Post Empirical Bayes Regression

Sheng-Kai Chang [1] Yu-Chang Chen [1] Shuo-Chieh Huang [2]
Shen-Hsun Liao [1]

[1]National Taiwan University

[2]University of Chicago

August 26, 2024

Introduction: a Simple Motivating Case

Regression of interest (θ_i unobserved):

$$y_{ij} = \beta_0 + \beta_1 \theta_i + \epsilon_{ij}$$
$$\mathbf{E}[\theta_i \epsilon_{ij}] = \mathbf{E}[\epsilon_{ij}] = 0$$

The EB model:

$$\theta_i \sim N(\mu_\theta, \sigma_\theta^2),$$
$$X_{ij} \mid \theta_i \sim N(\theta_i, \sigma_x^2)$$

- θ_i: teacher i's quality
- X_{ij}: student j's test score

Our question: let $\hat{\beta}^{EB} =$ OLS regression of y_{ij} on $\hat{\theta}_i^{EB}$. Is $\hat{\beta}^{EB}$ consistent?

Background

Empirical Bayes (EB) is a popular method for estimating fixed effects

- e.g., teacher quality, neighborhood effects, hospital effects, ...
- challenge: each unit only has a few observations $\Longrightarrow$ shrinkage

EB estimates are often used as inputs to other statistical procedures

- kernel density estimator, e.g., distribution of teacher quality
- regression, e.g., effect of teacher quality on labor market outcomes

Despite its prevalence, a formal theoretical analysis of such two-step procedure seems to be lacking

This Paper

We develop a two-step method that uses EB estimates in a regression: "Post Empirical Bayes Regression"

- allows general EB estimators, including nonparametric EB
- allows both linear and nonlinear regression

We show that the estimator is consistent and asymptotically normal

Our method provides a coherent framework to estimate

- the fixed effects (the $\theta's$)
- the distribution of fixed effects
- the regression model of interest

Outline of Today's Talk

EB in a nutshell

A simple case: normal-normal EB + simple regression

General setup: nonparametric EB + (possibly nonlinear) regression

Simulation

EB in a Nutshell

EB in a Nutshell: the Bayesian Part

Suppose that we want to estimate the mean of a normal distribution

$$X \sim N(\theta, \sigma_x^2),$$

in which the variance σ_x^2 is known

Bayesian imposes a prior distribution on the parameter θ, say,

$$\theta \sim N(\mu_\theta, \sigma_\theta^2),$$

in which both μ_θ and σ_θ^2 are known

EB in a Nutshell: the Bayesian Part (Cont'd)

Upon observing the realization $X = x$, Bayesian updates the prior

$$\theta \mid X = x \sim N(\cdot, \cdot),$$

where the posterior mean is

$$\frac{\sigma_x^2}{\sigma_x^2 + \sigma_\theta^2} \cdot \mu_\theta + \frac{\sigma_\theta^2}{\sigma_x^2 + \sigma_\theta^2} \cdot x$$

- shrinkage: the empirical evidence $X = x$ are shrunk to the prior mean μ_θ based on the signal-to-noise ratio

Under the squared loss, the posterior mean is the **Bayes rule** of μ_θ

From Bayes to Empirical Bayes

EB methods tries to "estimate" the prior distribution from data

- Then, proceed as Bayesian with the estimated prior

Given teacher quality θ_i, the test score X_{ij} follows normal distribution

$$X_{ij} \mid \theta_i \sim N(\theta_i, \sigma_x^2),$$

and, as in the Bayesian approach, we impose

$$\theta_i \sim N(\mu_\theta, \sigma_\theta^2)$$

- However, the parameters $(\mu_\theta, \sigma_x^2, \sigma_\theta^2)$ are not assumed to be known

Estimating the Prior

The model:

$$\theta_i \sim N(\mu_\theta, \sigma_\theta^2),$$
$$X_{ij} \mid \theta_i \sim N(\theta_i, \sigma_x^2).$$

How can we estimate $(\mu_\theta, \sigma_\theta^2, \sigma_x^2)$? Notice that

- μ_θ = average of teacher quality = average test score across teachers
- σ_θ^2 = variation of teacher qualities
- σ_x^2 = within teacher variation variation of test score

which can be estimated by their empirical analogues

The EB Estimator

Given $(\hat{\mu}_\theta, \hat{\sigma}^2_\theta, \hat{\sigma}^2_x)$, the EB estimator of teacher i's quality θ_i is

$$\hat{\theta}_i^{EB} = \frac{\hat{\sigma}^2_x/m}{\hat{\sigma}^2_x/m + \hat{\sigma}^2_\theta} \cdot \hat{\mu}_\theta + \frac{\hat{\sigma}^2_\theta}{\hat{\sigma}^2_x/m + \hat{\sigma}^2_\theta} \cdot \bar{x}_i,$$

where $\bar{x}_i$ is the class average test score and m is the class size

- EB estimator is the Bayes rule except the prior is estimated

$\hat{\theta}_i^{EB}$ is $\bar{x}_i$ but shrunk to the grand mean $\hat{\mu}_\theta$

- But its target and amount of shrinkage are decided by the data (instead of a prior imposed by the researcher)

A simple case: normal-normal EB + simple regression

EB Estimates as Regressors

Suppose, in addition to θ_i, we are also interested in the regression

$$y_{ij} = \beta_0 + \beta_1 \theta + \epsilon_{ij}$$

- e.g., effect of teacher quality θ_i on students' wage y_{ij}

Can we regress y_{ij} on $\hat{\theta}_i^{EB}$? In particular,

- $\hat{\theta}_i^{EB}$ is a noisy measurement of θ_i
- $\hat{\theta}_i^{EB}$ is biased for θ_i

Consistency: Attenuation Bias Perspective

$\hat{\beta}^{EB}$ is in fact consistent. Recall that if we regress y_{ij} on $\bar{x}_i$

$$\begin{aligned}\text{Attenuation Bias} &= \text{Signal-to-Noise Ratio} \\ &= \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_x^2/m} \\ &= \text{Shrinkage factor}\end{aligned}$$

Shrinkage happens to cancel out the attenuation bias

- First noted by Whittermore (1989)
- Gao and Ghosh (2012) calculates its MSE, proving consistency

Consistency: IV Perspective

Recall that we can address measurement errors by instrumental variable

Explanation II: note that EB is an estimate of the posterior mean

$$\hat{\theta}_i^{EB} = \hat{\mathbf{E}}[\theta_i | x_1, x_2, ..., x_J]$$

which can be thought of as fitted value of the first-stage fitted value of θ_i

- So regressing on $\hat{\theta}_i^{EB}$ effectively constitutes a 2SLS regresion

Extension to Non-linear Case

Replacing θ_i with $\hat{\theta}_i^{EB}$ in simple regression gives consistent estimates

Does it generalize to other cases?

- e.g., can I plug in $\hat{\theta}_i^{EB}$ into a probit?
- also, what if the EB estimates come from EB models other than the normal-normal specification

General Setup

Model Setup: the EB Part

Model for the EB estimator: write $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{im})$

$$\theta_i \overset{iid}{\sim} \pi,$$
$$\boldsymbol{x}_i|\theta_i \overset{iid}{\sim} \mu_\theta(\boldsymbol{x}_i; \gamma),$$

- π is the distribution of fixed effects
- $\mu_\theta(\boldsymbol{x}_i; \gamma)$ is the likelihood with unknown nuisance parameter γ

Example: beta-binomial models for hospital effects

- $\theta_i =$ hospital effect. $\theta_i \sim B(\alpha, \beta)$
- $x_{ij} =$ successful recovery. $x_{ij}|\theta_i \sim Ber(\theta_i)$

We can allow π to be non-parametrically specified (but $\mu_\theta(;\gamma)$ has to be known up to γ)

Model Setup: the Regression

Model of interest:

$$y_{ij} = g_1(\theta_i; \beta) + g_2(z_{ij}; \delta) + \epsilon_{ij}$$
$$\mathbf{E}[\epsilon_{ij}|\theta_i] = \mathbf{E}[\epsilon_{ij}|z_{ij}] = 0,$$

where z_{ij} is an (observed) covariate and $g_1(\cdot)$ and $g_2(\cdot)$ are the conditional mean functions. Examples of $g_1(\cdot)$:

- $g_1(\theta_i, \beta) = \beta_0 + \theta_i\beta_1 + \theta_i^2\beta_2$
- $g_1(\theta_i, \beta) = \Phi(\theta_i\beta)$

We will view $g_2(\cdot)$ and δ as nuisance

Can we replace θ_i with $\hat{\theta}_i^{EB}$ and run (possibly non-linear) regression?

- answer: no

EB as Regressors: Non-linear Models (Cont'd)

Replacing θ_i with $\hat{\theta}_i^{EB}$ generally leads to inconsistent estimators. Fundamental reason:

$$\mathbf{E}[g_1(\theta_i, \beta)|\boldsymbol{x}_i] \neq g_1(\mathbf{E}[\theta_i|\boldsymbol{x}_i], \beta),$$

i.e., posterior mean of the transformation is **NOT** the transformation of the posterior mean

Whereas in the linear model, $g_1(\theta_i, \beta) = \theta_i\beta$, and

$$\mathbf{E}[g_1(\theta_i, \beta)|\boldsymbol{x}_i] = \mathbf{E}[\theta_i\beta|\boldsymbol{x}_i] = \mathbf{E}[\theta_i|\boldsymbol{x}_i]\beta = g_1(\mathbf{E}[\theta_i|\boldsymbol{x}_i], \beta)$$

Proper Post-EB Regression

The correct way is to calculate the posterior mean of $g_1(\cdot)$

$$\hat{g}_{1,i}^{EB} = \mathbf{E}[g_1(\theta_i, \beta)|\boldsymbol{x}_i]$$

instead of incorrectly plugging in the posterior mean of θ_i in $g_1(\cdot)$

$$g_1(\hat{\theta}_i^{EB}, \beta)$$

- To calculate the posterior mean, we need to first estimate the prior π and the nuisance parameter γ in the likelihood $\mu_\theta(\boldsymbol{x}_i; \gamma)$

Step 1: Estimate the Nuisance Parameters

Recall the model:

$$\theta_i \overset{iid}{\sim} \pi, \quad \boldsymbol{x}_i|\theta_i \overset{iid}{\sim} \mu_\theta(\boldsymbol{x}_i;\gamma),$$
$$y_{ij} = g_1(\theta_i;\beta) + g_2(z_{ij};\delta) + \epsilon_{ij}$$

Step 1: estimate the nuisance parameters (γ, δ)

- For γ, use MLE or method of moments (e.g., within-class variation)
- For δ, run a regression of y_{ij} on z_{ij} with fixed effect $\tilde{\theta}_i = g_1(\theta_i;\beta)$

Step 2: Estimate the Prior

The EB model:

$$\theta_i \overset{iid}{\sim} \pi,$$
$$\boldsymbol{x}_i|\theta_i \overset{iid}{\sim} \mu_\theta(\boldsymbol{x}_i;\gamma)$$

Note that we only observe x_{ij}, whose likelihood function depend on π:

$$f(\boldsymbol{x}_i) = \int \mu_\theta(\boldsymbol{x}_i;\gamma)\mathrm{d}\pi(\theta)$$

- MLE: estimate π by finding the distribution π that maximizes the probability of observing $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$

Step 2: Estimate the Prior (Cont'd)

Likelihood function of $\boldsymbol{x}_i$:

$$f(\boldsymbol{x}_i) = \int \mu_\theta(\boldsymbol{x}_i;\gamma)\mathrm{d}\pi(\theta)$$

Step 2: estimate the prior π with nonparametric (mixture) MLE:

$$\hat{\pi} = \arg\max_{\pi\in\Pi} \sum_i ln\left[\int \mu_\theta(\boldsymbol{x}_i;\hat{\gamma})\mathrm{d}\pi(\theta)\right],$$

where Π is the set of all possible distributions

Computation can be challenging. Modern convex optimization techniques are needed for larger problems (# obs. $\geq 10^6$)

Step 3: Running the Regression

Step 3: given the estimated prior $\hat{\pi}$ and likelihood $\mu_\theta(\boldsymbol{x}; \hat{\gamma})$, calculate the posterior mean

$$\hat{g}_1^{EB}(\theta_i, \beta) = \hat{\mathbf{E}}[g_1(\theta_i, \beta)|\boldsymbol{x}_i],$$

and run the regression

$$y_{ij} - g_2(z_{ij}; \hat{\delta}) = g_1(\theta_i; \beta) + \epsilon_{ij},$$

i.e., solve

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[y_{ij} - g_2(z_{ij}; \hat{\delta}) - \hat{g}_1^{EB}(\theta_i; \boldsymbol{\beta}) \right]^2$$

Main Advantages of the Method

The main advantages of our method:

- can handle non-classical measurement errors
- requires no tuning parameter while being nonparametric for the π
- has simple asymptotic distribution

and our method also deliver estimates of fixed effects and its distribution

Our analysis can be used to examine empirical practice. We show that naive implementation of post-EB regression could be problematic

- e.g., use EB estimates as regressors in a logit model

Main Assumption: Identification

Assumption (Identification of π)

Let $F(\mathbf{x}; \pi, \boldsymbol{\gamma})$ be the cumulative distribution function of $\boldsymbol{x}$. For all $\boldsymbol{\gamma} \in \Gamma$, if $F(\mathbf{x}; \pi, \boldsymbol{\gamma}) = F(\mathbf{x}; \pi', \boldsymbol{\gamma})$ for all $\mathbf{x}$, then $D_{KW}(\pi, \pi') = 0$.

Assumption (Identification of $\boldsymbol{\beta}$)

Let $\boldsymbol{\beta}_o$ denote the true value of $\boldsymbol{\beta}$. We have

$$\mathbb{E}\left\{\mathbb{E}[g_1(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o) - g_1(\boldsymbol{\theta}_i; \boldsymbol{\beta})|\boldsymbol{x}_i]\right\}^2 = 0$$

if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_o$.

Main Assumption: Nuisance Parameters

Assumption (Consistent Estimation)

We have consistent estimators $\hat{\gamma}, \hat{\delta}$ *for* (γ, δ).

Assumption (Bounded Derivative for NPMLE)

Let $Q(\pi, \gamma) = \mathbb{E}_{\mathbf{x}}[\int_{\Theta} \mu_{\theta}(\mathbf{x}; \gamma) \mathrm{d}\pi(\boldsymbol{\theta})]$. *There exists a neighborhood* N_{γ_o} *of* γ_o *such that* $\sup_{\gamma \in N_{\gamma_o}} |\frac{\partial Q(\pi, \gamma)}{\partial \gamma}| \leq M$ *for some* $M > 0$.

Assumption (Bounded Derivative for NLS)

Let $Q_n(\beta; \delta)$ *denote the objective function of the regression. and* β_o *be the true value. We have (i)* $Q_n(\beta; \delta)$ *is twice continuously differentiable in a neighborhood* $\mathcal{N}$ *of* β_o*; (ii) there exists* $H(\beta)$ *that is continuous at* β_o *and* $\sup_{\beta \in \mathcal{N}} ||\nabla_{\beta\beta} Q_n(\beta; \delta) - H(\beta)|| < M$ *for some* $M > 0$*; (iii)* $H = H(\beta)$ *is nonsingular.*

Main Assumption: Regularity Conditions

Assumption (Contintuity)

For all $\gamma \in \Gamma$, the function $f(\mathbf{x}; \pi)$ is continuous in $\pi \in \Pi$.

Assumption (Moment Existence)

The expectations

$$\mathbb{E} \sup_{\beta \in \mathcal{B}} (g_1(\boldsymbol{\theta}_i; \beta))^2$$

$$\mathbb{E} \left\{ y_{ij} - \mathbb{E}[g_1(\boldsymbol{\theta}; \beta) | \boldsymbol{x}_i] \right\}^2$$

exist.

Theoretical Result

Under the aforementioned conditions and some other regularity conditions, we can show that

- The NPMLE $\hat{\pi}$ converges to the true prior π
- The estimated posterior mean $\hat{g}_1^{EB}(\theta_i, \beta)$ converges to $\mathbf{E}[g_1(\theta_i, \beta)|\boldsymbol{x}_i]$
- Post-EB regression $\hat{\beta}$ is consistent and asymptotically normal

Simulation

Proposed Method vs. Naive

Data generation process:

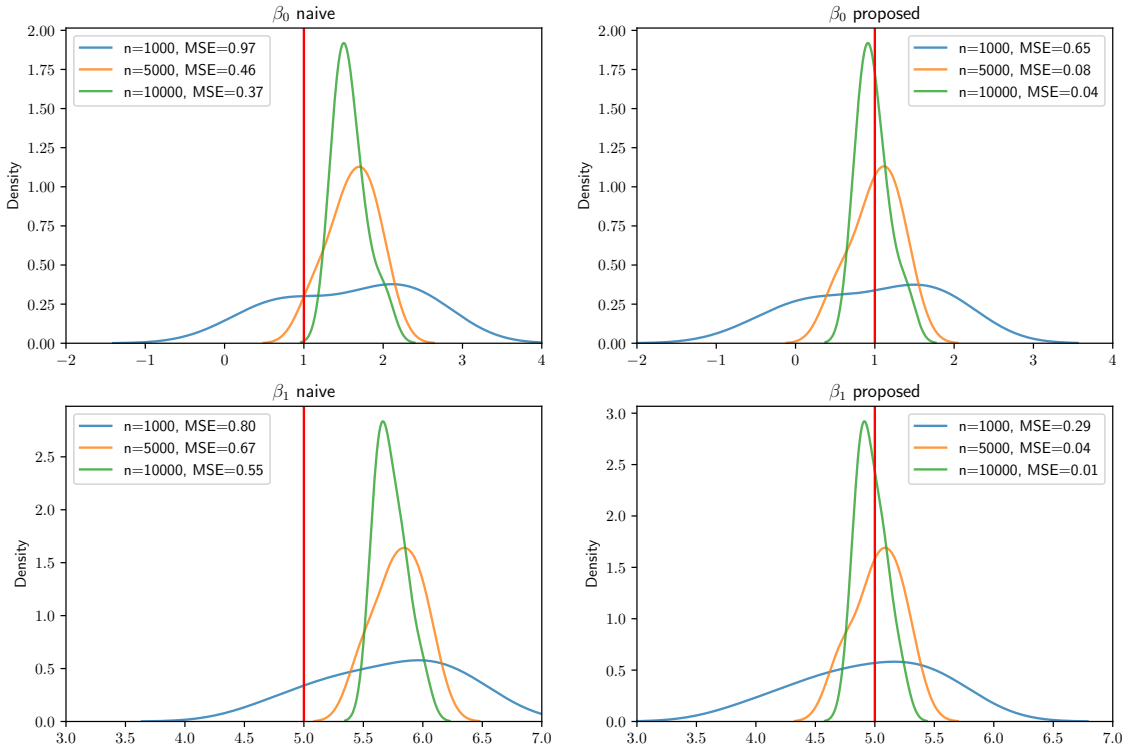
$$\theta_i \sim Beta(a, b)$$
$$10X_i \mid \theta_i \sim Bin(10, \theta_i)$$
$$Y_i = \beta_0 + \beta_1 \ln \theta_i + e_i, \; e_i \sim \mathcal{N}(0, \eta^2)$$

We compare our proposed method to the naive method

- naive method: regress y on $\ln \hat{\theta}_i^{EB}$ (biased for nonlinear models)
- proposed method: regress y on $\hat{\mathbb{E}}[\ln \theta_i | X_i]$

Simulation Results: Naive vs. Proposed Method

Naive vs. proposed method



Robustness against Nonparamtric Prior

Now we investigate our method's robustness against different priors

Data generation process:

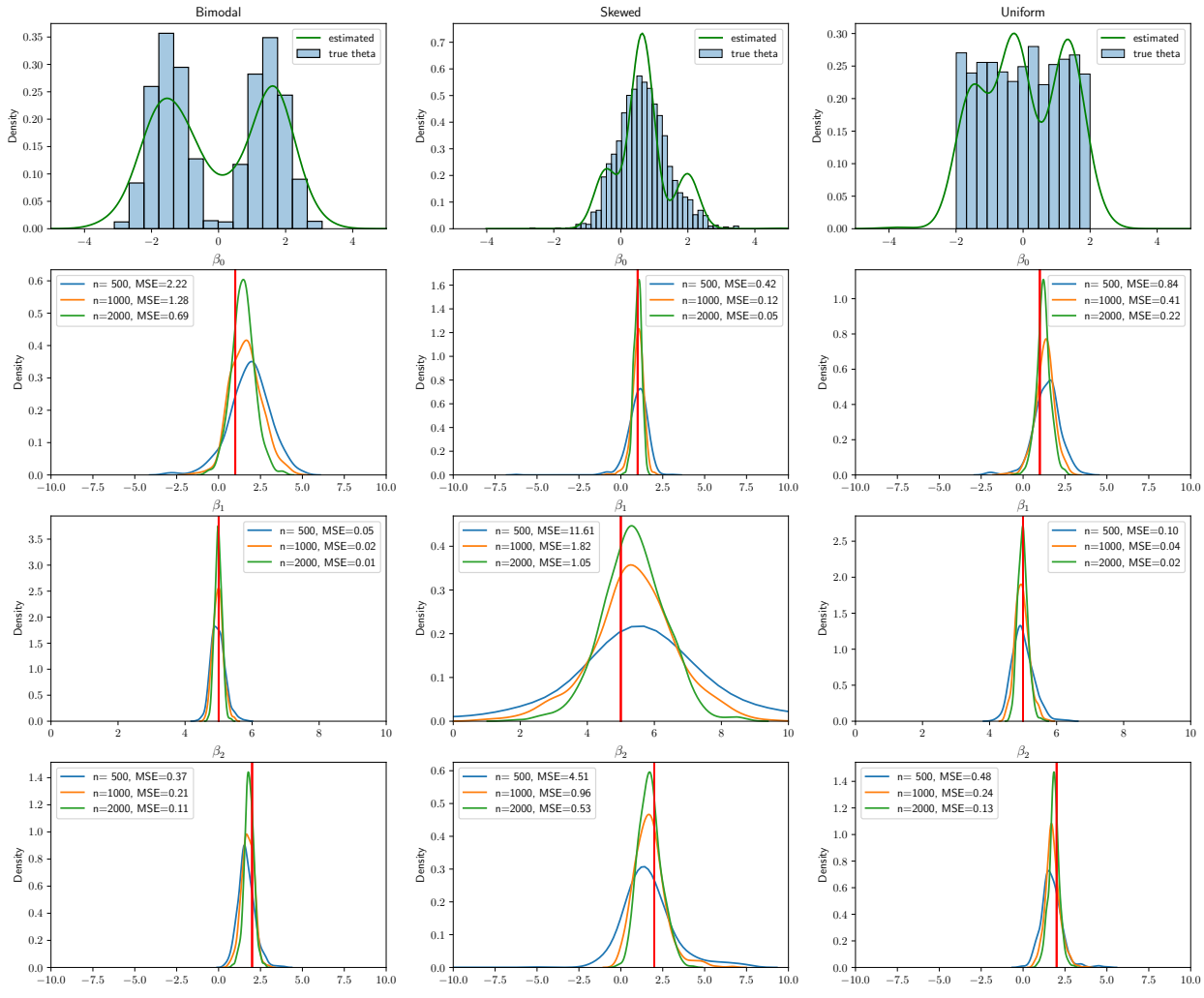
$$\theta_i \sim G$$

$$X_{ij} \mid \theta_i \sim \mathcal{N}(0, \sigma^2)$$

$$Y_{ij} = \beta_0 + \beta_1 \theta_i + \beta_2 \theta_i^2 + \alpha Z_{ij} + e_{ij},\ e_{ij} \sim \mathcal{N}(0, \eta^2)$$

- Estimate the prior G nonparametrically
- Prior $G =$ bimodal, skewed normal, uniform

Simulation Results: Nonparametric Prior



Conclusion

We propose a two-step method that

- allows nonparametric prior and nonlinear regression
- provides a coherent way to estimate both fixed effects and regression
- has consistency and asymptotic normality

Our results also serve as a benchmark to empirical practice regarding the use of EB estimates

Empirical Application

Empirical Application: Tennesse STAR

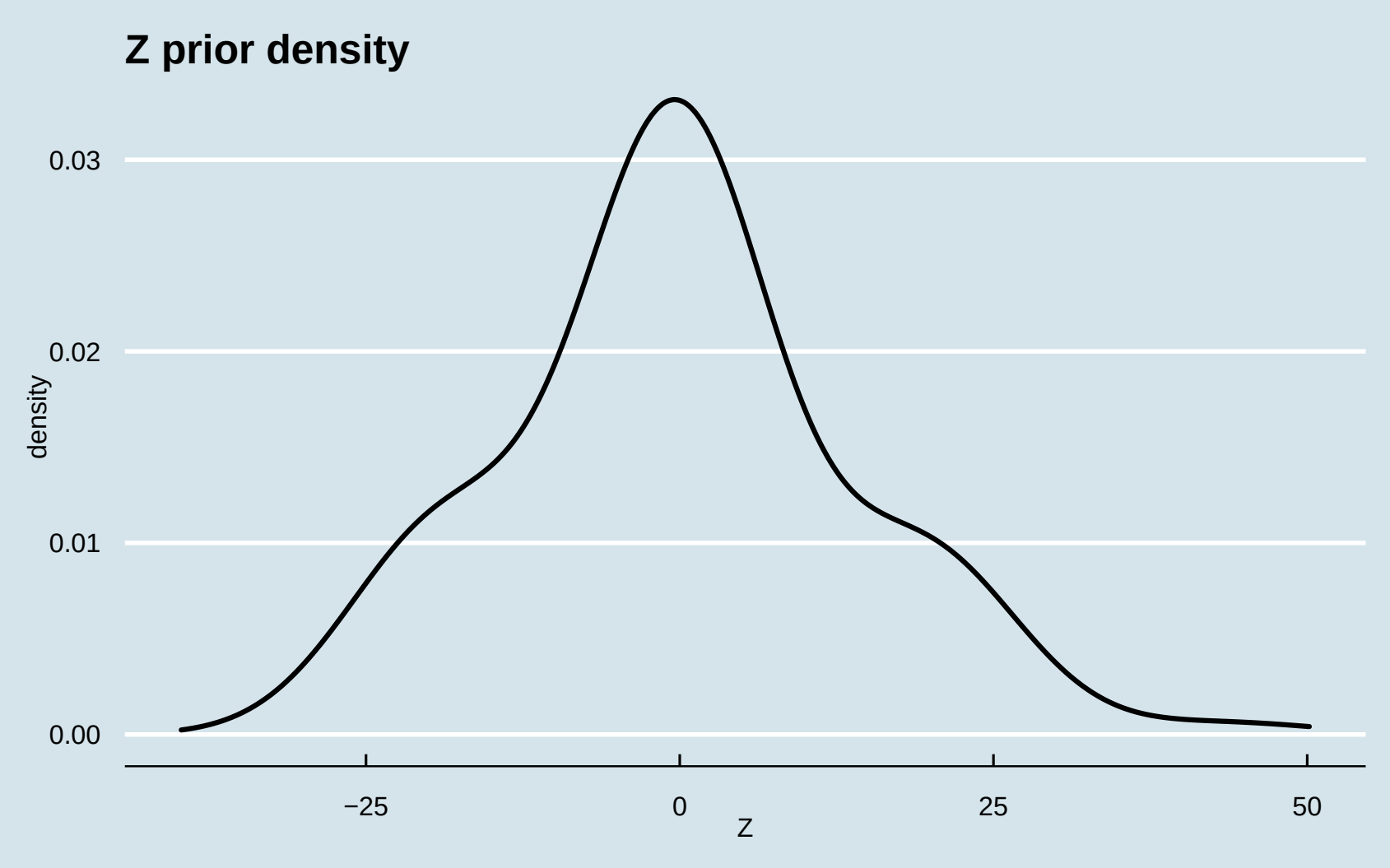
We use the Tennessee STAR data to demonstrate our methods

- panel data of students' academic performance from kindergarten
- gender, free lunch status, and teacher ID

We study the quality of kindergarten teachers

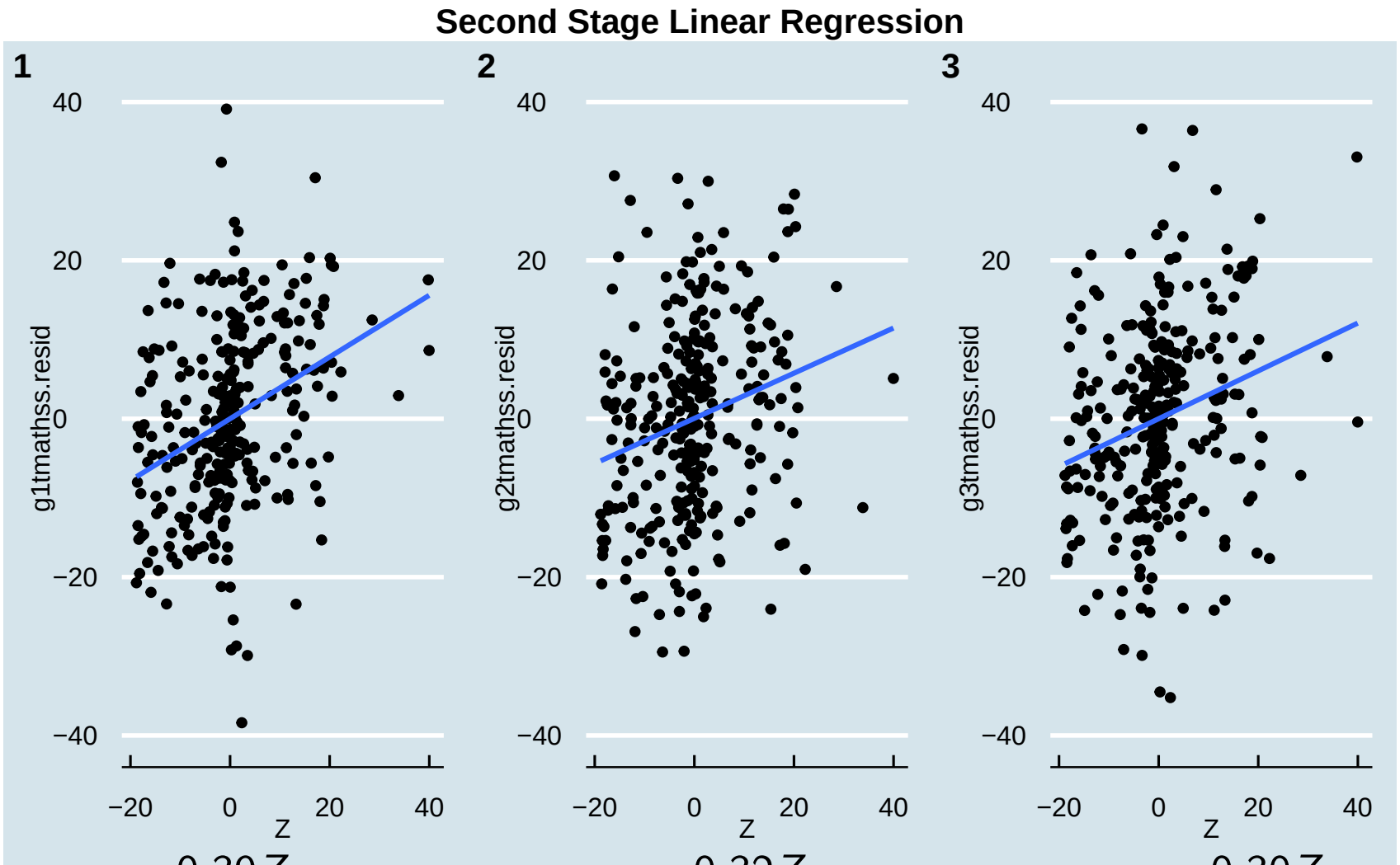
- and their effects on 1^{st}, 2^{nd} and 3^{rd} grade academic performance

Teacher Quality Distribution



μ : -0.003, σ : 12.7, skewness : 0.44, kurtosis : 3.53

Long-Term Effects of Kindergarten Teachers



$y_1 = 0.39Z$ $y_2 = 0.32Z$ $y_3 = 0.30Z$

Long-Term Effects of Kindergarten Teachers