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

Seonghun Lee

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 ψ 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 ψ, 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 ε 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 β_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 β_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 β_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 $\hat{Z}_{\text{P}}$. 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 β_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. Next, for each sampled observation, we randomly generate new binary values $\hat{Z}$ from a Bernoulli distribution with parameter $\hat{Z}_{\text{P}}$, 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 $\hat{\beta}_1^*$:

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

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

A natural question is why we generate a superpopulation and then resample datasets, rather than directly using bootstrapped samples with a single realization of binary assignments. The key reason is to account for the uncertainty in converting continuous posterior probabilities into binary values, which impacts the true variability of the regression coefficient β_1. If we were to estimate β_1 directly from bootstrapped samples without resampling from a superpopulation, the standard error of those estimates would be artificially small, failing to capture the variability introduced by the uncertainty inherent in the binary classification process.

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. 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:

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

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 $\hat{\beta}^*$:

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

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 σ, where $\hat{\sigma}^*_n$ is the variance of $\hat{\varepsilon}^*_{z_p}$ and σ is the variance of ε. 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 β^*_1 and $\hat{\beta}^*_1$ − "Bias" are approximately normally distributed around the population value of β_1, there exists $\sigma^2_m \approx \hat{\sigma}^2_m$ such that

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

where $\text{Var}(\beta^*_1) = \sigma^2_m$ and $\text{Var}(\hat{\beta}^*_1) = \hat{\sigma}^2_m$. The rationale for assuming $\sigma^2_m \approx \hat{\sigma}^2_m$ 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 β_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 β^*_1 and $\hat{\beta}^*_1$ denote estimators obtained from random resamples drawn from a superpopulation, and let F denote the distribution of β_1. If β^*_1 and $\hat{\beta}^*_1$ are unbiased estimators of β_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 β_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 R^p, and we define the Mallows distance as the infimum of $(E\|U - V\|^r)^{1/r}$ over all pairs of random vectors U and V with laws u and v, respectively.

Lemma 1. Assuming $X^T X = O_p(n)$,

$$\frac{1}{n}X^T\epsilon \xrightarrow{a.s.} 0.$$

Proof. Consider the subsequence of powers of 2, $n_m = 2^m$. For any $\lambda > 0$, using Kolmogorov's inequality and $X^T X = O_p(n)$ gives

$$P\left(\sup_{n_m \geq 1}\left|\frac{1}{n_m}\sum_{i=1}^{n_m} X_i^T\epsilon_i\right| \geq \lambda\right) \leq \frac{\sigma^2}{\lambda^2}\frac{\sum_{i=1}^{n_m} X_i^T X_i}{n_m^2} = O_p(2^{-m}).$$

Since $\sum_{m=1}^{\infty} O_p(2^{-m}) < \infty$, by the Borel-Cantelli lemma,

$$\frac{1}{n_m} \sum_{i=1}^{n_m} X_i^T \epsilon_i \xrightarrow{a.s.} 0.$$

For a general n, we can find m such that $n_m = 2^m \leq n < 2^{m+1} = n_{m+1}$ and $\frac{n_m}{n} \geq \frac{1}{2}$. Let $S_n = \sum_{i=1}^{n} X_i^T \epsilon_i$. Then,

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

Since $\frac{S_{n_m}}{n} \leq 2\frac{S_{n_m}}{n_m} \xrightarrow{a.s.} 0$ and the second term behaves similarly, we have $\frac{S_n}{n} \xrightarrow{a.s.} 0$.

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. Assuming $\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_{z_{p_i}} = 0$, then $d_2(\hat{F}_n, F_n) \xrightarrow{\text{a.s.}} 0$, where $\hat{F}_n$ is the empirical distribution of $\hat{\varepsilon}_{z_{p_1}}, \ldots, \hat{\varepsilon}_{z_{p_n}}$, and F_n is the empirical distribution of $\varepsilon_1, \ldots, \varepsilon_n$.

Proof. By Lemma 2,

$$d_2(\hat{F}_n, F_n)^2 \leq \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\varepsilon}_{z_{p_i}} - \varepsilon_i\right)^2 = \frac{1}{n} \left\|\hat{\varepsilon}_{z_p} - \varepsilon\right\|^2 \xrightarrow{\text{a.s.}} 0.$$

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 μ_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:

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

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 $\psi_n(\hat{F}_n)$ be the law of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$, where F is the common distribution of $\epsilon_1, \ldots, \epsilon_n$ and $\hat{F}_n$ is the common distribution of $\epsilon^*_{z_{p1}}, \ldots, \epsilon^*_{z_{pn}}$.

Use Theorem 2.1 of Freedman (1981) and the fact that $\sqrt{n}(\hat{\beta}^* - \hat{\beta}) = \sqrt{n}\left((X^*)^\top X^*\right)^{-1}(X^*)^\top \epsilon^*_{z_p}$,

$$\begin{aligned}d_2^p\left(\psi_n(\hat{F}_n), \psi_n(F)\right)^2 &\leq n\,\mathrm{trace}\left(\left((X^*)^\top X^*\right)^{-1}(X^*)^\top X(X^\top X)^{-1}\right)^{-1} d_2(F_n, F)^2 \\ &= \mathrm{trace}\left(\frac{1}{n}\left((X^*)^\top X^*\right)^{-1}(X^*)^\top X(X^\top X)^{-1}\right)^{-1} d_2(F_n, F)^2 \xrightarrow{a.s.} 0.\end{aligned}$$

by assumption and Lemma 4. Specifically, $\frac{1}{n}(X^*)^\top X^*$ is also positive definite, and $(X^*)^\top X$ is well-conditioned and non-singular. Thus, the trace is bounded away from zero, ensuring that its inverse remains finite.

The variance estimates of the estimated residuals and starred residuals are:

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

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

Proof. Using Jensen's inequality and Lemma 5,

$$\begin{aligned}
E\left[|\hat{\sigma}_n^\star - \sigma_n^\star| \,|\, Y_1, \ldots, Y_n\right]^2 &\leq E\left[(\hat{\sigma}_n^\star - \sigma_n^\star)^2 \,\middle|\, Y_1, \ldots, Y_n\right] \\
&\leq E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}^\star_{z_{p_i}} - \varepsilon^\star_{z_{p_i}}\right)^2 \,\middle|\, Y_1, \ldots, Y_n\right] \\
&= \frac{\hat{\sigma}_n}{n}\operatorname{trace}\left(X^\star(X^{\star\top}X^\star)^{-1}X^{\star\top}\right) \\
&= \frac{\hat{\sigma}_n p}{n} \xrightarrow{\text{a.s.}} 0,
\end{aligned}$$

because $\hat{\sigma}_n \xrightarrow{\text{a.s.}} \sigma$ by Lemmas 2 and 5. Specifically, since $\sigma_n \xrightarrow{\text{a.s.}} \sigma$, the Lemmas yield

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

Next, we show that the conditional distribution of $\sigma_n^\star$ is nearly a point mass at σ. By Lemma 4, $d_2(\hat{F}_n - F) \xrightarrow{\text{a.s.}} 0$, where $\varepsilon^\star_{z_{p_1}} \sim \hat{F}_n$ and $\varepsilon_i \sim F$. Applying this fact along with Lemmas 8.5 and 8.6 of Bickel and Freedman (1981), and using $\varphi(\varepsilon) = \varepsilon^2$ in Lemma 8.5, we obtain:

$$d_1\left(\frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon^\star_{z_{p_i}}\right)^2, \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right) \leq d_1\left(\left(\varepsilon^\star_{z_{p_i}}\right)^2, \varepsilon_i^2\right) \xrightarrow{\text{a.s.}} 0.$$

This result implies that the conditional law of $\frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon^{\star}_{z_{p_i}}\right)^2$ is nearly identical to the unconditional law of $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$. Consequently, the conditional law must concentrate around σ^2. Similarly, the conditional law of $\frac{1}{n}\sum_{i=1}^{n}\varepsilon^{\star}_{z_{p_i}}$ concentrates near 0.

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

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

Proof. Theorem 4.1 of Bickel and Freedman (1981) states that $\varphi_{n_1}(\hat{F}_n)$ converges weakly to the law of $B(F)$ using the facts that $\|\hat{F}_n - F\|_\infty \to 0$ almost surely by the Glivenko–Cantelli theorem, and that $\varphi_n(F)$ converges weakly to the law of $B(F)$ by Donsker's theorem, which is an application of the ordinary invariance principle.

Thus, we aim to show that

$$d_2(\hat{F}_n^{\star}, F_n^{\star}) \to 0 \quad \text{(a.s.)},$$

as this implies $\hat{F}_n^{\star}$ weakly converges to $F_n^{\star}$. We then apply the continuous mapping theorem to conclude that $\varphi_{n_2}(\hat{F}_n)$ also converges weakly to the law of $B(F)$.

To this end, we apply the result from the proof of Theorem 4 and use the same approach as in the proofs of Lemma 1, Lemma 2, and Lemma 3. Specifically, we utilize the fact that the conditional distribution of $\sigma_n^{\star}$ is nearly point mass at σ, which implies that

$$\frac{1}{n}X^{\star\top}\varepsilon^{\star}_{z_p} \to 0 \quad \text{(a.s.)}.$$

This leads to the result

$$d_2(\hat{F}_n^{\star}, F_n^{\star}) \le \frac{1}{n}\|\hat{\varepsilon}_{z_p} - \varepsilon^{\star}_{z_p}\|^2 \to 0 \quad \text{(a.s.)}.$$

Since $\hat{F}_n^{\star}$ weakly converges to $F_n^{\star}$ and $\varphi_{n_1}(\hat{F}_n)$ converges weakly to the law of $B(F)$, by the continuous mapping theorem, $\varphi_{n_2}(\hat{F}_n)$ also converges weakly to the law of $B(F)$.

Theorems 3 and 4 provide complementary insights that together offer a meaningful interpretation of our Bayesian latent modeling and inference using a superpopulation. Specifically, as the conditional distribution of $\hat{\sigma}_n^\star$ converges to a point mass at σ, the variability in the resampled residuals becomes more concentrated around the true residual variance. The convergence of residual variance and its relationship with $\hat{\beta}^\star$ indicates that the dispersion of the resampled regression coefficients $\hat{\beta}^\star$ accurately captures the true sampling variability, consistent with the asymptotic behavior $\hat{\beta}^\star \xrightarrow{d} \mathcal{N}(\hat{\beta}, \sigma^2(X^TX)^{-1})$. This gives us confidence in the precision of our estimator $\hat{\beta}^\star$. Taken together, these results reinforce our confidence in the overall modeling approach. The convergence of residual variance suggests that constructing a superpopulation using bootstrapped samples—and resampling from it— properly calibrates the variability of the regression coefficients. Furthermore, the results indicate that the Bayesian latent model effectively captures the latent variable Z, as the uncertainty in the regression residuals using $\hat{Z}_P^\star$ closely mirrors the uncertainty that would arise from the true latent variable Z.

The result of Theorem 5 is also meaningful because the weak convergence of $\varphi_{n_2}(\hat{F}_n)$ to the law of a Brownian bridge $B(F)$ establishes the asymptotic behavior of the empirical distribution of residuals. As the sample size increases, the difference between the empirical and true distribution becomes structured and predictable, following the pattern of a Brownian bridge. Specifically, since $\hat{F}_n$ converges uniformly to F, and the conditional distribution of $\sqrt{n}(\hat{F}_n^\star - \hat{F}_n)$ converges weakly to a Brownian bridge $B(F)$, the fluctuations of $\hat{F}_n^\star$ around $\hat{F}_n$ reflect the sampling variability around the true residual distribution. This asymptotic behavior, characterized by $B(F)$, supports the asymptotic validity of the superpopulation-based residual distribution. This validity provides a foundation for more accurate variability in the resampled regression coefficients $\hat{\beta}^*$, yielding improved standard error estimates and mitigating the risk of underestimation in finite samples. Thus, this result also reinforces the validity of our method.

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 β_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.

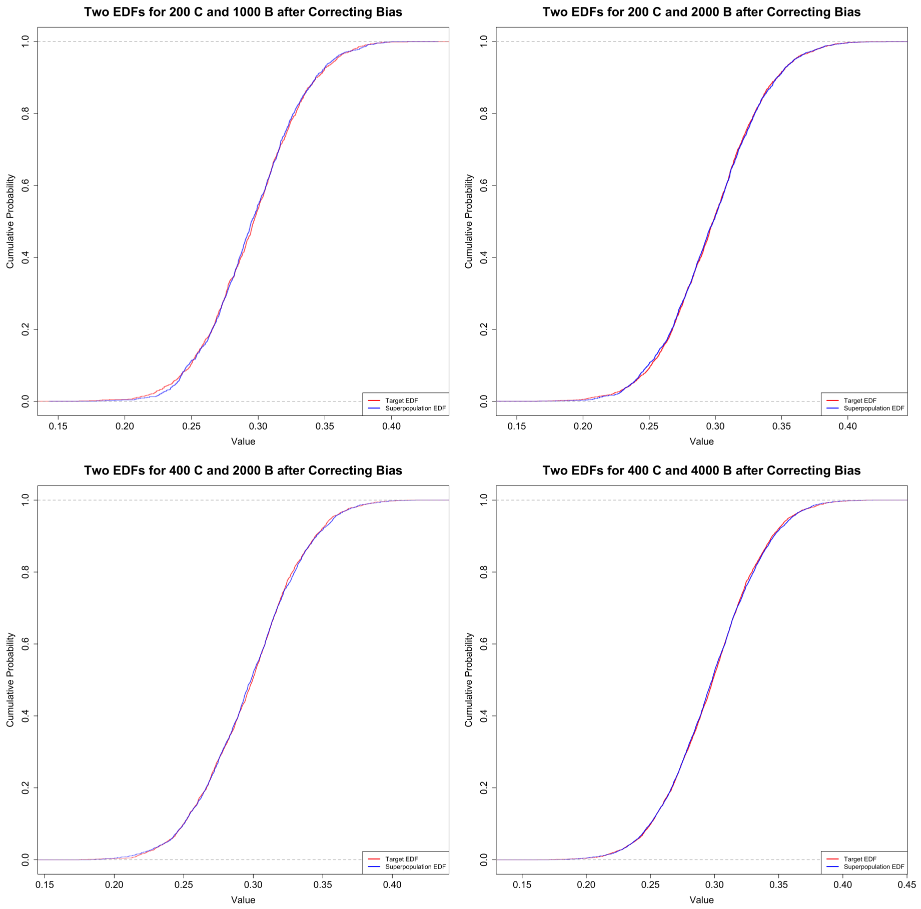

Figure 1: The empirical distribution functions of $\beta_1^\star$ and their target distributions from β_1.

To verify Theorem 2, we present our estimated results in Table 1, which demonstrates that the ratio $\sqrt{\frac{\mathrm{Var}(\beta_1^\star|F_B)}{\mathrm{Var}(\hat{\beta}_1^\star|\hat{F}_B)}} = \frac{0.0366}{0.0371} \approx 1$, 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 $\beta_1^\star$ based on 200 bootstrap and 2000 resamples.

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 $\beta_1^\star$ 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)

We now present the pooled results from 2000 resampled datasets to verify Theorem 4 and Theorem 5 using our randomly generated samples of size 3000. In Table 3, we expect the variance of $\hat{\sigma}_n^\star$ to approach zero as the sample size n grows, rather than remaining at 0.0006.

Table 3: Estimated results of $\hat{\sigma}_n^\star$ for $n = 3000$.

	Mean of $\hat{\sigma}_n^\star$	σ^2 of samples	σ^2 of population
Value	1.02 (0.0006)	1.01	1.00

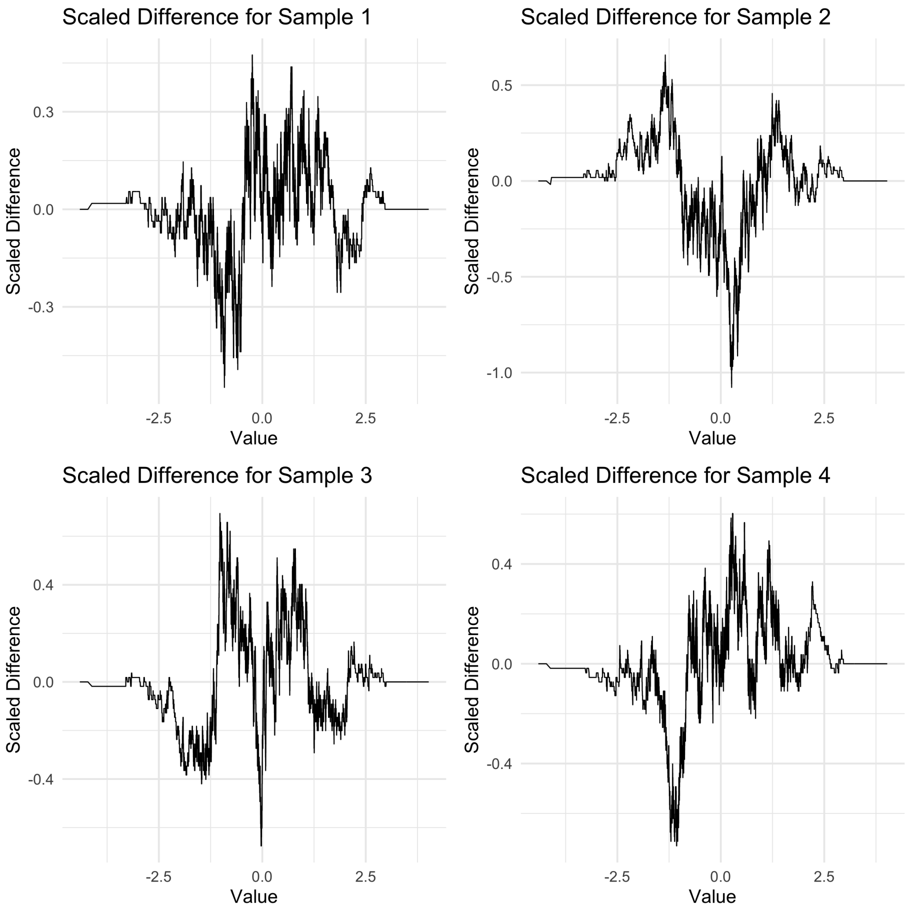

Figure 2: Checking that $\varphi_{n_2}(\hat{F}_n)$ converges weakly to the law of a Brownian bridge $B(F)$.

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 $t_{i-1} \le t < t_i$, we have:

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

and similarly,

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

Since

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

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

we conclude that

$$\lim \sup_t |\hat{F}_m(t) - F_m(t)| \le \epsilon.$$

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 σ^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 ψ 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 underreported CPS prevalence rates in FFCWS relative to national prevalence in NCANDS, the simulated data consists of true values of CPS involvement Z, a binary latent variable; misclassified self-reported CPS involvement A; and observed covariates X. We denote ψ as the probability of $Z = 1$ given our prior expectations of the prevalence of Z for each age-, race-/ethnicity-, and region-specific group. To reflect the features of real-world data, we generate a dataset of 100,000 observations to represent our population and randomly draw 3,000 observations to ensure the subpopulation is representative. The simulated data is generated in four steps:

1. Define the coefficients:

$$\begin{aligned}\alpha &= [-0.1,\ 1,\ 0.2,\ -0.2,\ -0.1,\ -1,\ -0.8,\ -0.9,\ -0.3] \\ \beta &= [1.1,\ 0.25,\ 0.1,\ -0.1,\ -0.15,\ -1,\ -0.7,\ -0.8,\ -0.2,\ 0.2]\end{aligned}$$

2. Generate covariates X:
   - Race/Ethnicity:
   $$x_{\text{race/ethnicity}} \sim \text{Multinomial}(0.15,\ 0.5,\ 0.25,\ 0.1)$$
   - Region:
   $$x_{\text{region}} \sim \text{Multinomial}(0.12,\ 0.18,\ 0.2,\ 0.27,\ 0.23)$$
   - Continuous:
   $$x_{\text{continuous}} \sim \text{Normal}(1,\ 2)$$
3. Generate true exposure and misclassified exposure:
   $$\begin{aligned} Z \mid X &\sim \text{Bernoulli}(\text{expit}(X^\top \alpha)), \\ A \mid X &\sim \text{Bernoulli}(\text{expit}(X^\top \beta + \beta_9 Z)). \end{aligned}$$
   Set $A = 0$ whenever $Z = 0$ to simulate only false negatives. This results in 382 misclassified cases out of 3,000, with no false positives.
4. Generate the outcome:
   $$Y \sim \text{Normal}(0.1 + 0.3Z + x_{\text{continuous}},\ 1).$$