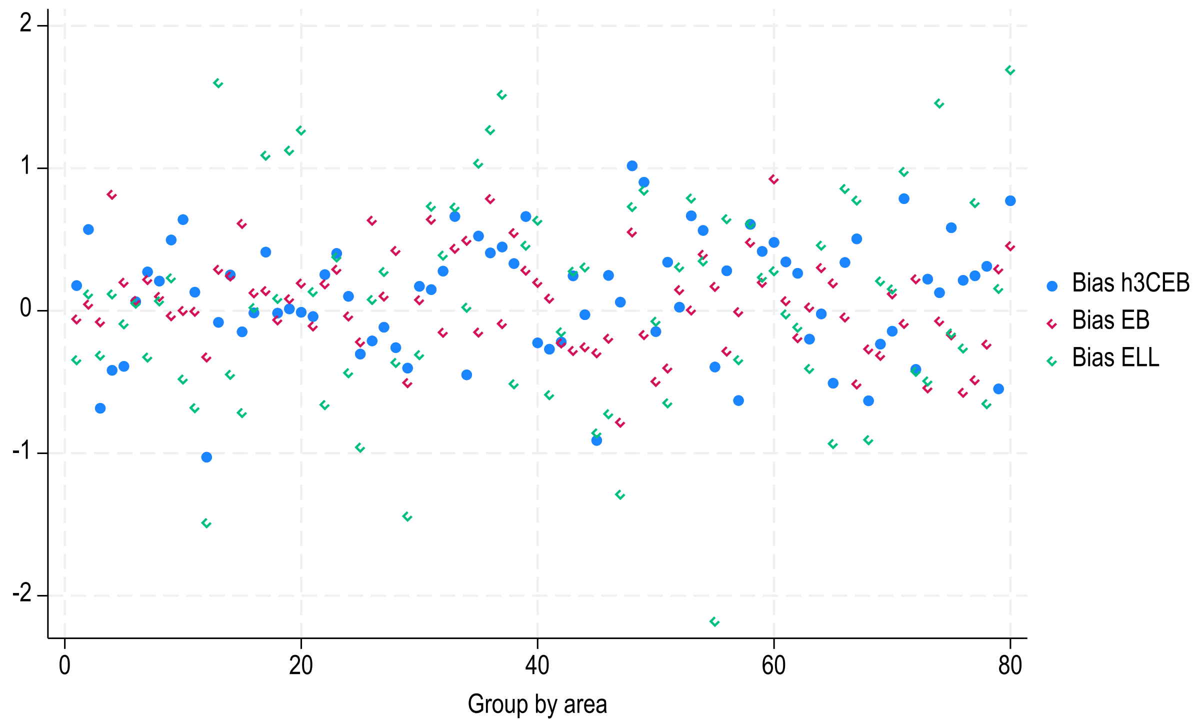
H3CEB vs ELL

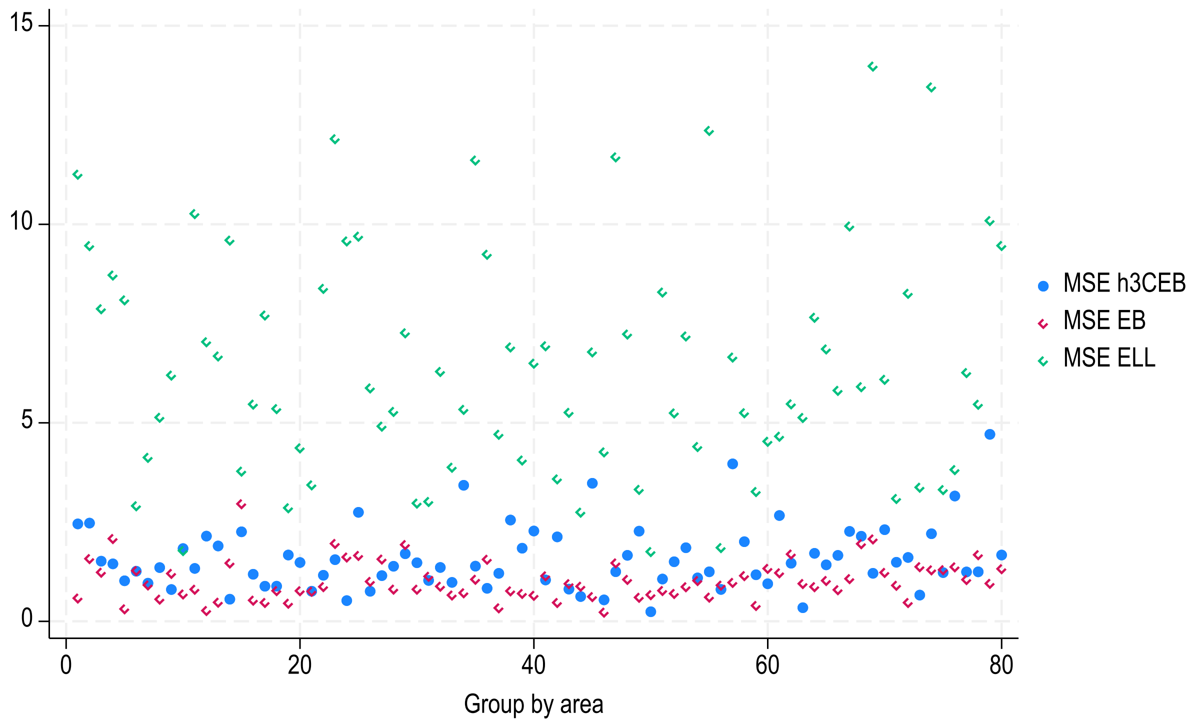




WORLD BANK GROUP

H3CEB vs EB vs ELL





Real-world example
Mexico

Mexico Data

Data Overview

- **Data source:** Mexican Intra-Censal Survey 2015 (*Encuesta Intracensal*), conducted by **INEGI** (Instituto Nacional de Estadística y Geografía).
- **Sample size:** 5.9 million households.
- **Coverage:** Nationally representative at multiple levels:
    - 32 states
    - 2,457 municipalities or delegations
- Localities with ≥50,000 inhabitants.

Adjustments for the SAE Application

- To build a realistic "census" for simulation:
    - Municipalities with fewer than 500 households were dropped.
    - 90% of zero-income households were removed to correct income distribution bias.
- The resulting dataset contains: 3.9 million households 1,865 municipalities (after filtering).

SAE Model Development Workflow

1️⃣ **Variable Preselection** – Use **LASSO regression** to identify key predictors of household welfare.
2️⃣ **Preliminary SAE (H3-GLS)** – Estimate initial fixed and random effects.
3️⃣ **Model Simplification** – Apply **stepwise backward elimination** to drop weak predictors.
4️⃣ **Refined SAE Model** – Re-estimate model after variable reduction.
5️⃣ **Multicollinearity Check** – Remove variables with **VIF > 5** to ensure stability.
6️⃣ **Post-VIF SAE Model** – Refit model after addressing collinearity.
7️⃣ **Influence Diagnostics** – Flag **outliers, high-leverage, and influential** points using residuals, leverage, and Cook's D.
8️⃣ **Alpha (Heteroskedasticity) Model** – Model residual variance via selected covariates using LASSO and VIF screening.
9️⃣ **Final SAE Model** – Fit **H3-GLS with α-model**, excluding flagged cases.
🔟 **Save Outputs** – Store final variable lists and flags as **mysvy.dta** for simulation.

Data Diagnostics - Normality

Transformation

log-shift transformation (lnskew) to make your welfare variable approximately normal.
The small skewness (≈ 0.0001) means the transformation successfully removed asymmetry — which is important for valid GLS assumptions.

Model structure

The estimation has **three parts**:

Stage	Model	Purpose
OLS model	Regresses transformed welfare on covariates	Gives baseline β coefficients
Alpha model	Regresses log(residual variance) on covariates	Captures heteroskedasticity (α-model)
GLS model	Re-estimates β using variance structure from α	Final Henderson III estimates

Data Diagnostics - outliers or influential observations

- Helps detect **outliers** (cases with unusually large residuals) and **influential observations** (cases that disproportionately affect regression coefficients).
- Ensures the **stability, robustness, and credibility** of regression and SAE model estimates.

Metric	Definition	Rule of Thumb
Outlier	An observation with a large residual — the actual value differs substantially from the predicted value.	|Studentized residual| > **2** (moderate), > **2.5–3** (severe)
Leverage	Measures how far a predictor value is from the mean; identifies points with extreme predictor combinations.	Leverage > **(2K + 2) / N** (where K = number of predictors, N = sample size)
Cook's Distance	Quantifies overall influence of each case by combining leverage and residual.	**D > 4 / N** → likely influential

WORLD BANK GROUP

Evaluating the Need for an α-Model in SAE

Test performed using alfatest(residual) with all independent variables *not included* in the β-model as variance predictors (zvar() option).

Diagnostic	Without α-Model	With α-Model (zvar = all other vars)
R-squared (α-model)	0.0000	**0.0187**
F-statistic (α-model)	< 1 (ns)	**6.59 (significant)**
Significant predictors	None	**7 variables**
Ratio σ_{ϱ}^2 / MSE	0.047	0.047 (stable)
Model fit (R^2 β-model)	0.489	0.489 (unchanged)

Interpretation

- The α-model reveals **significant heteroskedasticity** in residuals across municipalities.
- Variables such as **household size, education, and age composition** explain variance heterogeneity.
- Incorporating an α-model improves the **efficiency and robustness** of GLS estimation without altering mean predictions.

Understanding the Model parameters

OLS Model Diagnostics

- $R^2 = 0.48$:your covariates explain ~48% of welfare variation.
- All coefficients are **highly significant (p < 0.001)**.
- Signs make sense:
  Education (+), assets (+), household size (−), dependency (−).
- A good explanatory model — solid base for SAE simulation

Alpha Model Diagnostics

- $R^2 = 0.014$: small but expected; variance models rarely explain much.
- Positive coefficients (e.g. share_elderly, mun_hhsize) → areas with older or larger households tend to have higher variance.
- Negative coefficients (e.g. share_under15, max_secondary) → lower variance where education or younger population dominates.
- This step models **heteroskedasticity** across clusters (municipalities).
  It adjusts weights in the GLS estimation so small/unstable clusters don't dominate.

Understanding the Model parameters

GLS (Henderson III) Model

This is the **final stage** used in simulation.

Statistic	Value Meaning
$R^2 = 0.48$	good fit, same as OLS
Root MSE = 0.67	residual spread after modeling
$\sigma_\eta^2 = 0.025$	area-level (between-cluster) variance
$\sigma_\epsilon^2 = 0.425$	household-level (within-cluster) variance
$\rho = \dfrac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\epsilon^2} \approx 0.056$	ICC = 5.6% — modest spatial correlation

Meaning

- About 5–6% of welfare variation is due to **area effects** (clusters).
- SAE is justified — there's meaningful but not overwhelming heterogeneity across municipalities.
- All coefficients remain stable between OLS and GLS → your model is robust.

WORLD BANK GROUP

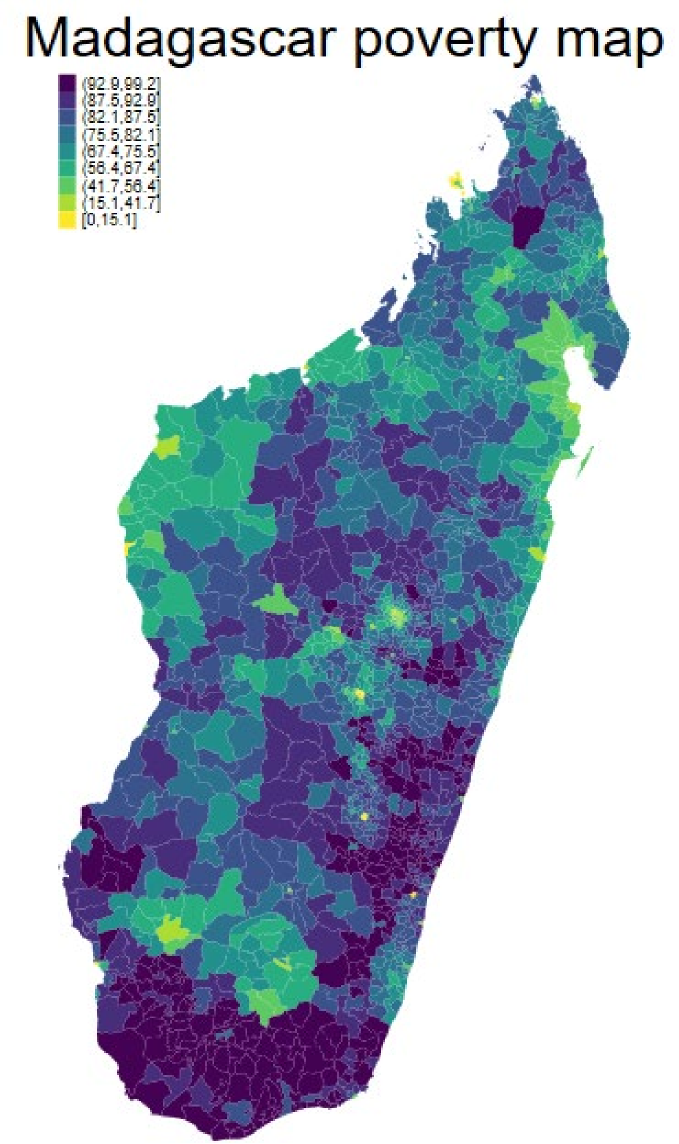
Figure generation

Shapefile data

- Google it

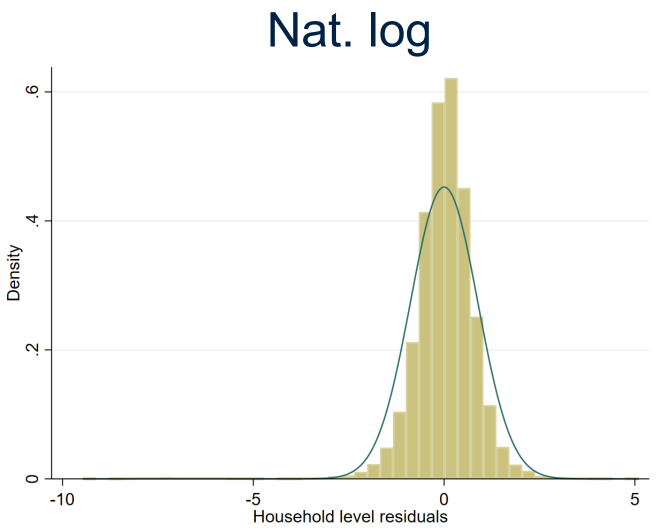
Stata code

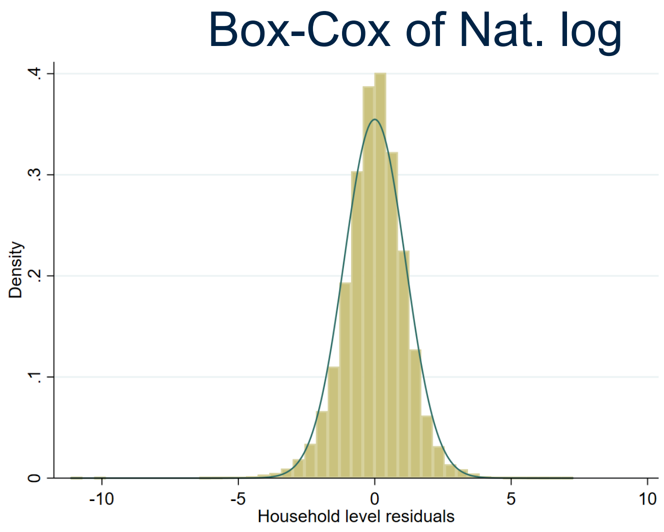
- Geoplot
- spmap

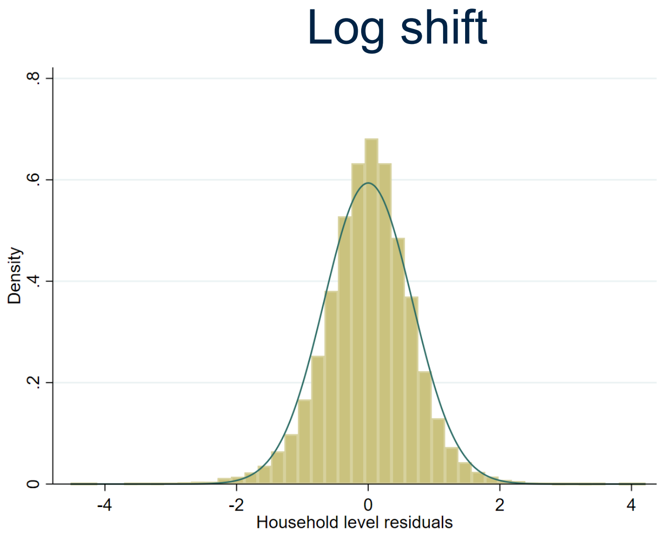


Appendix

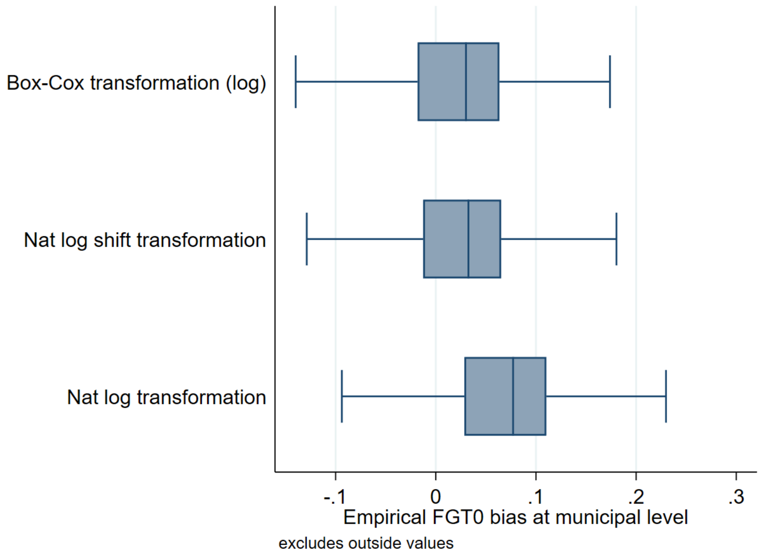
Transformation is necessary to achieve an approximately normal distribution



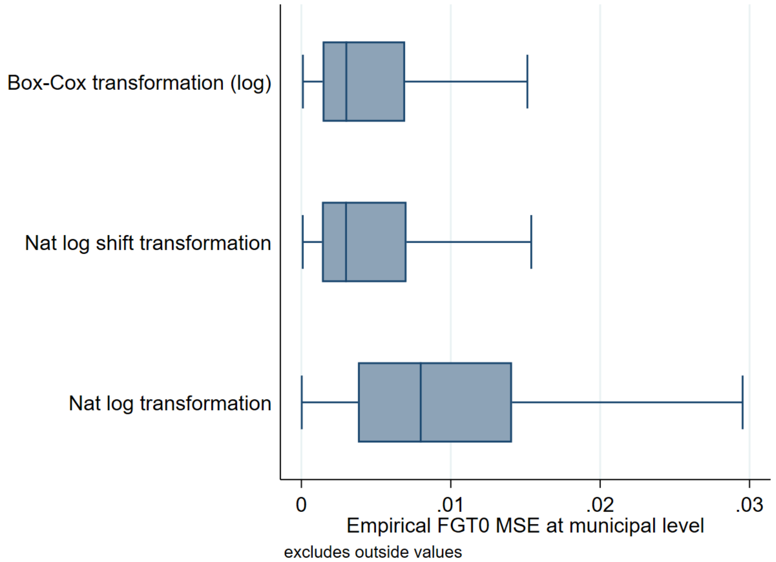




Bias (CEB_a)



MSE (CEB_a)



Transformation is necessary to achieve an approximately normal distribution

Back

The x-axis represents the percentile corresponding to a defined poverty line, the y-axis is the empirical bias. Each line corresponds to an area. See Corral et al. (2022) for details.

- The implemented SAE methodology relies on approximating as closely as possible the welfare distribution
- A model that that approximates the welfare distribution poorly may yield a decent estimate for a given poverty line, yet when judged at a different poverty line the method may completely fail
- You can see this in the figure on the left.
  - A model where the model's assumptions are imperfectly met may lead to the far-left figure. It will approximate values at certain thresholds but will do a poor job at others.
  - The model with good approximation has values which are adequate for every single area and threshold.
- New "sae" codes allow for transformations beyond the natural log so that the assumptions can be better met.

WORLD BANK GROUP

The Alpha Model — Modeling Residual Variance in SAE

Back

Context: Why an Alpha Model?

- In the **Henderson III GLS** framework, we first model the **mean** of the welfare indicator → that's the **β-model** (or *fixed-effects* model).
- But residuals often show **unequal variability** across households or areas → rural vs. urban, small vs. large households, etc.
- The **α-model** captures and explains this **heteroskedasticity**.

Component	Equation	Role
β-model (mean)	$y_{ij} = x'_{ij}\beta + u_i + e_{ij}$	Predicts the **expected welfare**
α-model (variance)	$\log(\sigma^2_{ij}) = z'_{ij}\alpha$	Explains **residual variance**

Interpretation

- The α-model refines GLS weights, improving efficiency.
- Accounts for heterogeneous populations (urban/rural, household size).
- Leads to **more precise predictions** and **better uncertainty estimates**.

WORLD BANK GROUP

Model Selection: Stepwise vs. LASSO

Back

Stepwise Selection

•Traditional inclusion/exclusion based on significance (AIC, BIC, p-value).
•Produces one "best" model through iterative testing.
•Simple and interpretable, but sensitive to noise and correlations.
•Common in ELL and EB applications (default in WB SAE Stata package).

LASSO Regression

•Penalized regression: adds an ℓ_1 penalty that shrinks weak coefficients to zero.
•Performs variable selection and regularization simultaneously.
•More stable in high-dimensional data and robust to multicollinearity.
•Increasingly used for model validation and robustness checks.

In short:

Stepwise = rule-based selection; LASSO = data-driven shrinkage.
Both aim to simplify the model and improve predictive power, but LASSO is more modern and statistically stable.

LASSO

Back

In LASSO, we estimate the regression coefficients by solving this optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

This formula has **two parts**:

The Loss Term — Fit the Data

This is the **ordinary least squares (OLS)** objective: minimize the sum of squared residuals.

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

If this were the only term, you would get the same result as OLS.

WORLD BANK GROUP

LASSO

Back

The L1 Penalty Term — Shrink Coefficients

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

This part is what makes LASSO special.

- It **adds a cost** for having large coefficients.
- The penalty is based on the **absolute value** of the coefficients.
- When λ increases → the penalty becomes stronger → coefficients shrink more.

Small λ → minimal shrinkage → β ≈ OLS → risk of overfitting.
Moderate λ → some β shrink, some become exactly zero → automatic variable selection.
Large λ → strong shrinkage → most β = 0 → underfitting.
LASSO shrinks coefficients **unequally**, removing weak or collinear predictors first.
Cross-validation chooses λ that balances **bias–variance tradeoff** for best prediction.

Stage 1 — Monte Carlo Simulation (with fixed parameters)

- This step **treats the model parameters as fixed** at their estimated values:

$$\hat{\beta}, \hat{\sigma}_\eta^2, \hat{\sigma}_e^2$$

  Using these fixed parameters, we:
  Simulate random area effects

$$\eta_c^* \sim N(0, \hat{\sigma}_\eta^2)$$

  Simulate household residuals

$$e_{ch}^* \sim N(0, \hat{\sigma}_e^2)$$

  Generate synthetic consumption for every census household

$$y_{ch}^* = x_{ch}\hat{\beta} + \eta_c^* + e_{ch}^*$$

  This produces **one synthetic census** consistent with the model.

- Purpose: **Recover the welfare distribution in each area.**

Stage 2 — Parametric Bootstrap

Back

- Monte Carlo alone **ignores uncertainty in the estimated parameters**

$$\hat{\beta}, \hat{\sigma}_\eta^2, \hat{\sigma}_e^2$$

Therefore, a **parametric bootstrap** is added:

- **For each bootstrap iteration $b = 1, \dots, B$:**
  1. **Generate synthetic survey data** using the model.
  2. **Re-estimate parameters**

$$\hat{\beta}^{(b)},\ \hat{\sigma}_\eta^{2(b)},\ \hat{\sigma}_e^{2(b)}$$

  3. **Run Monte Carlo again** with these newly estimated parameters.
  4. **Compute welfare indicators**.

- Across the B bootstrap rounds, compute:
  Mean estimate (EB predictor)
  MSE estimate (uncertainty)

Parameter Bootstrap

What is Parameter Bootstrap:
Parameter bootstrap is a resampling method where we simulate new datasets using the estimated model parameters, instead of resampling the original data directly.

Why we need it:
In SAE, the census does not contain consumption, and the model contains random components (η_c, e_{ch}) that we never observe. To quantify prediction uncertainty, we simulate these random parts using the estimated model parameters.

Parameter Bootstrap

Idea:

Simulate new datasets **from the fitted model**, not by resampling the original data.

Steps:

1. Use estimated model parameters

$$\hat{\beta}, \hat{\sigma}_{\eta}^{2}, \hat{\sigma}_{e}^{2}$$

2. Draw random area effects and random household errors:

$$\eta_c^{*(b)} \sim N\left(0, \hat{\sigma}_{\eta}^{2}\right) \qquad e_{ch}^{*(b)} \sim N\left(0, \hat{\sigma}_{e}^{2}\right)$$

3. Generate synthetic welfare

$$y_{ch}^{*(b)} = x_{ch}'\hat{\beta} + \eta_c^{*(b)} + e_{ch}^{*(b)}$$

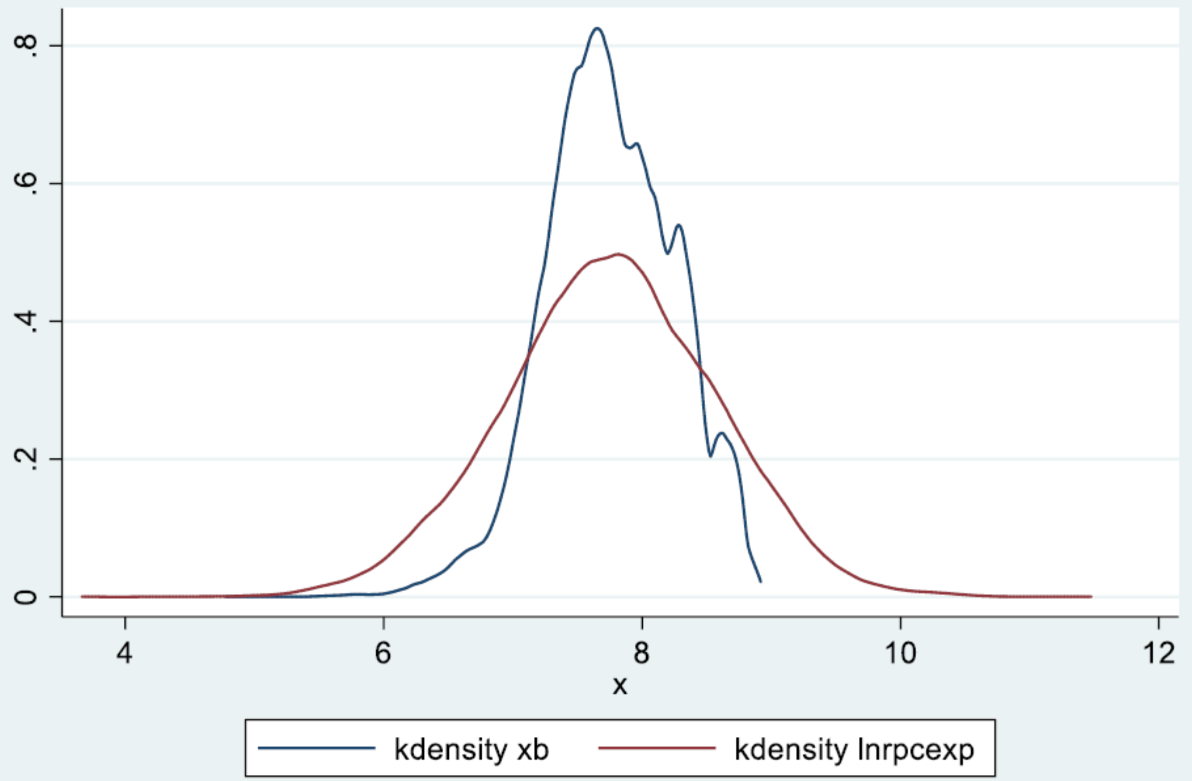
4. Re-estimate parameters (EB/CensusEB)

Captures **parameter uncertainty**.

5. Repeat 100–500 times to obtain:

- Mean poverty estimate
- Standard error (uncertainty)
- MSE
- Bias diagnostics

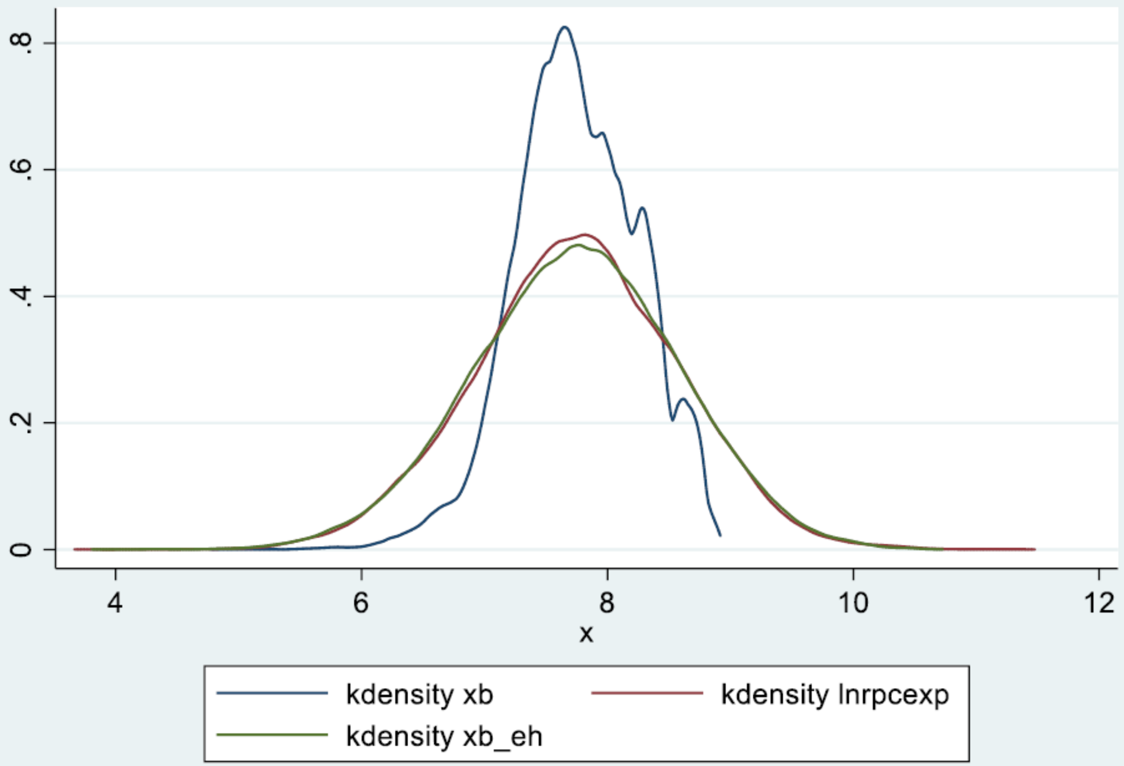
Why do we simulate residuals?



Note how the distribution is tighter on the XB, even if the average is the same.
The shape is also very different

- Note how the shapes of the distributions differ
- If poverty were to happen at a value of 6, we would miss this entirely if we just used $X\beta$
  - **Imputation is not prediction**
  - Prediction ignores inherent uncertainty of the unknown census welfare

Why do we simulate residuals?



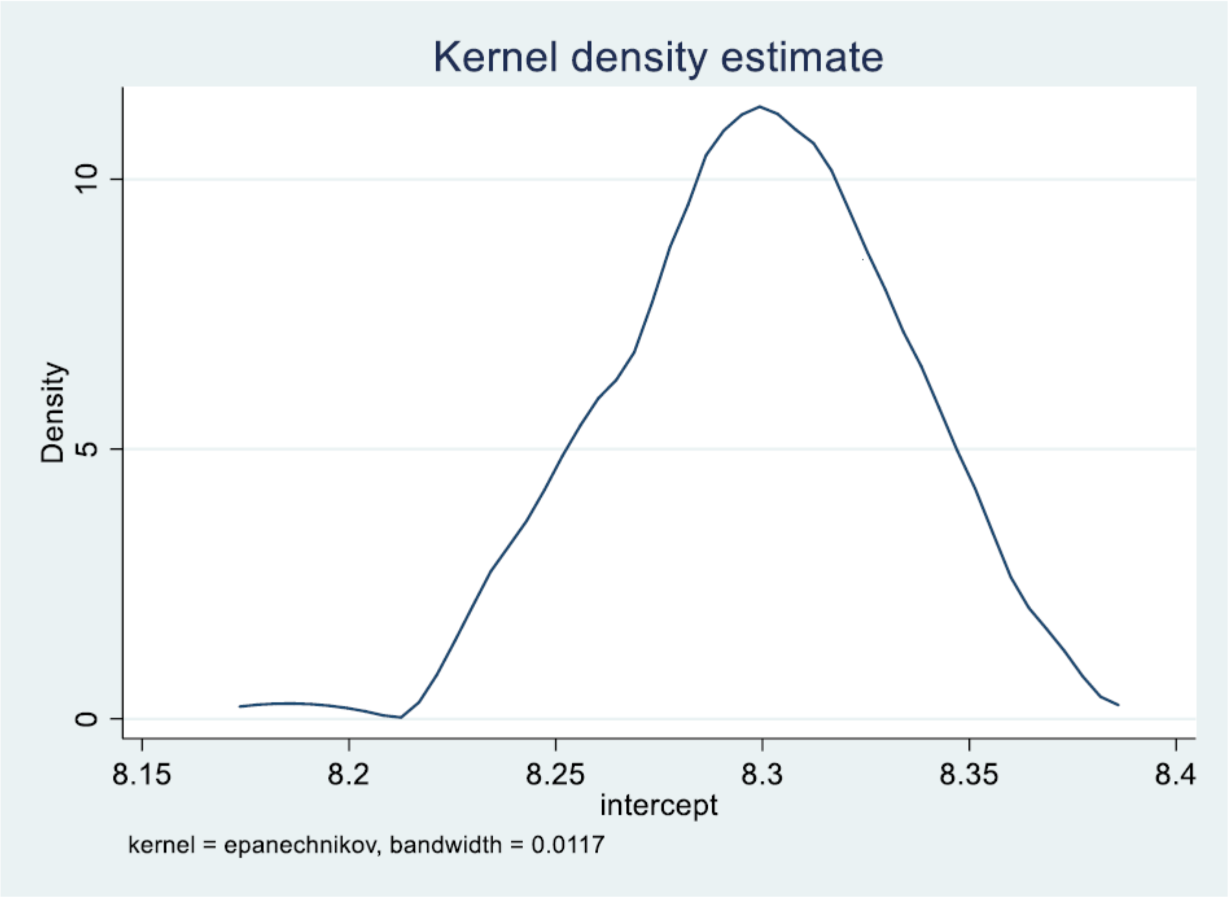
See how the distributions are now much more aligned!

When we simulate the residuals, the distributions are more aligned

- Even in the tails!

Why do we simulate $\boldsymbol{\beta}$s and other parameters?

Back



- Only adding the noise parameter requires us to know the true underlying parameters
- If you had drawn a different sample of the data from the same population, these parameters would change
- Simulation process allows us to resample data and re-estimate parameters using the resampled data

Back

Model Estimation: GLS vs. REML

GLS (Generalized Least Squares)

- Estimates regression coefficients (β) using a **known or assumed variance structure**.
- **Supports** both heteroskedasticity (via the α-model) and survey weights.
- Can **underestimate variance** if the variance components are estimated rather than known.
- Produces results nearly identical to REML if weights and heteroskedasticity are not specified

REML (Restricted Maximum Likelihood)

- Estimates variance components ($\sigma_\eta^2, \sigma_e^2$) **first**, accounting for the loss of degrees of freedom from estimating β.
- Provides **unbiased and more stable** estimates of variance parameters.
- **Does not** support survey weights or heteroskedasticity.

Practical Guidance

- **REML** is ideal for *academic or simulation exercises*—simple, clean, unweighted data.
- **GLS** is preferred for *real-world poverty mapping*, where survey weights and heteroskedasticity are relevant.
- Under simple conditions, **both produce nearly identical results**.