# Beyond the Bootstrap: Interpreting Bickel and Freedman's Theorems via Superpopulation Resampling After Bayesian Latent Modeling

Seonghun Lee

## Contents

*Abstract.* Misclassification of exposure variables is a pervasive problem in observational studies and can lead to biased regression coefficients and systematically underestimated uncertainty. Bayesian latent variable models provide a principled framework for correcting misclassification by incorporating prior population information, but the sampling properties of regression estimators based on posterior quantities remain poorly understood. In this paper, we develop a superpopulation-based bootstrap framework to study the asymptotic behavior of regression inference following Bayesian latent modeling. Our approach constructs a superpopulation by repeatedly resampling the observed data and generating multiple latent exposure realizations from posterior probabilities, thereby explicitly accounting for latent-state uncertainty that is ignored by single imputation or plug-in estimators. Building on the asymptotic theory of Bickel and Freedman, we establish weak convergence results for regression coefficients, variance estimators, and residual empirical processes under this superpopulation bootstrap. In particular, we show that the conditional distribution of the bootstrap variance estimator converges to a point mass at the true error variance, and that the empirical distribution of resampled residuals converges weakly to a Brownian bridge. These results provide a theoretical justification for the dispersion of regression coefficients obtained from posterior-based resampling and explain why continuous posterior exposure probabilities yield well-calibrated variability even when individual latent classifications are imperfect. Simulation studies confirm that the proposed method recovers the correct sampling variability, avoids the underestimation of uncertainty induced by single latent realizations, and closely approximates inference based on the true unobserved exposure. Together, our findings demonstrate that superpopulation resampling after Bayesian latent modeling yields asymptotically valid and practically reliable regression inference in the presence of exposure misclassification.

# 1 Introduction

Accurately estimating regression coefficients in the presence of misclassification bias remains a fundamental challenge in statistical modeling, particularly in observational studies. Bayesian latent modeling offers a promising solution by leveraging prior information to correct for misclassification, but assessing the reliability of this approach requires rigorous evaluation. In this paper, we propose a bootstrap-based superpopulation resampling method to examine the asymptotic properties of regression coefficients estimated from Bayesian latent models, specifically focusing on the precision of the estimated exposure variable and its corresponding regression coefficients.

Our approach extends Bickel and Freedman's asymptotic results by applying them to resampled datasets drawn from a superpopulation, constructed through bootstrapping, instead of directly from initial bootstrap samples. We fit a linear regression model using information from our Bayesian latent model and extract residuals, which are appended to the original dataset for resampling. This strategy allows us to assess the variability of regression estimates and residual distributions,

enabling the evaluation of asymptotic properties, such as the convergence of variance estimates and the empirical distribution of residuals.

By adapting Bickel and Freedman's results and incorporating Mallows distance to assess distributional discrepancies, we provide a robust theoretical foundation for our approach. The convergence of the conditional distribution of variance estimates to a point mass at the true variance justifies the dispersion of regression coefficients, ensuring they reflect the true variability. Additionally, the weak convergence of the empirical process to a Brownian bridge captures the asymptotic behavior of residuals, ensuring that resampling from the empirical distribution accurately reflects true variability. These results highlights the ability of the Bayesian latent model to effectively capture the latent variable, with the uncertainty in the regression residuals closely mirroring that from the true latent variable.

The remainder of this paper is organized as follows. We first describe the methodology underlying our bootstrap residual approach, detailing the construction of resampled datasets and the estimation procedures. We then present key theoretical results supporting our framework and discuss their implications. Finally, we conclude with empirical applications and simulations that validate our theoretical findings and illustrate the practical advantages of our approach in addressing misclassification bias.

# 2 Method

## 2.1 Bayesian Latent Variable Model

Since this section is secondary to our main discussion, we provide a brief overview of our Bayesian latent model. First, we establish notations. Denote $Z$ as a binary latent variable, $A$ the observed and misclassified binary variable, $X$ observed covariates, and $\psi$ the probability that $Z = 1$ given our prior information. The data-generating process for simulation analysis is described in the Appendix. The primary goal is to estimate the probability that the true value of $Z$ is 1 given the observed covariates and prior information, $\Pr(Z_i = 1 \mid A_i, X_i, \psi_i)$. To this end, we treat $Z$ as a latent variable and fit a Bayesian latent model using the Bayesian inference program Stan, which

provides flexibility in specifying priors and fitting models, as well as high-performance statistical computation via Markov Chain Monte Carlo (MCMC) methods. We first specify the following joint model:

$$p(A_i, Z_i, \psi_i \mid X_i) = p(A_i \mid Z_i, \psi_i, X_i)\, p(Z_i \mid \psi_i, X_i)\, p(\psi_i \mid X_i),$$

which, under conditional independence assumptions, simplifies to:

$$p(A_i, Z_i, \psi_i \mid X_i) = p(A_i \mid Z_i, X_i)\, p(Z_i \mid \psi_i)\, p(\psi_i \mid X_i).$$

Here, $A$ is estimated as a function of $Z$ and $X$ (the covariates), and the prior model assumes $Z \sim \text{Bernoulli}(\psi)$. The Stan model includes the following conditional distributions:

$$\begin{aligned} A_i \mid Z_i, X_i &\sim \text{Bernoulli}(p_i), \\ Z_i \mid \psi_i &\sim \text{Bernoulli}(\psi_i), \\ \psi_i \mid X_i &\sim \text{Beta}(A_i, B_i), \end{aligned}$$

where $p_i = \text{expit}(X_i^\top \beta + \beta_9 Z_i)$ and $A_i$ and $B_i$ are specified in the prior. In this model, prior information is incorporated through the estimation of $\psi$, which follows a Beta distribution with parameters $A$ and $B$. For example, the value of $A$ is determined based on the expected (a priori) number of respondents with CPS involvement using available population information by age, race, and city. The value of $B$ corresponds to the expected number of respondents without CPS involvement. If the population prevalence of CPS involvement for a given subgroup is 0.43, the expected number of cases with CPS involvement in a sample of 3,000 is 1,287, leading to $A = 1287$ and $B = 1713$. We then marginalize over the latent variable $Z$ by summing over its possible values. Finally, the posterior probability $\Pr(Z_i = 1 \mid A_i, X_i, \psi_i)$ is computed using the estimated likelihood and prior probabilities via Bayes' rule at the final stage of our Stan model. The detailed derivation of $\Pr(Z_i = 1 \mid A_i, X_i, \psi_i)$ can be found in the Appendix.

## 2.2 Resampling from Original Population

Let $Y$ represent the outcome of interest. To generate the outcome variable for the population, we randomly sample $\varepsilon$ from a standard normal distribution. Consequently, the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independently and identically distributed according to a common distribution $F$. We then create the target distribution of a linear regression coefficient for exposure, denoted as $\beta_1$, using a resampling-based approach. Specifically, we draw repeated samples of size $n = 3000$, with replacement, from our constructed population of size 100,000. Each resampled dataset matches the size of the original sample. Using each resampled dataset, we regress the outcome $Y$ on the true exposure variable $Z$ and covariates $X$. This procedure is repeated $M$ times, resulting in $M$ estimates of $\beta_1$. Let each resampled dataset be denoted as:

$$\left\{ \left( Z_1^{*(i)}, X_1^{*(i)}, Y_1^{*(i)} \right), \ldots, \left( Z_n^{*(i)}, X_n^{*(i)}, Y_n^{*(i)} \right) \right\}, \quad i = 1, \ldots, M.$$

We fit a linear regression model to each dataset, obtaining $M$ values of the regression coefficient for the exposure variable $Z$, denoted as:

$$\left( \beta_1^{*(1)}, \beta_1^{*(2)}, \ldots, \beta_1^{*(M)} \right).$$

Note that that we do not use the hat notation because we are resampling from the original population and fitting the model using the true exposure $Z$. The empirical distribution of $\beta_1^{*(i)}$ over $i = 1, \ldots, M$ defines our target distribution, denoted by the empirical distribution function $F_m$. We summarize this distribution using its sample mean and variance:

$$\bar{\beta}_1^* = \frac{1}{M} \sum_{i=1}^{M} \beta_1^{*(i)},$$

$$\mathrm{Var}(\beta_1^*) = \frac{1}{M-1} \sum_{i=1}^{M} \left( \beta_1^{*(i)} - \bar{\beta}_1^* \right)^2.$$

This approach allows us to quantify the variability in $\beta_1$ when the true exposure $Z$ is known and used in regression. In later sections, we compare this target distribution to alternative estimators

derived from estimated binary and continuous proxy variables for $Z$.

### 2.3 Bootstrap and Superpopulation Resampleing

Now, we use continuous posterior probabilities to construct a superpopulation and employ a resampling approach to evaluate our method by comparing the estimated distribution of $\hat{\beta}_1^*$ to the target empirical distribution. Specifically, we assign each sample a binary value, denoted as $\hat{Z}$, based on the continuous posterior probabilities $\hat{Z}_{\mathrm{P}}$. Theoretically, for a sample size of $n = 3000$, the total number of possible binary combinations is $2^n$, which approximates $10^{903}$. However, the actual number of combinations is smaller, as some individuals have posterior probabilities equal to 1 (see Appendix). Nonetheless, directly sampling from this immense superpopulation is computationally prohibitive.

To address this, we first generate 200 and 400 bootstrap samples of size $n = 3000$ from our original sample, forming superpopulations of size 600,000 and 1,200,000 observations, respectively. We then repeatedly draw samples, with replacement, from each superpopulation as we did earlier to construct the target empirical distribution for $\beta_1$. We set $M$ to be five times the number of bootstrap samples, resulting in 1000 and 2000 resampled datasets of size $n = 3000$, respectively. This choice has no inferential significance but ensures that each posterior probability $\hat{Z}_{P_i}$ contributes multiple latent realizations across the resampled datasets, stabilizing the resulting empirical distribution without incurring substantial computational burden.

Next, for each sampled observation, we randomly generate new binary values $\hat{Z}$ from a Bernoulli distribution with parameter $\hat{Z}_{\mathrm{P}}$, yielding datasets of the form:

$$\left\{\left(\hat{Z}_1^{*(i)}, X_1^{*(i)}, Y_1^{*(i)}\right), \ldots, \left(\hat{Z}_n^{*(i)}, X_n^{*(i)}, Y_n^{*(i)}\right)\right\}, \quad i = 1, \ldots, M.$$

We then fit a linear regression model on each resampled dataset to estimate the regression coefficient corresponding to $\hat{Z}^*$, producing $M$ values of $\hat{\beta}_1^*$:

$$\left(\hat{\beta}_1^{*(1)}, \hat{\beta}_1^{*(2)}, \ldots, \hat{\beta}_1^{*(M)}\right).$$

Using these estimates, we construct the empirical distribution function $\hat{F}_m$ and compute the sample mean and variance of $\hat{\beta}_1^*$:

$$\bar{\hat{\beta}}_1^* = \frac{1}{M}\sum_{i=1}^{M} \hat{\beta}_1^{*(i)},$$
$$\mathrm{Var}(\hat{\beta}_1^*) = \frac{1}{M-1}\sum_{i=1}^{M} \left(\hat{\beta}_1^{*(i)} - \bar{\hat{\beta}}_1^*\right)^2.$$

**Sources of Variability Relevant to Superpopulation Construction.** The dominant source of uncertainty in estimating the regression coefficient for exposure is latent-state variability: the true exposure $Z$ is unobserved, and each subject is represented only through its posterior probability $\hat{Z}_{P_i}$. A single binary realization of $\hat{Z}$ fails to reflect this uncertainty and leads to underestimated dispersion of $\hat{\beta}_1$. For this reason, we generate many bootstrap samples and repeatedly draw binary values $\hat{Z}_i^* \sim \text{Bernoulli}(\hat{Z}_{P_i})$, allowing the superpopulation to approximate the full distribution of plausible binary configurations consistent with the posterior probabilities. Without this additional layer of resampling, estimates of $\beta_1$ would depend on only one latent configuration and would therefore exhibit artificially small standard errors, failing to capture the variability introduced by the uncertainty in the binary classification process.

Although bootstrap resampling traditionally reflects sampling variability, in our setting its main role is to provide multiple index vectors so that each $\hat{Z}_{P_i}$ yields many latent-state realizations. Nonetheless, because the procedure resamples observations with replacement, it also subtly incorporates the sampling error inherent in survey data. This secondary benefit arises even though the construction is primarily motivated by latent-state uncertainty. Model-error variability does not enter at this point. Its contribution to the variability of $\hat{\beta}_1$ is evaluated later in Section 2.5 using a residual bootstrap, which examines how uncertainty in the regression errors propagates through the estimator.

## 2.4 Residual Adjustment Using Resampled Datasets

Within the context of social science, biostatistics, and other disciplinary research, interpretability is often a key concern. Thus, using the binary variable $\hat{Z}$ is preferred as a predictor over the continuous

probabilities $\hat{Z}_P$. However, using $\hat{Z}$ leads to biased regression coefficients, as the latent modeling is not optimal for individual classification due to limited population information. Additionally, information is lost when continuous probabilities are converted into binary values. To improve interpretability—at the cost of increased standard error—we propose the following method. First, we fit linear regression models using each resampled dataset:

$$\hat{Z}^* = \hat{Z}_P^* \beta + \varepsilon_z$$

where $\hat{Z}^*$ and $\hat{Z}_P^*$ come from each resampled dataset. Next, we extract the residuals $\hat{\varepsilon}_z$ from each model and append them to the corresponding dataset as predictors, yielding the following structure:

$$\left\{ \left( \hat{Z}_1^{*(i)}, X_1^{*(i)}, Y_1^{*(i)}, \hat{\varepsilon}_{z_1}^{*(i)} \right), \ldots, \left( \hat{Z}_n^{*(i)}, X_n^{*(i)}, Y_n^{*(i)}, \hat{\varepsilon}_{z_n}^{*(i)} \right) \right\}, \quad i = 1, \ldots, M.$$

Subsequently, we regress $Y^*$ on $\hat{Z}^*$, $X^*$, and $\hat{\varepsilon}_z^*$, collecting $M$ estimates of the $\hat{Z}^*$ coefficients. This approach works because $\hat{\varepsilon}_z$ captures the variation in $\hat{Z}$ that is unexplained by $\hat{Z}_P$; hence, the resulting estimates should closely mirror those obtained from regressing $Y^*$ on $\hat{Z}_P^*$ and $X^*$.

### 2.5 Bootstrap Residuals for Asymptotic Theory

To evaluate the performance of the Bayesian latent modeling approach in estimating $\hat{Z}_P$ and the precision of the corresponding regression coefficient, we use a bootstrap and resampling strategy based on residuals. This approach applies Freedman's theorems in the context of resampled datasets from a superpopulation, rather than bootstrapped samples from a single observed dataset. We first fit the following linear regression model using the original sample:

$$Y = X\beta + \varepsilon_{z_p}$$

where $X$ includes $\hat{Z}_P$ along with other covariates. We use the notation $\varepsilon_{z_p}$ to emphasize that the regression model is fit using the continuous posterior probabilities $\hat{Z}_P$ as the exposure variable, rather than a single binary realization.

Next, we extract the residuals $\hat{\varepsilon}_{z_p}$ and append them to the original sample. The bootstrap is then applied to this sample, forming a superpopulation and generating resampled datasets as described in the previous section. The resulting data structure is:

$$\left\{\left(X_1^{*(i)}, \hat{\varepsilon}_{z_p,1}^{*(i)}\right), \ldots, \left(X_n^{*(i)}, \hat{\varepsilon}_{z_p,n}^{*(i)}\right)\right\}, \quad i = 1, \ldots, M.$$

Each resampled dataset is then used to generate the starred outcome using the regression coefficient estimate $\hat{\beta}$ from the original model:

$$Y^* = X^*\hat{\beta} + \hat{\varepsilon}_{z_p}^*$$

We then estimate the regression coefficients from the generated data:

$$\hat{\beta}^* = \left(X^{*\top}X^*\right)^{-1} X^{*\top}Y^*$$

Next, we compute the starred residuals using $Y^*$ and $\hat{\beta}^*$:

$$\hat{\varepsilon}_{z_p}^\star = Y^* - X^*\hat{\beta}^*$$

The purpose of estimating the starred residuals is to assess the performance of $\hat{Z}_P$ relative to the true latent variable $Z$ by analyzing the asymptotic behavior of $\hat{\varepsilon}^*$ using the framework established by Bickel and Freedman (Bickel and Freedman, 1981; Freedman, 1981). Specifically, we examine whether the conditional distribution of $\hat{\sigma}_n^*$ converges to a point mass at $\sigma$, where $\hat{\sigma}_n^*$ is the variance of $\hat{\varepsilon}_{z_p}^*$ and $\sigma$ is the variance of $\varepsilon$. Additionally, we investigate whether the asymptotic distribution of the empirical distribution function of $\hat{\varepsilon}_{z_p}^*$ converges to a Brownian bridge. The next section revisits these theoretical results and explains their relevance in our context.

## 3 Theoretical Results and Inference

Since the fitted coefficients $\beta_1^*$ and $\hat{\beta}_1^* -$ "Bias" are approximately normally distributed around the population value of $\beta_1$, there exists $\sigma_m^2 \approx \hat{\sigma}_m^2$ such that

$$\sqrt{m}(\beta_1^* - \beta_1) \xrightarrow{D} N(0, \sigma_m^2), \quad \sqrt{m}(\hat{\beta}_1^* - \beta_1) \xrightarrow{D} N(\text{"Bias"}, \hat{\sigma}_m^2),$$

where $\text{Var}(\beta_1^*) = \sigma_m^2$ and $\text{Var}(\hat{\beta}_1^*) = \hat{\sigma}_m^2$. The rationale for assuming $\sigma_m^2 \approx \hat{\sigma}_m^2$ will be discussed shortly. Recall that $m$ is the number of resampled datasets.

Now, we assume that the constructed empirical distribution function $F_m$ converges almost surely to $F$ by the strong law of large numbers, the cumulative distribution function of $\beta_1$. This is reasonable considering that the true latent variable $Z$ was used to construct the target distribution $F_m$. Then, if the Bayesian latent model gives perfect correction for every individual, $\hat{F}_m$ also converges almost surely to $F$ by the strong law of large numbers. Although this condition may appear challenging to satisfy, there exists only a location shift between $F_m$ and $\hat{F}_m$, even in the presence of misclassified cases of $\hat{Z}$. This outcome arises from the relationship that $\text{Var}(\beta_1^*) \approx \text{Var}(\hat{\beta}_1^*)$. These insights lead us to the following theorem:

**Theorem 1.** Let $\beta_1^*$ and $\hat{\beta}_1^*$ denote estimators obtained from random resamples drawn from a superpopulation, and let $F$ denote the distribution of $\beta_1$. If $\beta_1^*$ and $\hat{\beta}_1^*$ are unbiased estimators of $\beta_1$ and $\text{Var}(\beta_1^*|F_m)/\text{Var}(\hat{\beta}_1^*|\hat{F}_m) \approx 1$, then

$$\|\hat{F}_m - F_m\|_\infty \xrightarrow{a.s.} 0.$$

A proof of Theorem 1 can be established using the idea of the Glivenko-Cantelli theorem, as outlined in the Appendix.

**Theorem 2.** Let $\hat{Z}_P$ represent continuous posterior probabilities and $Z$ denote the i.i.d. true

binary exposure.

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Z}_{P_i} \approx \frac{1}{n}\sum_{i=1}^{n} Z_i \quad \Rightarrow \quad \sqrt{\frac{\mathrm{Var}(\beta_1^* \mid F_m)}{\mathrm{Var}(\hat{\beta}_1^* \mid \hat{F}_m)}} \approx 1.$$

This theorem suggests that a variance ratio of 1 can be achieved if the mean of the estimated continuous probabilities closely approximates the mean of the true binary exposure, even if individual misclassifications are not fully corrected. This result is influenced by the fact that the standard error of the linear regression coefficient is inversely proportional to the sum of squares of the predictor values, which appears in the denominator of its formula. If $\hat{Z}_P \approx Z$, then the proportion of 0's and 1's in the binary variable $\hat{Z}$ will be approximately the same as in $Z$, and hence

$$\sum_{i=1}^{n}(\hat{Z}_i - \hat{Z}_P)^2 \approx \sum_{i=1}^{n}(Z_i - Z)^2.$$

Consequently, when estimating $\beta_1^*$ and $\hat{\beta}_1^*$ from bootstrap samples, their variability remains very similar. It is important to note that $\hat{Z}_P \approx Z$ is ensured in our Bayesian latent modeling approach, which leverages prior information from the population data. The proof is in the Appendix.

Now, we revisit key results from Bickel and Freedman, adapting them to our specific context and providing a more detailed explanation to highlight their relevance. While the original results were established in the context of bootstrapped samples, our focus is on their implications for resampled datasets drawn from a superpopulation constructed through bootstrapping.

**Mallows distance.** Mallows (1972) introduced a metric on the space of probability distributions. We employ this metric to quantify the discrepancy between two measures. Let $u$ and $v$ be probability measures in $\mathbb{R}^p$, and we define the Mallows distance as the infimum of $(\mathbb{E}\|U - V\|^r)^{1/r}$ over all pairs of random vectors $U$ and $V$ with laws $u$ and $v$, respectively.

**Lemma 1.** Assume that $(\varepsilon_i)_{i\geq 1}$ are independent with $E(\varepsilon_i) = 0$, $(\varepsilon_i) = \sigma^2 < \infty$, independent of $(X_i)_{i\geq 1}$, and that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^\top X_i = O_p(1).$$

Then

$$\frac{1}{n}X^\top \varepsilon = \frac{1}{n}\sum_{i=1}^{n} X_i^\top \varepsilon_i \xrightarrow{\text{a.s.}} 0.$$

**Proof.** Let $S_n = \sum_{i=1}^{n} X_i^\top \varepsilon_i$. We first prove that $S_{n_m}/n_m \to 0$ almost surely along the dyadic subsequence $n_m = 2^m$, and then extend to all $n$.

Fix $\lambda > 0$ and consider $n_m = 2^m$. Using Kolmogorov's inequality conditionally on $(X_i)_{i\geq 1}$, we have

$$P\left(\max_{1\leq k\leq n_m} |S_k| \geq \lambda n_m \,\middle|\, (X_i)_{i\geq 1}\right) \leq \frac{(S_{n_m} \mid (X_i))}{\lambda^2 n_m^2} = \frac{\sigma^2 \sum_{i=1}^{n_m} X_i^\top X_i}{\lambda^2 n_m^2}.$$

Taking expectations and using the tower property,

$$P\left(\max_{1\leq k\leq n_m} |S_k| \geq \lambda n_m\right) \leq \frac{\sigma^2}{\lambda^2} E\left(\frac{1}{n_m^2}\sum_{i=1}^{n_m} X_i^\top X_i\right).$$

By the assumption $\frac{1}{n}\sum_{i=1}^{n} X_i^\top X_i = O_p(1)$ and standard arguments (e.g. uniform integrability or a moment bound), there exists a constant $C < \infty$ such that for all $m$,

$$E\left(\frac{1}{n_m^2}\sum_{i=1}^{n_m} X_i^\top X_i\right) \leq \frac{C}{n_m}.$$

Hence

$$P\left(\max_{1\leq k\leq n_m} |S_k| \geq \lambda n_m\right) \leq \frac{C\sigma^2}{\lambda^2}\frac{1}{n_m} = \frac{C'}{2^m},$$

for some constant $C'$. Therefore

$$\sum_{m=1}^{\infty} P\left(\max_{1\leq k\leq n_m} |S_k| \geq \lambda n_m\right) < \infty.$$

By the Borel–Cantelli lemma,

$$\max_{1\leq k\leq n_m} \frac{|S_k|}{n_m} \xrightarrow{\text{a.s.}} 0, \quad \text{and in particular} \quad \frac{S_{n_m}}{n_m} \xrightarrow{\text{a.s.}} 0.$$

Now let $n$ be arbitrary. Define $m = \lfloor \log_2 n \rfloor$, so that $m \leq \log_2 n < m + 1$. Exponentiating with

base 2 gives $2^m \le n < 2^{m+1}$. Hence we can choose $m$ such that $n_m = 2^m \le n < 2^{m+1} = n_{m+1}$. Then

$$\frac{S_n}{n} = \frac{S_{n_m}}{n} + \frac{1}{n}\sum_{i=n_m+1}^{n} X_i^\top \varepsilon_i.$$

For the first term,

$$\left|\frac{S_{n_m}}{n}\right| = \frac{n_m}{n}\left|\frac{S_{n_m}}{n_m}\right| \le \left|\frac{S_{n_m}}{n_m}\right| \xrightarrow{\text{a.s.}} 0,$$

since $n_m/n \le 1$ and $S_{n_m}/n_m \to 0$ a.s.

For the second term, define

$$T_n = \sum_{i=n_m+1}^{n} X_i^\top \varepsilon_i.$$

By the same conditional Kolmogorov argument as above, with $n$ in place of $n_m$ and using $n < 2n_m$, we obtain

$$P\left(\left|\frac{T_n}{n}\right| \ge \lambda\right) \le \frac{C''}{n}$$

for some constant $C''$ (details are analogous and omitted for brevity). Thus $\sum_{n=1}^{\infty} P(|T_n/n| \ge \lambda) < \infty$ and another application of Borel–Cantelli yields $T_n/n \to 0$ almost surely. Combining the two parts gives

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} 0.$$

Since $S_n = X^\top \varepsilon$, this proves the lemma.

**Lemma 2.** Assuming $\frac{1}{n}X^T X$ is positive definite,

$$\frac{1}{n}\|\hat{\epsilon}_{(z_P)} - \epsilon\|^2 \xrightarrow{a.s.} 0.$$

**Proof.** Use $\frac{1}{n}\left\|\hat{\varepsilon}_{(Z_P)} - \varepsilon\right\|^2 = \varepsilon^\top X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top \varepsilon$,

$$\frac{1}{n}\left\|\hat{\varepsilon}_{(Z_P)} - \varepsilon\right\|^2 = \left[\frac{1}{n}\varepsilon^\top X\right]\left[\frac{1}{n}X^\top X\right]^{-1}\left[\frac{1}{n}X^\top \varepsilon\right].$$

By assumption and Lemma 1, the proof is complete.

**Lemma 3.** Assume $\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_{z_p i}=0$. Define the empirical distributions

$$\hat{F}_n=\frac{1}{n}\sum_{i=1}^{n}\delta_{\hat{\varepsilon}_{z_p i}},\qquad F_n=\frac{1}{n}\sum_{i=1}^{n}\delta_{\varepsilon_i},$$

where $\delta_x$ denotes the Dirac point mass at $x$. Then

$$d_2(\hat{F}_n,F_n)\xrightarrow{\text{a.s.}}0,$$

where $d_2$ denotes the 2-Wasserstein distance on $\mathbb{R}$.

**Proof.** Define the coupling

$$\pi_n=\frac{1}{n}\sum_{i=1}^{n}\delta_{(\hat{\varepsilon}_{z_p i},\,\varepsilon_i)},$$

whose first and second marginals are $\hat{F}_n$ and $F_n$, respectively.

Then $\pi_n\in\Pi(\hat{F}_n,F_n)$ and

$$\int(x-y)^2\,d\pi_n(x,y)=\frac{1}{n}\sum_{i=1}^{n}(\hat{\varepsilon}_{z_p i}-\varepsilon_i)^2=\frac{1}{n}\|\hat{\varepsilon}_{z_p}-\varepsilon\|^2.$$

By definition of $d_2$ as an infimum over couplings,

$$d_2(\hat{F}_n,F_n)^2\le\int(x-y)^2\,d\pi_n(x,y)=\frac{1}{n}\|\hat{\varepsilon}_{z_p}-\varepsilon\|^2.$$

Lemma 2 implies $\frac{1}{n}\|\hat{\varepsilon}_{z_p}-\varepsilon\|^2\xrightarrow{\text{a.s.}}0$, and hence $d_2(\hat{F}_n,F_n)\xrightarrow{\text{a.s.}}0$.

**Lemma 4.** $d_2(\hat{F}_n,F)\xrightarrow{\text{a.s.}}0$.

**Proof.** By Lemma 8.4 of Bickel and Freedman (1981),

$$d_2(F_n,F)\xrightarrow{\text{a.s.}}0.$$

Then by Lemma 3 and the triangle inequality,

$$d_2(\hat{F}_n, F) \leq d_2(\hat{F}_n, F_n) + d_2(F_n, F) \xrightarrow{\text{a.s.}} 0.$$

**Lemma 5.** Let $\mu_i$ and $v_i$ be real numbers. Define

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i, \quad s_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2,$$

and similarly define $\bar{v}$ and $s_v^2$ for $v_i$. Then,

$$(s_\mu - s_v)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(\mu_i - v_i)^2.$$

**Proof.** Note that $s_\mu = \frac{\|\mu - \bar{\mu}\|}{\sqrt{n}}$ and $s_v = \frac{\|v - \bar{v}\|}{\sqrt{n}}$, so

$$\begin{aligned}(s_\mu - s_v)^2 &\leq \frac{1}{n}\left\|(\mu - \bar{\mu}) - (v - \bar{v})\right\|^2 \\ &= \frac{1}{n}\left(\|\mu - v\|^2 - n(\bar{\mu} - \bar{v})^2\right) \\ &= \frac{1}{n}\|\mu - v\|^2.\end{aligned}$$

For the next three theorems, we assume that $\hat{Z}_P^*$ closely approximates the information of the unobserved true latent variable $Z$ within the regression model framework discussed in the section on bootstrap residuals for asymptotic theory. Additionally, note that the weak convergence results of these theorems hold for all sample sequences except for a set of sequences with probability zero under the true data-generating process. In the following theorem, recall the starred data generated using bootstrapped residuals as discussed in Section 2.5:

$$Y^* = X^*\hat{\beta} + \hat{\epsilon}^*_{z_p}, \quad \hat{\beta}^* = \left((X^*)^\top X^*\right)^{-1}(X^*)^\top Y^*, \quad \hat{\epsilon}^\star_{z_p} = Y^* - X^*\hat{\beta}^*$$

**Theorem 3.** Assuming $V = \frac{1}{n}X^\top X$ is positive definite, given the i.i.d. $Y_1, \ldots, Y_n$, the conditional distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ converges weakly to $N(0, \sigma^2 V^{-1})$.

**Proof.** Let $\psi_n(F)$ be the law of $\sqrt{n}(\hat{\beta} - \beta)$ and $\psi_n(\hat{F}_n)$ the law of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$, where $F$ is the common distribution of $\varepsilon_1, \ldots, \varepsilon_n$ and $\hat{F}_n$ is the (centered) empirical distribution of the bootstrap residuals $\hat{\varepsilon}^*_{z_p 1}, \ldots, \hat{\varepsilon}^*_{z_p n}$. Conditionally on the design matrix $X$, the bootstrap estimator satisfies

$$\sqrt{n}(\hat{\beta}^* - \hat{\beta}) = A_n \varepsilon^*, \qquad A_n := \sqrt{n}\big((X^*)^\top X^*\big)^{-1}(X^*)^\top,$$

where $\varepsilon^*$ has i.i.d. coordinates with common law $\hat{F}_n$.

By Theorem 2.1 of Freedman (1981), for any two error laws $G_1, G_2$ on $\mathbb{R}$, if $\psi_n(G_k)$ denotes the law of $A_n \varepsilon^{(k)}$, where the coordinates of $\varepsilon^{(k)}$ are i.i.d. with distribution $G_k$, then

$$d_2\big(\psi_n(G_1), \psi_n(G_2)\big)^2 \;\le\; \operatorname{trace}(A_n A_n^\top)\, d_2(G_1, G_2)^2,$$

Applying this with $G_1 = \hat{F}_n$ and $G_2 = F$ gives

$$d_2\big(\psi_n(\hat{F}_n), \psi_n(F)\big)^2 \;\le\; \operatorname{trace}(A_n A_n^\top)\, d_2(\hat{F}_n, F)^2.$$

Next compute the trace factor:

$$A_n A_n^\top = n\big((X^*)^\top X^*\big)^{-1}(X^*)^\top X^* \big((X^*)^\top X^*\big)^{-1} = n\big((X^*)^\top X^*\big)^{-1},$$

so

$$\operatorname{trace}(A_n A_n^\top) = \operatorname{trace}\left(\left(\tfrac{1}{n}(X^*)^\top X^*\right)^{-1}\right).$$

Let $V := \frac{1}{n} X^\top X$ and $V_n^* := \frac{1}{n}(X^*)^\top X^*$. By assumption $V$ is positive definite. Since $X^*$ is obtained by resampling the rows of $X$ with replacement, $V_n^*$ converges (conditionally on $X$) to $V$ almost surely, and is itself positive definite for all large $n$. Hence the eigenvalues of $V_n^*$ are bounded away from zero and infinity, and therefore

$$\sup_n \operatorname{trace}\big((V_n^*)^{-1}\big) < \infty \quad \text{a.s.}$$

That is, $\operatorname{trace}(A_n A_n^\top)$ is almost surely bounded.

By Lemma 4 we have $d_2(\hat{F}_n, F) \xrightarrow{\text{a.s.}} 0$, so the preceding inequality yields

$$d_2\big(\psi_n(\hat{F}_n), \psi_n(F)\big) \xrightarrow{\text{a.s.}} 0.$$

Finally, by the usual regression asymptotics under $V$ positive definite, $\psi_n(F)$ converges weakly to $N(0, \sigma^2 V^{-1})$. Combining these two facts shows that, conditionally on $Y_1, \ldots, Y_n$, the distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ converges weakly to $N(0, \sigma^2 V^{-1})$, which proves the theorem.

Having established the asymptotic validity of the bootstrap estimator $\hat{\beta}^*$, we now turn to the behavior of the associated variance estimators. In particular, we consider the empirical variance of the estimated residuals and of the bootstrap residuals, defined respectively as

$$\begin{aligned}
\hat{\sigma}_n &= \frac{1}{n}\sum_{i=1}^n \big(\hat{\varepsilon}_{z_{p_i}}\big)^2 - \left(\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_{z_{p_i}}\right)^2, \\
\hat{\sigma}_n^* &= \frac{1}{n}\sum_{i=1}^n \left(\hat{\varepsilon}^*_{z_{p_i}}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}^*_{z_{p_i}}\right)^2, \\
\hat{\sigma}_n^\star &= \frac{1}{n}\sum_{i=1}^n \left(\hat{\varepsilon}^\star_{z_{p_i}}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}^\star_{z_{p_i}}\right)^2.
\end{aligned}$$

**Theorem 4.** Given $Y_1, \ldots, Y_n$, the conditional distribution of $\hat{\sigma}_n^\star$ converges to a point mass at $\sigma^2$.

**Proof.**

**Step 1: Consistency of $\hat{\sigma}_n$.** By Lemma 5,

$$(\hat{\sigma}_n - \sigma_n)^2 \;\le\; \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_{z_{p_i}} - \varepsilon_i\right)^2 \xrightarrow{\text{a.s.}} 0.$$

By Lemma 2, $\sigma_n = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right)^2 \xrightarrow{\text{a.s.}} \sigma^2$. Thus,

$$\hat{\sigma}_n \xrightarrow{\text{a.s.}} \sigma^2.$$

**Step 2: Comparing the starred and resampled estimators.** Using Jensen's inequality,

$$\mathbb{E}[|\hat{\sigma}_n^\star - \hat{\sigma}_n^*|\,|\,Y_1,\ldots,Y_n]^2 \;\le\; \mathbb{E}\Big[(\hat{\sigma}_n^\star - \hat{\sigma}_n^*)^2 \,\Big|\, Y_1,\ldots,Y_n\Big].$$

Apply Lemma 5 to the sequences $\left(\hat{\varepsilon}^\star_{z_{p_i}}\right)$ and $\left(\hat{\varepsilon}^*_{z_{p_i}}\right)$:

$$(\hat{\sigma}_n^\star - \hat{\sigma}_n^*)^2 \;\le\; \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}^\star_{z_{p_i}} - \hat{\varepsilon}^*_{z_{p_i}}\right)^2.$$

Conditional on $Y_1,\ldots,Y_n$, the vector $\hat{\varepsilon}^\star_{z_p}$ is obtained by a linear transformation of $\hat{\varepsilon}^*_{z_p}$ with hat matrix $H^* = X^*(X^{*\top}X^*)^{-1}X^{*\top}$. Thus,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}^\star_{z_{p_i}} - \hat{\varepsilon}^*_{z_{p_i}}\right)^2 \,\middle|\, Y_1,\ldots,Y_n\right] = \frac{\hat{\sigma}_n}{n}\,\operatorname{trace}(H^*) = \frac{\hat{\sigma}_n p}{n}.$$

Since $\hat{\sigma}_n \xrightarrow{\text{a.s.}} \sigma^2$, we conclude

$$\hat{\sigma}_n^\star - \hat{\sigma}_n^* \xrightarrow{\text{a.s.}} 0.$$

**Step 3: Convergence of the law of $\hat{\sigma}_n^*$.** By Lemma 4,

$$d_2(\hat{F}_n, F) \xrightarrow{\text{a.s.}} 0,$$

where $\hat{F}_n$ is the empirical distribution of $\hat{\varepsilon}_{z_{p_1}},\ldots,\hat{\varepsilon}_{z_{p_n}}$ and $F$ is the common distribution of $\varepsilon_1,\ldots,\varepsilon_n$. This is the condition required to apply Lemma 8.5 of Bickel and Freedman (1981).

Taking $p = 2$, $p' = 1$, and $\varphi(\varepsilon) = \varepsilon^2$, Lemma 8.5 then yields

$$d_1\Big(\mathcal{L}\big((\hat{\varepsilon}^*_{z_{p_i}})^2\big), \mathcal{L}(\varepsilon_i^2)\Big) \xrightarrow{\text{a.s.}} 0.$$

Next, by Lemma 8.6, applied to $U_j = (\hat{\varepsilon}^*_{z_{p_j}})^2$ and $V_j = \varepsilon_j^2$, we obtain

$$d_1\left(\mathcal{L}\Big(\frac{1}{n}\sum_{i=1}^n (\hat{\varepsilon}^*_{z_{p_i}})^2\Big), \mathcal{L}\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\Big)\right) \le \frac{1}{n}\sum_{i=1}^n d_1\Big(\mathcal{L}\big((\hat{\varepsilon}^*_{z_{p_i}})^2\big), \mathcal{L}(\varepsilon_i^2)\Big) \xrightarrow{\text{a.s.}} 0.$$

Hence, conditional on $Y_1, \ldots, Y_n$, the law of

$$\frac{1}{n}\sum_{i=1}^n \left(\hat{\varepsilon}^*_{z_{p_i}}\right)^2$$

is asymptotically identical to the unconditional law of $\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2$, which converges to $\sigma^2$. Similarly,

$$\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}^*_{z_{p_i}} \xrightarrow{\text{a.s.}} 0.$$

Combining the two components yields

$$\hat{\sigma}_n^* \xrightarrow{\mathcal{L}|Y_1,\ldots,Y_n} \sigma^2.$$

**Step 4: Concluding for $\hat{\sigma}_n^\star$.** From Step 2,

$$\hat{\sigma}_n^\star - \hat{\sigma}_n^* \xrightarrow{\text{a.s.}} 0,$$

and from Step 3,

$$\hat{\sigma}_n^* \xrightarrow{\mathcal{L}|Y_1,\ldots,Y_n} \sigma^2.$$

Thus, by Slutsky's lemma applied conditionally,

$$\hat{\sigma}_n^\star \xrightarrow{\mathcal{L}|Y_1,\ldots,Y_n} \sigma^2.$$

This establishes that the conditional distribution of $\hat{\sigma}_n^\star$ converges to a point mass at $\sigma^2$.

For the next theorem, recall that $\hat{F}_n$ is the common distribution of $\hat{\varepsilon}^*_{z_{p_1}}, \ldots, \hat{\varepsilon}^*_{z_{p_n}}$, and the unknown distribution function $F$ is to be estimated by $\hat{F}_n$. Denote $\hat{F}_n^*$ as the empirical distribution function of $\hat{\varepsilon}^*_{z_{p_1}}, \ldots, \hat{\varepsilon}^*_{z_{p_n}}$ and $\hat{F}_n^\star$ as the empirical distribution function of $\hat{\varepsilon}^\star_{z_{p_1}}, \ldots, \hat{\varepsilon}^\star_{z_{p_n}}$. Additionally, let $\varphi_n(F)$ denote the law of $\sqrt{n}(\hat{F}_n - F)$, $\varphi_{n_1}(\hat{F}_n)$ the law of $\sqrt{n}(\hat{F}_n^* - \hat{F}_n)$, and $\varphi_{n_2}(\hat{F}_n)$ the law of $\sqrt{n}(\hat{F}_n^\star - \hat{F}_n)$.

**Theorem 5.** Given $\hat{\varepsilon}_{z_{p_1}}, \ldots, \hat{\varepsilon}_{z_{p_n}}$, $\sqrt{n}(\hat{F}_n^\star - \hat{F}_n)$ converges weakly to a Brownian bridge $B(F)$.

**Proof.** By the Glivenko–Cantelli theorem,

$$\|\hat{F}_n - F\|_\infty \xrightarrow{\text{a.s.}} 0,$$

and by Donsker's theorem,

$$\varphi_n(F) = \mathcal{L}\Big(\sqrt{n}(\hat{F}_n - F)\Big) \Rightarrow \mathcal{L}(B(F)).$$

Theorem 4.1 of Bickel and Freedman (1981) then implies that the conditional law of the standard residual bootstrap satisfies

$$\varphi_{n_1}(\hat{F}_n) = \mathcal{L}\Big(\sqrt{n}(\hat{F}_n^* - \hat{F}_n) \,\Big|\, Y_1, \ldots, Y_n\Big) \Rightarrow \mathcal{L}(B(F)).$$

Thus, to transfer this limit to the superpopulation bootstrap, it suffices to show

$$d_2(\hat{F}_n^\star, \hat{F}_n^*) \xrightarrow{\text{a.s.}} 0.$$

Consider the coupling that pairs $\hat{\varepsilon}^\star_{z_{p_i}}$ with $\hat{\varepsilon}^*_{z_{p_i}}$ for each $i = 1, \ldots, n$. Using the definition of the $d_2$ distance,

$$d_2^2(\hat{F}_n^\star, \hat{F}_n^*) \leq \frac{1}{n}\sum_{i=1}^n \left(\hat{\varepsilon}^\star_{z_{p_i}} - \hat{\varepsilon}^*_{z_{p_i}}\right)^2 = \frac{1}{n}\left\|\hat{\varepsilon}^\star_{z_p} - \hat{\varepsilon}^*_{z_p}\right\|^2.$$

From the proof of Theorem 4, the conditional distribution of $\hat{\sigma}_n^\star$ is nearly a point mass at $\sigma^2$.

Repeating the arguments of Lemmas 1–3 along the superpopulation bootstrap then yields

$$\frac{1}{n}\left\|\hat{\varepsilon}^{\star}_{z_p} - \hat{\varepsilon}^{*}_{z_p}\right\|^2 \xrightarrow{\text{a.s.}} 0.$$

Consequently,

$$d_2(\hat{F}^{\star}_n, \hat{F}^{*}_n) \xrightarrow{\text{a.s.}} 0.$$

Finally, consider the map

$$T_n(G) = \sqrt{n}\,(G - \hat{F}_n).$$

This map is Lipschitz under the Wasserstein distance $d_2$, since

$$d_2(T_n(G_1), T_n(G_2)) = \sqrt{n}\, d_2(G_1, G_2).$$

Therefore $T_n$ is continuous, and by the continuous mapping theorem,

$$d_2(\hat{F}^{\star}_n, \hat{F}^{*}_n) \to 0 \quad \Longrightarrow \quad d_2\Big(\sqrt{n}(\hat{F}^{\star}_n - \hat{F}_n), \sqrt{n}(\hat{F}^{*}_n - \hat{F}_n)\Big) \to 0.$$

Since $\varphi_{n_1}(\hat{F}_n) \Rightarrow \mathcal{L}(B(F))$, the continuous mapping theorem further implies

$$\varphi_{n_2}(\hat{F}_n) = \mathcal{L}\Big(\sqrt{n}(\hat{F}^{\star}_n - \hat{F}_n)\Big) \Rightarrow \mathcal{L}(B(F)).$$

This proves the theorem.

Theorems 3 and 4 jointly clarify the asymptotic behavior of the superpopulation bootstrap under our Bayesian latent-variable framework. The convergence of the conditional distribution of $\hat{\sigma}^{\star}_n$ to a point mass at $\sigma$ implies that the resampled residuals become increasingly concentrated around the true residual variance. This stabilization of the residual variance, together with its connection to the dispersion of the resampled coefficients $\hat{\beta}^{\star}$, shows that the bootstrap replicates the correct sampling variability of the regression estimator. In particular, the behavior

$$\hat{\beta}^{\star} \xrightarrow{d} \mathcal{N}\Big(\hat{\beta},\, \sigma^2 (X^{\top}X)^{-1}\Big)$$

indicates that the superpopulation-based resampling scheme accurately mimics the large-sample distribution of $\hat{\beta}$. Together, these results justify the precision of $\hat{\beta}^\star$ and confirm that constructing a superpopulation from bootstrapped samples calibrates the coefficient variability correctly. They further demonstrate that the Bayesian latent model captures the underlying exposure variable $Z$ effectively, since the uncertainty in the residuals formed using $\hat{Z}_P^\star$ closely aligns with what would be obtained if the true $Z$ were observed.

Theorem 5 provides a complementary insight by establishing the asymptotic behavior of the empirical distribution of the residuals. The weak convergence

$$\sqrt{n}\,(\hat{F}_n^\star - \hat{F}_n) \;\Rightarrow\; B(F)$$

shows that the bootstrap fluctuations of the empirical distribution mimic those of the true residual distribution, where $B(F)$ is the limiting Brownian bridge from classical empirical process theory. Since $\hat{F}_n$ converges uniformly to $F$, the result implies that the superpopulation residual distribution is asymptotically valid: it reproduces the correct form of randomness around $F$. This property guarantees that the superpopulation bootstrap accurately captures the variability in regression coefficients—yielding reliable standard errors and reducing the risk of underestimation in finite samples.

Taken together, Theorems 3–5 confirm that the proposed superpopulation-based bootstrap is theoretically well-calibrated. It reproduces the correct asymptotic behavior of both residuals and regression coefficients, ensuring internally coherent inference within the Bayesian latent-variable framework.

# 4 Results

To verify Theorem 1, we estimate $F_B$ and $\hat{F}_B$ using our randomly generated samples of size 3000 and overlay the two empirical distribution functions in the plot. It compares the empirical distribution functions of $\beta_1^\star$ and their target distributions from $\beta_1$ based on 200 bootstrap samples with 1000 and 2000 resampled datasets, as well as 400 bootstrap samples with 2000 and 4000 resampled datasets.

Note that bias is subtracted from $\hat{\beta}_1^\star$ to correct for bias before generating the plot.

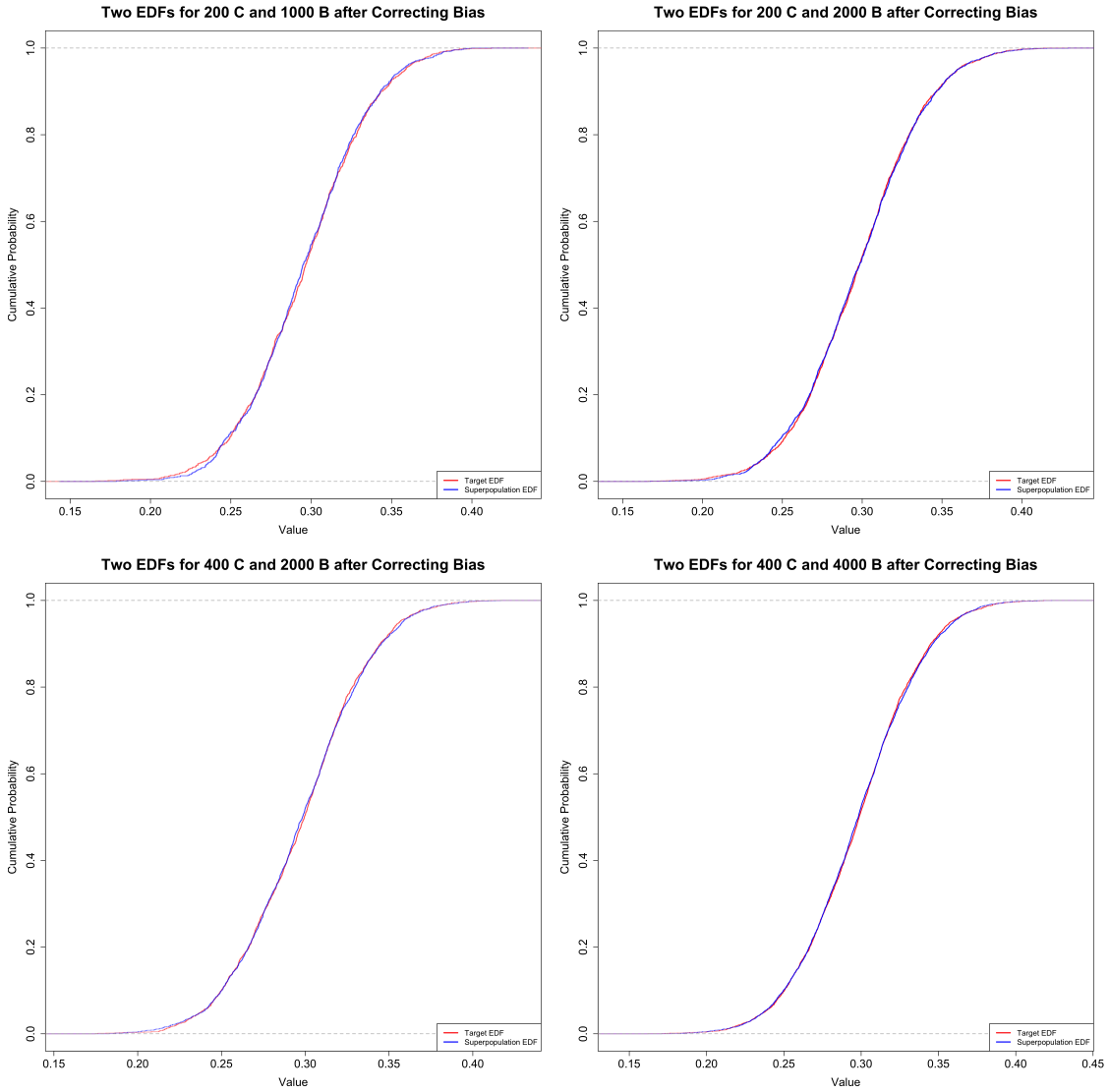


Figure 1: The empirical distribution functions of $\beta_1^\star$ and their target distributions from $\beta_1$.

To verify Theorem 2, we present our estimated results in Table 1, which demonstrates that the ratio $\sqrt{\frac{\text{Var}(\beta_1^\star|F_B)}{\text{Var}(\hat{\beta}_1^\star|\hat{F}_B)}} = \frac{0.0366}{0.0371} \approx 1$, even when $\hat{\beta}_1^\star$ is biased. Additionally, Table 2 presents the results for different sample sizes under the scenario where population prevalence information of all relevant variables in the true data-generating process of the latent variable $Z$ is available.

Table 1: Estimated results of $\beta_1$ for 200 B and 2000 resamples of n = 3000.

| Method | Effect of $Z$ on $Y$ |
|---|---|
| True Population Value | 0.30 |
| A (misclassified) | 0.23 (0.041) |
| True $Z$ | 0.30 (0.036) |
| Binary $\hat{Z}$ | 0.17 (0.037) |
| Continuous $\hat{Z}_P$ | 0.28 (0.049) |

Table 2: Estimated results of $\beta_1$ using all relevant population priors for $Z$.

| Method | $n = 1000$ | $n = 3000$ | $n = 5000$ | $n = 10000$ |
|---|---|---|---|---|
| Population Value | 0.30 | 0.30 | 0.30 | 0.30 |
| True $Z$ | 0.32 (0.0624) | 0.31 (0.0370) | 0.30 (0.0279) | 0.30 (0.0202) |
| Binary $\hat{Z}$ | 0.19 (0.0632) | 0.19 (0.0377) | 0.18 (0.0284) | 0.18 (0.0200) |
| Continuous $\hat{Z}_P$ | 0.28 (0.078) | 0.32 (0.048) | 0.31 (0.036) | 0.30 (0.025) |

To assess the finite-sample performance of the bootstrap variance estimator $\hat{\sigma}_n^\star$ and to illustrate Theorem 4 and the process-level result in Theorem 5, we generated 2000 resampled datasets for each sample size and computed $\hat{\sigma}_n^\star$ in every replication. Table 3 reports the Monte Carlo mean and standard deviation of $\hat{\sigma}_n^\star$ for $n = 3000$ and $n = 10000$, together with the true variance $\sigma^2 = 1.00$. The mean of $\hat{\sigma}_n^\star$ is very close to the population value (1.008 for $n = 3000$ and 0.998 for $n = 10000$), while the standard deviation decreases from 0.013 to 0.006 as the sample size increases. This shrinking dispersion is consistent with Theorem 4, which states that the conditional distribution of $\hat{\sigma}_n^\star$ converges to a point mass at $\sigma^2$, so that $\hat{\sigma}_n^\star$ becomes increasingly concentrated around the true variance in larger samples.

Table 3: Estimated results of $\hat{\sigma}_n^\star$.

| | Mean of $\hat{\sigma}_n^\star$ | SD of $\hat{\sigma}_n^\star$ | $\sigma^2$ of population |
|---|---|---|---|
| n = 3000 | 1.008 | 0.013 | 1.00 |
| n= 10000 | 0.998 | 0.006 | 1.00 |

Figure 2 presents four independent realizations of the process $\sqrt{n}\,(\hat{F}_n^\star - \hat{F}_n)$ for $n = 3000$. Each path exhibits the qualitative features of a Brownian bridge: the process fluctuates randomly around zero in the interior of the support, with excursions that shrink toward zero in the tails where the

empirical distributions align more closely.

This visual behavior is consistent with Theorem 5, which states that $\sqrt{n}\,(\hat{F}_n^\star - \hat{F}_n)$ converges weakly to a Brownian bridge $B(F)$. The oscillatory structure in all four sample paths mirrors the mean-zero, tied-down nature of a Brownian bridge: while the paths display random variation, they return toward zero near the boundary regions. Although finite-sample realizations remain noisy, the overall shape provides empirical support for the theoretical convergence claimed in Theorem 5.

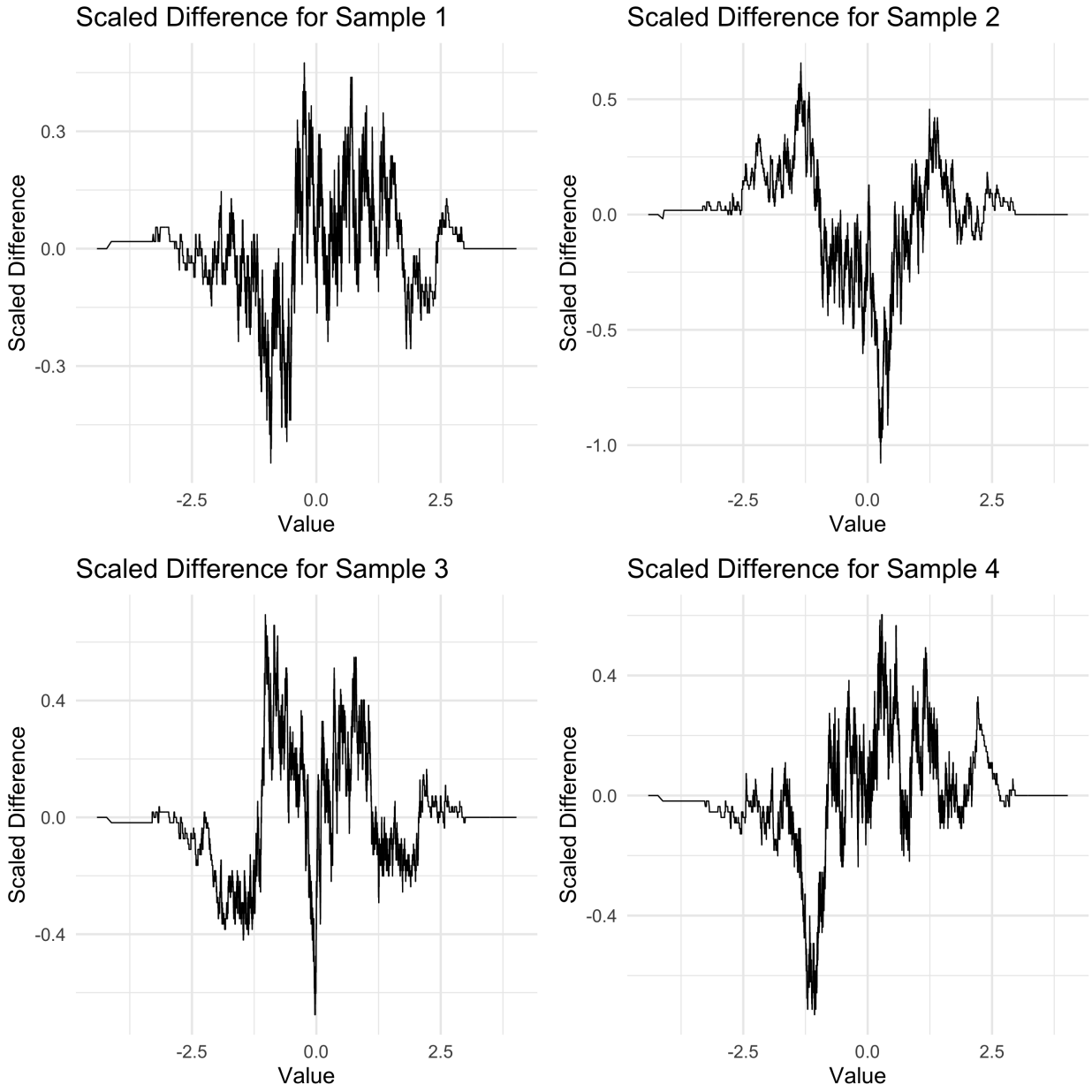


Figure 2: Checking that $\varphi_{n_2}(\hat{F}_n)$ converges weakly to the law of a Brownian bridge $B(F)$.

# 5 Appendix

**Proof of Theorem 1**

By the strong law of large numbers, $\hat{F}_m(t)$ and $F_m(t)$ converge almost surely to $F(t)$, and $\hat{F}_m(t^-)$ and $F_m(t^-)$ converge almost surely to $F(t^-)$. Now, we set points at which $F$ jumps more than $\epsilon/2$ as points of the partition. Given a fixed $\epsilon > 0$, there exists a partition $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ such that $F(t_i^-) - F(t_{i-1}) < \epsilon/2$ for every $i$. Now, for $t_{i-1} \leq t < t_i$, we have:

$$\begin{aligned}
\hat{F}_m(t) - F_m(t) &= (\hat{F}_m(t) - F(t)) - (F_m(t) - F(t)) \\
&\leq (\hat{F}_m(t_i^-) - F(t_i^-) + F(t_i^-) - F(t)) - (F_m(t_i^-) - F(t_i^-) + F(t_i^-) - F(t)) \\
&\leq (\hat{F}_m(t_i^-) - F(t_i^-) + \epsilon/2) - (F_m(t_i^-) - F(t_i^-) - \epsilon/2),
\end{aligned}$$

and similarly,

$$\begin{aligned}
\hat{F}_m(t) - F_m(t) &\geq (\hat{F}_m(t_{i-1}) - F(t_{i-1}) + F(t_{i-1}) - F(t)) - (F_m(t_{i-1}) - F(t_{i-1}) + F(t_{i-1}) - F(t)) \\
&\geq (\hat{F}_m(t_{i-1}) - F(t_{i-1}) - \epsilon/2) - (F_m(t_{i-1}) - F(t_{i-1}) + \epsilon/2).
\end{aligned}$$

Since

$$\max_{i \in \{1,\ldots,k\}} \left\{ |\hat{F}_m(t_i^-) - F(t_i^-)|, |\hat{F}_m(t_{i-1}) - F(t_{i-1})| \right\} \to 0,$$

$$\max_{i \in \{1,\ldots,k\}} \left\{ |F_m(t_i^-) - F(t_i^-)|, |F_m(t_{i-1}) - F(t_{i-1})| \right\} \to 0,$$

we conclude that

$$\limsup_t |\hat{F}_m(t) - F_m(t)| \leq \epsilon.$$

It holds for any $\epsilon > 0$ and hence $\|\hat{F}_m - F_m\|_\infty \to 0$.

**Proof of Theorem 2**

The standard error of a regression coefficient in a linear regression model is

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}},$$

where $\sigma^2$ is the residual variance of the outcome conditional on the predictors. This standard error approximates the standard deviation of the coefficient estimates obtained through bootstrapping. Therefore,

$$\sqrt{\mathrm{Var}(\beta_1^\star | F_m)} \approx \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}} \quad \text{and} \quad \sqrt{\mathrm{Var}(\hat{\beta}_1^\star | \hat{F}_m)} \approx \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(\hat{Z}_i - \bar{\hat{Z}})^2}}.$$

If $\bar{\hat{Z}} \approx \bar{Z}$, then the proportion of 0's and 1's in the binary variable $\hat{Z}$ will be approximately the same as in $Z$ and hence

$$\sum_{i=1}^{n}(\hat{Z}_i - \bar{\hat{Z}})^2 \approx \sum_{i=1}^{n}(Z_i - \bar{Z})^2,$$

and we assume $\sigma^2 \approx \hat{\sigma}^2$ given that $\hat{Z}$ still captures most of the information in $Z$. Therefore,

$$\sqrt{\frac{\mathrm{Var}(\beta_1^\star | F_m)}{\mathrm{Var}(\hat{\beta}_1^\star | \hat{F}_m)}} \approx \sqrt{\frac{\sum_{i=1}^{n}(\hat{Z}_i - \bar{\hat{Z}})^2}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}} \approx 1.$$

**Derivation of** $\Pr(Z = 1 \mid A, X, \psi)$

Using the assumption that $A$ depends on $Z$ and $X$, $Z \perp X \mid \psi$, and $\psi$ depends on $X$, we have:

$$\begin{aligned}
\Pr(A \mid X, \psi) &= \sum_{Z=0}^{1} \Pr(A, Z \mid X, \psi) \\
&= \Pr(A \mid Z = 1, X) \Pr(Z = 1 \mid \psi) + \Pr(A \mid Z = 0, X) \Pr(Z = 0 \mid \psi).
\end{aligned}$$

Using the above result and the perfect specificity assumption $\Pr(A = 1 \mid Z = 0) = 0$ (i.e., no false positives), we get:

$$\Pr(A \mid X, \psi) = \begin{cases} \Pr(A = 1 \mid Z = 1, X) \Pr(Z = 1 \mid \psi), & \text{if } A = 1, \\ \Pr(A = 0 \mid Z = 1, X) \Pr(Z = 1 \mid \psi) + \Pr(Z = 0 \mid \psi), & \text{if } A = 0. \end{cases}$$

Using Bayes' rule,

$$\begin{aligned}\Pr(Z = 1 \mid A, X, \psi) &= \frac{\Pr(A \mid Z = 1, X, \psi)\Pr(Z = 1, X, \psi)}{\Pr(A, X, \psi)} \\ &= \frac{\Pr(A \mid Z = 1, X)\Pr(Z = 1 \mid \psi)}{\Pr(A \mid X, \psi)}.\end{aligned}$$

Let $\rho = \Pr(A = 1 \mid Z = 1, X)$ and $\psi = \Pr(Z = 1 \mid \psi)$. Then:

$$\Pr(Z = 1 \mid A, X, \psi) = \begin{cases} \frac{\rho\psi}{\rho\psi}, & \text{if } A = 1, \\ \frac{(1-\rho)\psi}{(1-\rho)\psi+(1-\psi)}, & \text{if } A = 0. \end{cases}$$

**Data Simulation**

Motivated by documented underreporting of CPS involvement in the Future of Families and Child Wellbeing Study (FFCWS), our simulation design follows the framework introduced in Berger et al. (2025), recently accepted in *Social Science Review*. That paper compares FFCWS self-reported CPS involvement to national benchmarks from NCANDS and provides a detailed description of the data-generating process used to replicate realistic rates of underreporting.

Briefly, the simulated dataset consists of: (i) a latent true CPS involvement indicator $Z$, (ii) a misclassified self-reported CPS indicator $A$, and (iii) observed covariates $X$ capturing age, race/ethnicity, region, and continuous socioeconomic characteristics.

To mirror real-world prevalence patterns, we first generate a population of 100,000 observations and then draw a representative subsample of 3,000 units for analysis. The full specification of the simulation model—including the covariate distributions, misclassification mechanism, and outcome-generating process—is provided in Berger et al. (2025). We refer readers to that paper for complete methodological details.

## References


Berger, L. M., Dickerson, T., Gelman, A., Jung, H.-M., Lee, S., Thomas, M., and Waldfogel, J. (2025). Adjusting for underreporting of Child Protective Services involvement in the Future of Families and Child Wellbeing Study and assessing its empirical implications through illustrative analyses of young adult disconnection. *Social Science Review*, forthcoming.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics* **9**(6), 1196–1217.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, **9**(6), 1218–1228.

Mallows, C. L. (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics* **43**(2), 508–515.

Sarndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Valliant, R., Dever, J. A. and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Chapman & Hall/CRC, Boca Raton.