Advances in statistical inference for longitudinal generalized linear models

flipscores Two-Stage Summary Statistics

Angela Andreella
angela.andreella@unitn.it

BEYOND Summer School

May 29, 2025

So . . .

We introduce some concepts about GLMMs and GEEs. We have seen that both have advantages and disadvantages. Let's focus on the disadvantages again ☻:

GLMM:

- Uncertainty about how to specify the random effects structure,
- Invalid inference under model misspecification (both random effects and fixed effects structure),
- Convergence issues.

GEE:

- Problems with small sample sizes,
- Difficulties with unbalanced designs,
- Presence of endogenous covariates — i.e., variables correlated with the error term — when assuming independence,
- Non-randomly missing data.

Our proposal

Psychometrika (2025), 0, 1–23
doi:10.1017/psy.2024.22

PSYCHOMETRIC SOCIETY

APPLICATION REVIEWS AND CASE STUDIES

Robust Inference for Generalized Linear Mixed Models: A "Two-Stage Summary Statistics" Approach Based on Score Sign Flipping

Angela Andreella[1], Jelle Goeman[2], Jesse Hemerik[3] and Livio Finos[4]

[1]University of Trento, Italy; [2]Leiden University Medical Center, The Netherlands; [3]Erasmus University Rotterdam, The Netherlands; [4]University of Padova, Italy

Corresponding author: Angela Andreella; Email: angela.andreella@unitn.it

(Received 24 October 2024; accepted 11 December 2024)

This manuscript is part of the special section *Model Identification and Estimation for Longitudinal Data in Practice.* We thank Drs. Carolyn J. Anderson and Donald Hedeker for serving as co-Guest Editors.

flip2sss

Ingredients for the **flip2sss** (flipscores two-stage summary statistics):

- Permutation theory (flip)
- Score test (scores)
- Two-stage summary statistics approach (2sss)

The proposed method is efficiently implemented in the `R` package `jointest`, which is compatible with large datasets and complex statistical models.

flipscores (brief recap)

Consider n independent observations $y_1, \ldots, y_n$ following a GLM from the exponential dispersion family:

$$f(y_i; \theta_i; \phi_i) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right\}$$

where θ_i is the canonical parameter, ϕ_i is the dispersion parameter, and $\mu_i = \mathbb{E}(y_i) = b'(\theta_i)$. The GLM is then defined as:

$$\boxed{g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}}$$

where $g(\cdot)$ is the link function, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$, $\boldsymbol{X}$ is the design matrix, and $\boldsymbol{\beta}$ is the vector of q parameters.

flipscores (brief recap)

The main focus is to test the null hypothesis for a given element d of $\boldsymbol{\beta}$, while still accounting for all nuisance parameters, i.e.,

$$H_0 : \beta_d = \beta_0 \mid \beta_1, \ldots, \beta_{d-1}, \beta_{d+1}, \ldots, \beta_q, \phi_1, \ldots, \phi_n.$$

The *effective score* is given by:

$$\boxed{S = n^{-1/2} \boldsymbol{X}_d^\top \boldsymbol{W}^{1/2} (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{V}^{1/2} (\boldsymbol{y} - \hat{\boldsymbol{\mu}})} \tag{1}$$

where:

- $\boldsymbol{V} = \text{diag}\{\text{Var}(y_i)\}$,
- $\boldsymbol{W} = \text{diag}\left\{\frac{\partial \mu_i}{\partial \eta_i}\right\} \boldsymbol{V}^{-1} \text{diag}\left\{\frac{\partial \mu_i}{\partial \eta_i}\right\}$,
- $\boldsymbol{H} = \boldsymbol{W}^{1/2} \boldsymbol{X}_{-d} (\boldsymbol{X}_{-d}^\top \boldsymbol{W} \boldsymbol{X}_{-d})^{-1} \boldsymbol{X}_{-d}^\top \boldsymbol{W}^{1/2}$,
- $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$ are the fitted values of the model under H_0.

flipscores (brief recap)

To improve small-sample reliability, we consider the standardized version:

$$\boxed{S^\star = \frac{S}{\mathrm{var}\{S\}^{1/2}}},$$

where

$$\mathrm{Var}\{S\} = n^{-1}\boldsymbol{X}_d^\top \boldsymbol{W}^{1/2}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{W}^{1/2}\boldsymbol{X}_d + o_p(1).$$

$S^\star$ is asymptotically **valid under any variance misspecification**.

The statistic S in Equation (1) can also be rephrased as a sum of n components: the ***effective score* contributions**.

flipscores (brief recap)

The p-value is computed by randomly flipping the signs of these score contributions. $\Rightarrow$ matrix $\boldsymbol{F}$ with diagonal elements -1 or 1 sampled with equal probability.

Sign-flipping the score contribution: multiplying the effective score by $\boldsymbol{F}$:

$$S(\boldsymbol{F}) = n^{-1/2}\boldsymbol{X}_d^\top \boldsymbol{W}^{1/2}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{V}^{-1/2}\boldsymbol{F}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}).$$

The corresponding standardized version is

$$\boxed{S^\star(\boldsymbol{F}) = \frac{S(\boldsymbol{F})}{\mathrm{var}\{S(\boldsymbol{F})\}^{1/2}}},$$

where

$$\mathrm{Var}\{S(\boldsymbol{F})\} = n^{-1}\boldsymbol{X}_d^\top \boldsymbol{W}^{1/2}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{F}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{F}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{W}^{1/2}\boldsymbol{X}_d + o_p(1).$$

flipscores (brief recap)

Considering B independent sign flip transformations, where $S_1^\star = S^\star(\boldsymbol{I})$ is the observed test statistic, we reject the null hypothesis $H_0 : \beta_d = \beta_0$ versus the alternative $H_1 : \beta_d > \beta_0$ at significance level α if:

$$S_1^\star > S^\star_{(\lceil (1-\alpha)B \rceil)},$$

where $S^\star_{(1)} \leq S^\star_{(2)} \leq \cdots \leq S^\star_{(B)}$ are the sorted statistics and $\lceil \cdot \rceil$ is the ceiling function.
In the same way, we reject H_0 versus $H_1 : \beta_d < \beta_0$ if:

$$S_1^\star < S^\star_{(\lceil \alpha B \rceil)},$$

and versus $H_1 : \beta_d \neq \beta_0$ if:

$$S_1^\star < S^\star_{(\lceil (\alpha/2)B \rceil)} \cup S_1^\star > S^\star_{(\lceil (1-\alpha/2)B \rceil)}.$$

adolong data

Adopted children measured longitudinally over `Time`:

- Health issues: response variable `Unhealth` (binary variable)
- `Sex` (binary variable)
- `Age` (age of the child when they arrived in the family, in years)
- `Country` (country of origin)

A possible model of interest is:

$$\texttt{Unhealth} \sim \texttt{1 + Country + Age + Sex + Time + Sex:Time}.$$

Consider again n observations $y_1, \ldots, y_n$ and n_j observations in cluster (i.e., child) j, where $n = \sum_j^N n_j$ and N is the total number of clusters (i.e., children).

The **exchangeability assumption** used to compute the null distribution of the standardized score test statistic $S^\star$ does not hold
$\Rightarrow$ **"two-stage summary statistics" approach**.

In short, we reduce the hierarchical complexity of the data by computing summary measures at the cluster level in the first stage. These summary measures are then analyzed as a response random variable in the second stage.

Therefore, the main assumption is that we have (at least asymptotically) unbiased estimators of these summary measures.

First stage

A GLM is fitted separately for each subject j including only the h covariates $\boldsymbol{K}_{ij} \in \mathbb{R}^{1\times h}$ that vary within-cluster j, i.e.,

$$g(\mu_{ij}) = \boldsymbol{K}_{ij}\boldsymbol{\tau}_j \tag{2}$$

where $\boldsymbol{\tau}_j \in \mathbb{R}^h$ is the vector of h parameters for cluster j.

Fitting the model defined in Equation (2) leads to a vector of estimated parameters $\hat{\boldsymbol{\tau}}_j \in \mathbb{R}^h$ for each cluster j corresponding to the design matrix $\boldsymbol{K}_j$ that varies within-subject j.

adolong data

First stage

We fit a logistic regression for each child where `Time` is the only within-cluster covariate. So:

$$\text{logit}[\Pr(\texttt{Unhealth}_{ij} = 1)] = \boldsymbol{K}_{ij}\boldsymbol{\tau}_j$$

where $\boldsymbol{K}_{ij} \in \mathbb{R}^{1\times 2}$:

$$\boldsymbol{K}_{ij} = \begin{bmatrix} 1 & \texttt{Time}_{ij} \end{bmatrix}$$

→ $\hat{\boldsymbol{\tau}}_j$ is our **"summary statistics"**: estimated intercept and slope for the variable `Time` for child j.

Second stage

The N vectors $\hat{\boldsymbol{\tau}}_j$ are collected in a $N \times h$ matrix $\boldsymbol{T}$ and modeled by a linear model with the between-cluster variables as predictors:

$$\begin{bmatrix} \hat{\boldsymbol{\tau}}_1^\top \\ \vdots \\ \hat{\boldsymbol{\tau}}_N^\top \end{bmatrix} = \boldsymbol{T} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E} \tag{3}$$

where

- $\boldsymbol{X} \in \mathbb{R}^{N \times l}$ is the matrix of between-cluster variables,
- $\boldsymbol{\beta} \in \mathbb{R}^{l \times h}$ are the corresponding parameters,
- $\boldsymbol{E} \in \mathbb{R}^{N \times h}$ is the matrix of errors.

Second stage

Recalling our initial model: `Unhealth` $\sim$ `1 + Country + Age + Sex + Time + Sex:Time`

- The intercept of within-cluster effects (i.e., the first column of $\boldsymbol{T}$) is modeled by the between-cluster effects `Intercept`, `Country`, `Age`, and `Sex`.
- The estimated slope related to `Time` (i.e., the second column of $\boldsymbol{T}$) is modeled only by `Intercept` and `Sex` covariates.

Therefore, in the second column of $\boldsymbol{T}$, only those coefficients of `Intercept` and `Sex` covariates will be estimated, while the ones associated with `Country` and `Age` are set to 0.

adolong data

Second stage

$\boldsymbol{\beta} \in \mathbb{R}^{12 \times 2}$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{\texttt{Intercept}} & 0 \\ \beta_{\texttt{Age}} & 0 \\ \beta_{\texttt{Sex Male}} & 0 \\ \beta_{\texttt{Country Cambodia}} & 0 \\ \beta_{\texttt{Country China}} & 0 \\ \beta_{\texttt{Country Colombia}} & 0 \\ \beta_{\texttt{Country Ethiopia}} & 0 \\ \beta_{\texttt{Country India}} & 0 \\ \beta_{\texttt{Country Thailand}} & 0 \\ \beta_{\texttt{Country Vietnam}} & 0 \\ 0 & \beta_{\texttt{Intercept}} \\ 0 & \beta_{\texttt{Sex}} \end{bmatrix}$$

adolong data

Second stage

$\boldsymbol{X} \in \mathbb{R}^{N \times 12}$, here we represent the first three child:

Unhealth	Age	Sex	Time	Country	Subj
1	7	Male	3	Ethiopia	1
0	2	Female	6	Vietnam	3
0	8	Male	1	Colombia	5
⋮	⋮	⋮	⋮	⋮	⋮

$$\boldsymbol{X} = \begin{bmatrix} 1 & 7 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 8 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Second stage

In summary, the problem is now rewritten as a multiple multivariate linear model where each observation is the **summary statistics**:

$$\hat{\boldsymbol{\tau}}_j \sim (\boldsymbol{x}_j \boldsymbol{\beta}, \boldsymbol{\psi}_j)$$

with

- $\boldsymbol{x}_j$ is the j-th row of $\boldsymbol{X}$
- $\boldsymbol{\psi}_j = \text{Var}(\boldsymbol{\epsilon}_j)$ is the variance of the independent rows of the error matrix $\boldsymbol{E} = [\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_N^\top]^\top$

$\boldsymbol{\psi}_j$ can be decomposed as a sum of two sources of variability:

- $\boldsymbol{\Sigma}$: variance of the random effects $\Rightarrow$ captures the between-subject variability due to the presence of **random effects**
- $\boldsymbol{\Sigma}_j$: conditional variance of $\hat{\boldsymbol{\tau}}_j \Rightarrow$ the second represents the **within-cluster variability**.

flip2sss

Outside the fully balanced designs with homoscedastic errors, the ψ_j varies among clusters, making the standard LM tools unreliable.

The literature on 2sss focuses on the estimate of ψ_j
$\Rightarrow$ making assumptions on the data and being constrained to a precise specification of the random quantities in the model.
$\Rightarrow$ choose whether intercept and slope are random coefficients and also among independent or correlated random effects.
$\Rightarrow$ possible **misspecification**!

➔ ***standardized flipscores*** *approach in the second stage!*

$H_0 : \beta = \beta_0$ is then tested with the standardized score test, applying the same sign-flipping transformation across all h models, still avoiding the complexities associated with formulating the random component of the model.

Any **summary measures** at the cluster level can be used in the second stage; the only required property is the (asymptotic) unbiasedness of these summary measures.

As an example, in the application and in the simulations, we will adopt and evaluate the use of the Firth correction in the fitting of the cluster-level binomial models; this helps in reducing the finite-sample bias of the maximum likelihood estimates.

Although the unbiasedness property holds, the selection of summary measures is guided by the **researcher's specific objectives**. For instance, in clinical trials, common choices include post-treatment means or mean changes relative to baseline.

In our example, we focus on the effect of time on health conditions.

Simulations

Simulated model

$\boldsymbol{y}$ is simulated as a Bernoulli rv with the following mean:

$$\mu_{ij} = \text{logit}^{-1}(X_{ij}\beta + Z_j\gamma + U_j + D_j X_{ij})$$

where

- $\boldsymbol{Z}$ design matrix of the nuisance parameters γ with $Z_j \sim \mathcal{N}(0, 1)$
- $\boldsymbol{X}$ design matrix of the tested parameters β with $X_{ij} = O_{ij} + Z_j/2$ where $O_{ij} \sim \mathcal{N}(0, 1)$
- $\boldsymbol{U}$ defines the subject-specific random effect (i.e., random intercept) where $U_j \sim \mathcal{N}(0, 0.5)$
- $\boldsymbol{D}$ the subject-specific random slope of $\boldsymbol{X}$ where $D_j \sim \mathcal{N}(0, 0.5)$

Simulations

Setting

The aim is to test $H_0 : \beta = 0$. We compare GLMM, GEE and flip2sss in terms of:

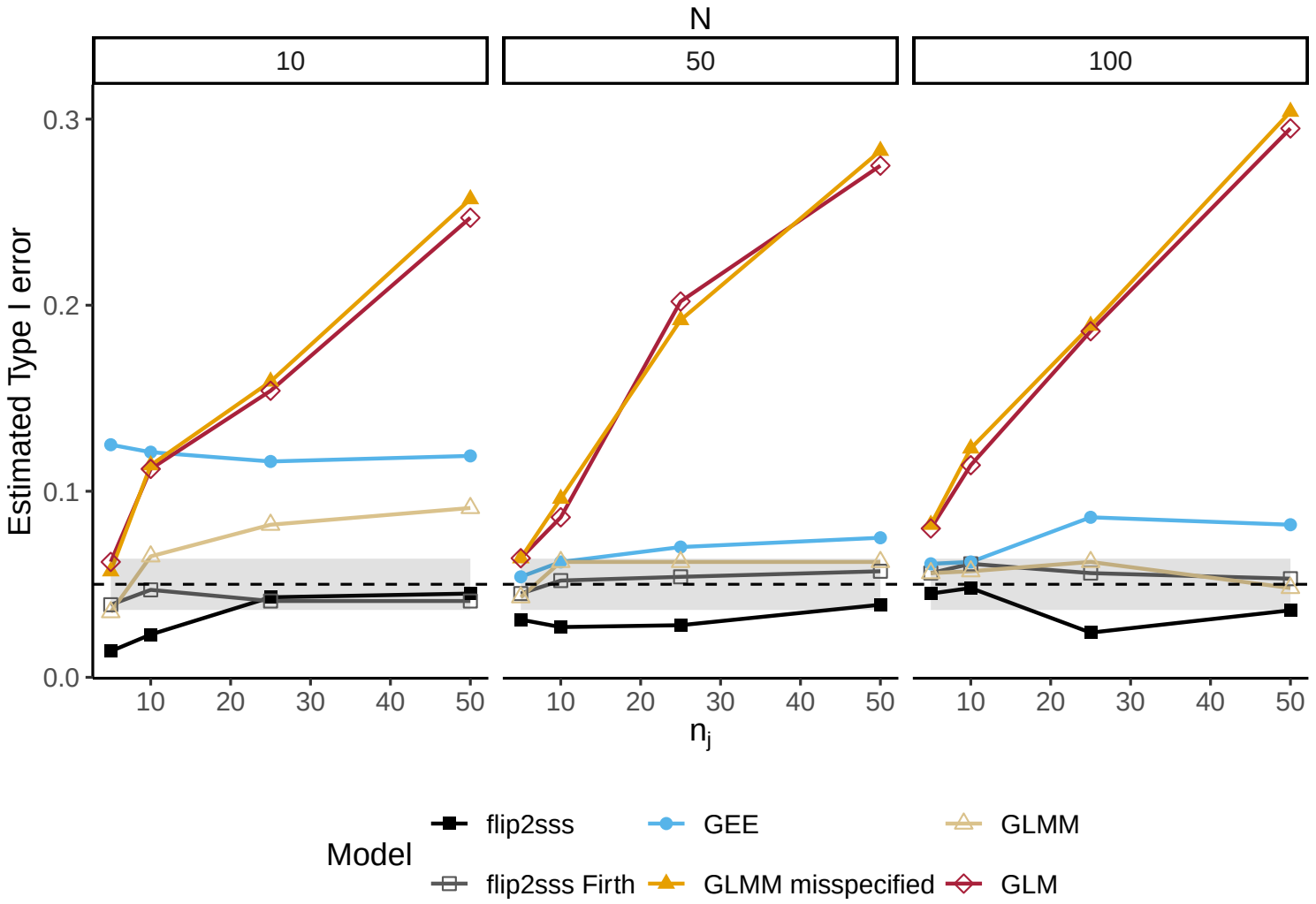
- type I error control $\Rightarrow \beta = 0, \gamma = 2$
- power $\Rightarrow \beta = 2, \gamma = 2$

In both scenarios, 1000 simulations and 100 permutations are performed.

- **flip2sss**: using MLE of GLM and Firth correction in the first stage
- **GLMM**: correctly specified (random intercept and slope) and misspecified (only random slope)
- **GEE**: independent working correlation matrix (to assure consistency) which is typically a safe choice

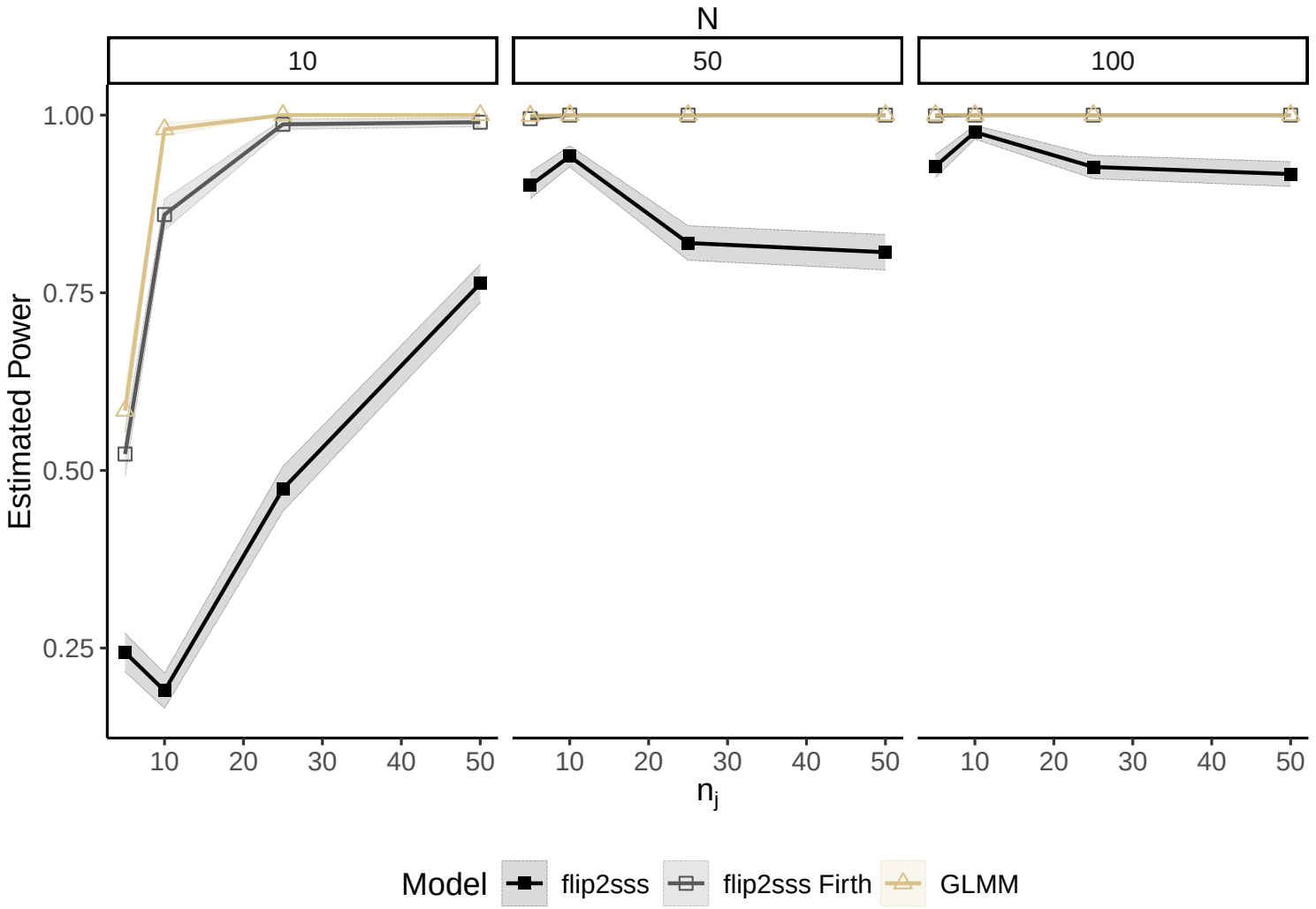
Simulations

Balanced data



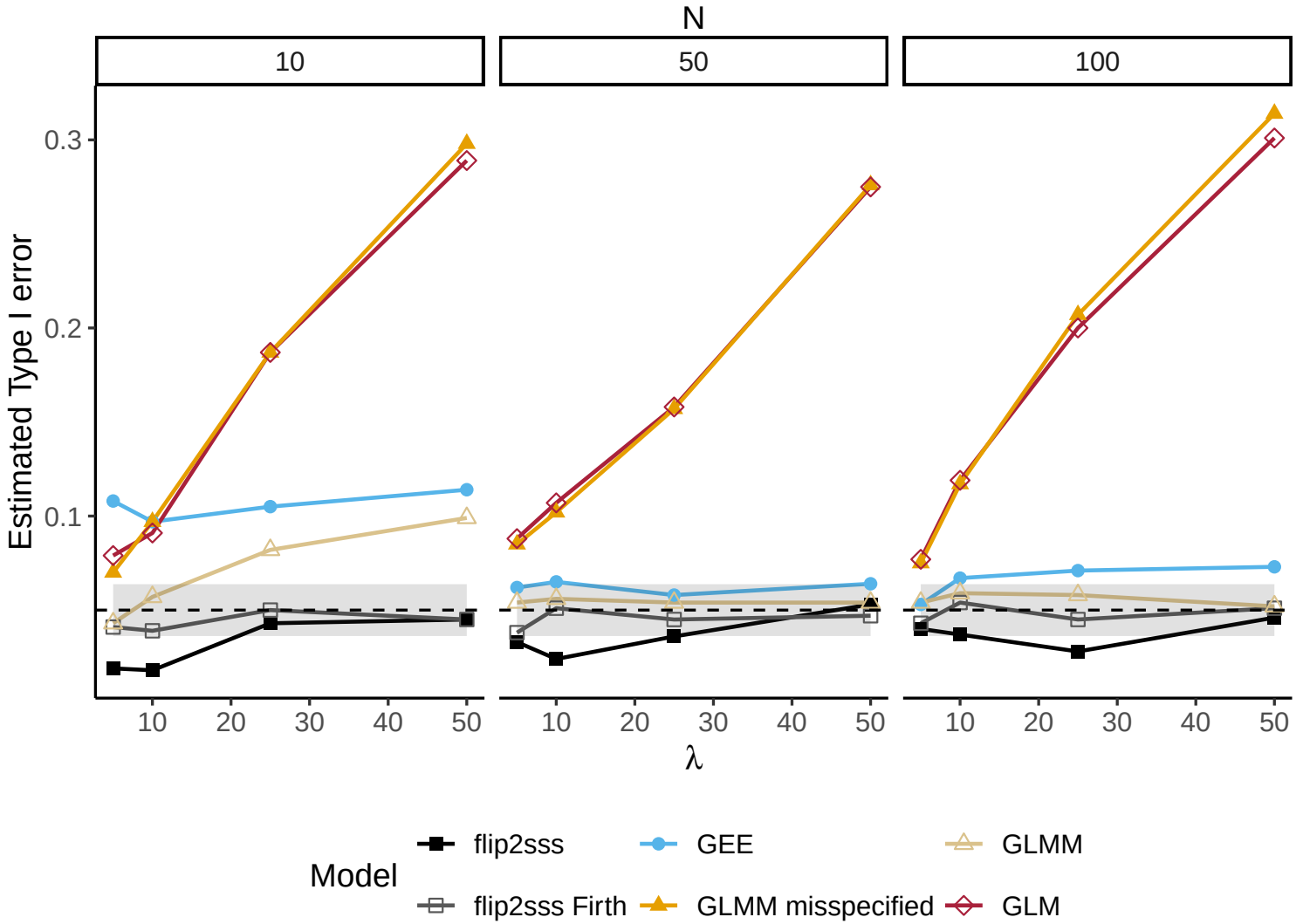
Simulations

Balanced data



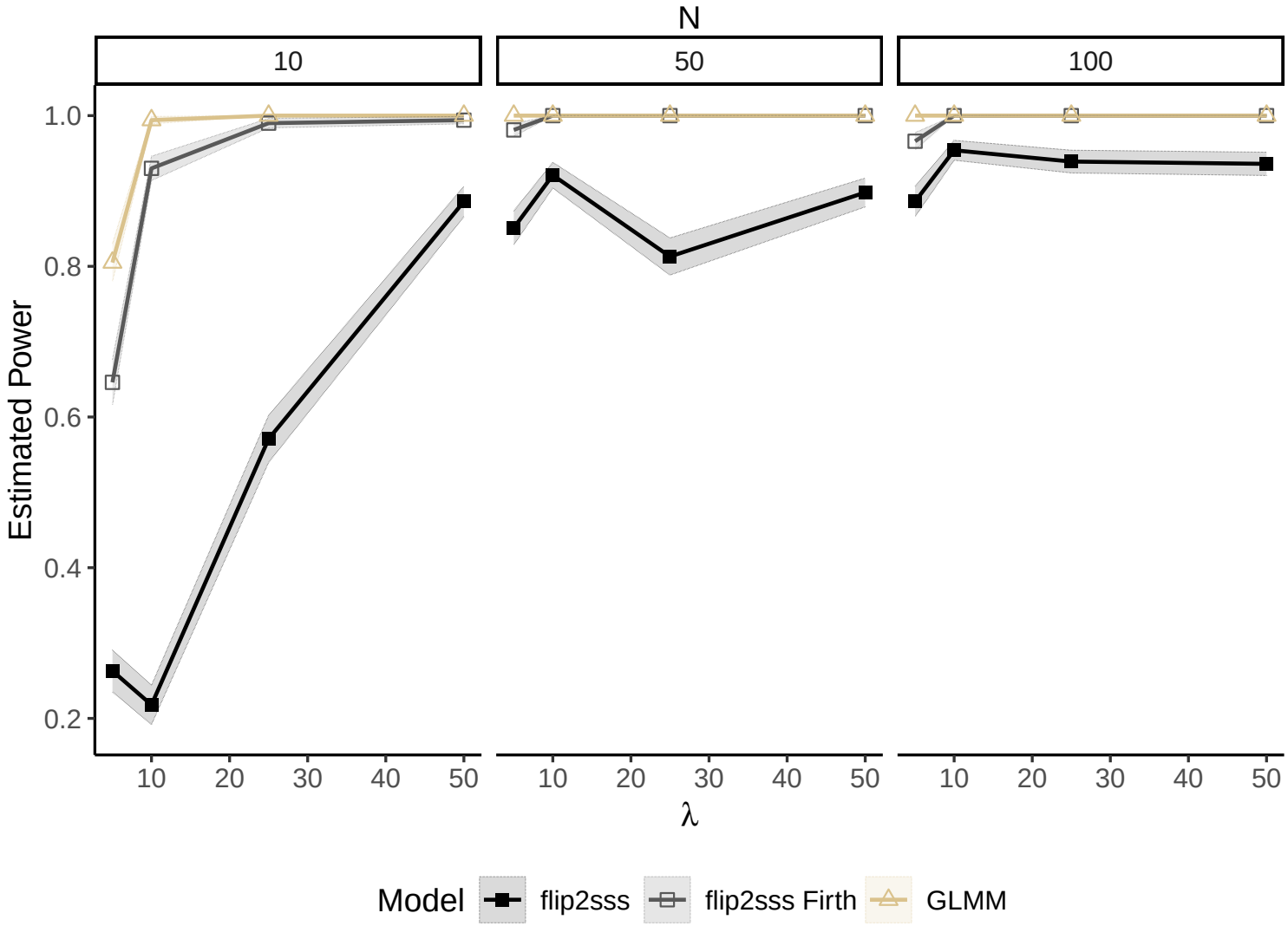
Simulations

Unbalanced data: $n_j \sim Pois(\lambda)$ with $\lambda \in \{5, 10, 25, 50\}$.



Simulations

Unbalanced data: $n_j \sim Pois(\lambda)$ with $\lambda \in \{5, 10, 25, 50\}$.



To sum up

`flip2sss`, addresses heteroscedasticity, variance misspecification, and within-subject dependence, key aspects in statistical modeling.

★ Key **advantages** include:

- No need to specify random effect structures but only the clusters.
- Flexibility in the first stage with different estimators (e.g., maximum likelihood and Firth correction).
- Easy extension to multivariate cases (multiple dependent variables).

💡 **Limitations** and **future research**:

- Bias in the first stage summary measure estimator impacts statistical power in the second stage.
- Handling more complex correlation structures, such as crossed-random effects.

Application

Let's see it in R!

Bibliography

- Diggle, P. (2002). Analysis of longitudinal data. Oxford university press.
- Andreella, A., Goeman, J., Hemerik, J., & Finos, L. (2024). Robust inference for Generalized Linear Mixed Models: a "Two-Stage Summary Statistics" approach based on score sign flipping. Psychometrika, 1-42.
- De Santis, R., Goeman, J. J., Davenport, S. J., Hemerik, J., & Finos, L. (2025). Inference in generalized linear models with robustness to misspecified variances. Journal of the American Statistical Association, 1-16.