Logistic Regression for Massive Data with Rare Events

Haiying Wang

Lifeng Liu

USTC

March 8th, 2022

Content

Introduction

- Big data with rare events in binary responses, also called imbalanced data, are data in which the number of events is much smaller than the number of non-events.
- Cases: Events; Controls: Nonevents.
- A commonly approach: under-sampling and/or over-sampling.
- Theoretical analyses of the effects of under-sampling and over-sampling in terms of parameter estimation are still rare.

Introduction

- Many articles obtained theoretical results based on the regular assumption that the probability of event occurring is fixed and does not go to zero.
- In this paper, we obtain convergence rates and asymptotic distributions of parameter estimators under the assumption that both the number of cases and the number of controls are random.
- This is the first study that provides distributional results for rare events data with a decaying event rate.

Introduction

Main contributions:

- Derive the asymptotic distribution of the maximum likelihood estimator (MLE) of the unknown parameter, which shows that the asymptotic variance convergences to zero in a rate of the inverse of the number of the events instead of the inverse of the full data sample size.
- Prove that under-sampling a small proportion of the nonevents, the resulting under-sampled estimator may have identical asymptotic distribution to the full data MLE.
- Show that over-sampling(replicate) approach may even result in efficiency loss in terms of parameter estimation.

Model

Let $\mathcal{D}_n = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ be independent data of size n from a logistic regression model,

$$\mathbb{P}(y = 1 \mid \mathbf{x}) = p(\alpha, \boldsymbol{\beta}) = \frac{e^{\alpha + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}}{1 + e^{\alpha + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}}. \tag{1}$$

- $\mathbf{x} \in \mathbb{R}^d$ is the covariate, $\mathbf{z} = (1, \mathbf{x}^T)^T$.
- y $\in \{0, 1\}$ is the binary class label, 1 for cases and 0 for controls.
- $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^T)^T$.

This paper focuses on estimating the unknown $\boldsymbol{\theta}$.

Model

- Let n_1 and n_0 be the numbers cases (observations with $y_i = 1$) and controls (observations with $y_i = 0$). And here, n_1 and n_0 are random, because they are summary statistics about the observed data, i.e., $n_1 = \sum_{i=1}^{n} y_i$ and $n_0 = n - n_1$.
- For rare events data n_1 is much smaller than n_0. Thus, for asymptotic investigations, it is reasonable to assume that $n_1/n_0 \to 0$, or equivalently $n_1/n \to 0$ in probability, as $n \to \infty$.
- For big data with rare events, there should be a fair amount of cases observed, so it is appropriate to assume that $n_1 \to \infty$ in probability.

Model

To model this scenario, we assume that the marginal event probability $\mathbb{P}(y=1)$ satisfies that as $n \to \infty$,

$$\mathbb{P}(y=1) \to 0 \quad \text{and} \quad n\mathbb{P}(y=1) \to \infty. \tag{2}$$

We accommodate this condition by assuming that the true value of $\boldsymbol{\beta}$, denoted as $\boldsymbol{\beta}_t$, is fixed while the true value of α, denoted as α_{nt}. Specifically, we assume $\alpha_{nt} \to -\infty$ as $n \to \infty$ in a rate such that

$$\begin{aligned} \frac{n_1}{n} &= \mathbb{P}(y=1)\left\{1 + o_P(1)\right\} \\ &= \mathbb{E}\left(\frac{e^{\alpha_{nt}+\beta_t^{\mathrm{T}}\mathbf{x}}}{1+e^{\alpha_{nt}+\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}}\right)\left\{1 + o_P(1)\right\}. \end{aligned} \tag{3}$$

Model

The MLE based on the full data $\mathcal{D}_n$, say $\hat{\boldsymbol{\beta}}$, is the maximizer of

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{\pi} \left\{ y_i \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta} - \log\left(1 + e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta}}\right) \right\}, \tag{4}$$

which is also the solution to the following equation,

$$\dot{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ y_i - p_i(\alpha, \boldsymbol{\beta}) \right\} \mathbf{z}_i = 0, \tag{5}$$

where $\dot{\ell}(\boldsymbol{\theta})$ is the gradient of the log-likelihood $\ell(\boldsymbol{\theta})$.

Model

The following Theorem gives the asymptotic normality of the MLE $\hat{\beta}$ for rare events data.

Theorem 1

If $\mathbb{E}\left(e^{t\|\mathbf{x}\|}\right) < \infty$ for any $t > 0$ and $\mathbb{E}\left(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\right)$ is a positive-definite matrix, then under the conditions in (2) and (3), as $n \to \infty$,

$$\sqrt{n_1}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{nt}\right) \longrightarrow \mathbb{N}\left(\mathbf{0}, \mathbf{V}_f\right), \tag{6}$$

in distribution, where

$$\begin{aligned} \mathbf{V}_f &= \mathbb{E}\left(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\right)\mathbf{M}_f^{-1}, \quad \text{and} \\ \mathbf{M}_f &= \mathbb{E}\left(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\right) = \mathbb{E}\left\{e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\begin{pmatrix} 1 & \mathbf{x}^{\mathrm{T}} \\ \mathbf{x} & \mathbf{x}\mathbf{x}^{\mathrm{T}} \end{pmatrix}\right\}. \end{aligned} \tag{7}$$

Under-sampled Estimator

Questions:

- Convergence rate?
- Estimation efficiency loss (an enlarged asymptotic variance)?

Under-sampled Estimator

From the full data set $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, we want to use all the cases while only select a subset for the controls. Let π_0 be the probability that each data points with $y_i = 0$ is selected in the subset, and $\delta_i \in \{0, 1\}$ be the binary indicator variable that signifies if the i-th observation is included in the subset.Here, we define the sampling plan by assigning

$$\delta_i = y_i + (1 - y_i) I (u_i \leq \pi_0), \quad i = 1, \ldots, n, \tag{8}$$

where $u_i \sim \mathbb{U}(0, 1), i = 1, \ldots, n$.

Under-sampled Weighted Estimator

The sampling inclusion probability given the full data $\mathcal{D}_n$ for the i-th data point is

$$\pi_i = \mathbb{E}\left(\delta_i \mid \mathcal{D}_n\right) = y_i + \left(1 - y_i\right)\pi_0 = \pi_0 + \left(1 - \pi_0\right) y_i.$$

the under-sampled weighted estimator, $\hat{\boldsymbol{\theta}}^{w}_{under}$, is the maximizer of

$$\ell^{\mathrm{w}}_{\text{under}}\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{n} \frac{\delta_i}{\pi_i}\left\{y_i \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta} - \log\left(1 + e^{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta}}\right)\right\}. \tag{9}$$

Under-sampled Weighted Estimator

We present the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{under}^{w}$ in the following theorem

Theorem 2

If $\mathbb{E}(e^{t\|\mathbf{x}\|}) < \infty$ for any $t > 0, \mathbb{E}(e^{\boldsymbol{\theta}_{nt}^{\mathrm{T}}\mathbf{x}}\mathbf{z}\mathbf{z}^{\mathrm{T}})$ is a positive-define matrix, and $c_n = e^{\alpha_{nt}}/\pi_0 \to c$ for a constant $c \in [0, \infty)$, then under the conditions in (2) and (3), as $n \to \infty$,

$$\sqrt{n_1}\left(\hat{\boldsymbol{\theta}}_{\text{under}}^{\text{w}} - \boldsymbol{\theta}_{nt}\right) \longrightarrow \mathbb{N}\left(\mathbf{0}, \mathbf{V}_{\text{under}}^{\text{w}}\right), \tag{10}$$

in distribution, where

$$\begin{aligned} \mathbf{V}_{\text{under}}^{\text{w}} &= \mathbb{E}\left(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\right)\mathbf{M}_f^{-1}\mathbf{M}_{\text{under}}^{\text{w}}\mathbf{M}_f^{-1}, \quad \text{and} \\ \mathbf{M}_{\text{under}}^{\text{w}} &= \mathbb{E}\left\{e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\left(1 + ce^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\right)\mathbf{z}\mathbf{z}^{\mathrm{T}}\right\}. \end{aligned} \tag{11}$$

Under-sampled Weighted Estimator

Remark. If $\mathbb{E}(e^{t\|\mathbf{x}\|}) < \infty$ for any $t > 0$, then from (3) and the dominated convergence theorem, we know that $n_1 = ne^{\alpha_{nt}}\mathbb{E}(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}})\{1 + o_P(1)\}$. Thus

$$c_n\mathbb{E}\left(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\right) = \frac{n_1}{n\pi_0}\{1 + o_P(1)\} = \frac{n_1}{n_0\pi_0}\{1 + o_P(1)\}.$$

$c\mathbb{E}(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}})$ can be interpreted as the asymptotic ratio of the number of cases to the number of controls in the under-sampled data. Therefore, since $\mathbb{E}(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}) > 0$ is a fixed constant, the value of c has the following intuitive interpretations.

- $c = 0$: take much more controls than cases;
- $0 < c < \infty$: the number of controls to take is at the same order of the number of cases;
- $c = \infty$: take much fewer control than cases.

Under-sampled Unweighted Estimator

Based on the control under-sampled data, if we obtain an estimator from an unweighted objective function, say

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}^{\mathrm{u}}_{\mathsf{under}} &= \arg\max_{\boldsymbol{\theta}} \ell^{\mathrm{u}}_{\mathrm{under}}(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \delta_i \left[y_i \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta} - \log\left\{ 1 + e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta}} \right\} \right],
\end{aligned}
$$

where $\tilde{\boldsymbol{\theta}}^{\mathrm{u}}_{\mathsf{under}} = (\hat{\alpha}^{u}_{under}, \hat{\boldsymbol{\beta}}^{u}_{under})^{\mathrm{T}}$, $\hat{\alpha}^{u}_{under}$ is the intercept estimator and $\hat{\boldsymbol{\beta}}^{u}_{under}$ is the slope estimator.

Under-sampled Unweighted Estimator

According to Fithian & Hastie, 2014, Wang, 2019, the intercept estimator $\hat{\alpha}^{u}_{under}$is asymptotically biased while the slope estimator $\hat{\boldsymbol{\beta}}^{u}_{under}$is still asymptotically unbiased. We define the under-sampled unweighted estimator with bias correction $\hat{\boldsymbol{\theta}}^{ubc}_{under}$ as

$$\hat{\boldsymbol{\theta}}^{ubc}_{under} = \tilde{\boldsymbol{\theta}}^{\mathrm{u}}_{\textsf{under}} + \mathbf{b}, \tag{12}$$

where

$$\mathbf{b} = \{\log(\pi_0), 0, \ldots, 0\}^{\mathrm{T}}. \tag{13}$$

Under-sampled Unweighted Estimator

The following theorem gives asymptotic distribution of $\hat{\boldsymbol{\theta}}_{under}^{ubc}$

Theorem 3

If $\mathbb{E}(e^{t\|\mathbf{x}\|}) < \infty$ for any $t > 0, \mathbb{E}(e^{\boldsymbol{\theta}_{nt}^{\mathrm{T}}\mathbf{x}}\mathbf{z}\mathbf{z}^{\mathrm{T}})$ is a positive-define matrix, and $e^{\alpha_{nt}}/\pi_0 \to c$ for a constant $c \in [0, \infty)$, then under the conditions in (2) and (3), as $n \to \infty$,

$$\sqrt{n_1}\left(\hat{\boldsymbol{\theta}}_{\text{under}}^{\text{ubc}} - \boldsymbol{\theta}_{nt}\right) \longrightarrow \mathbb{N}\left(\mathbf{0}, \mathbf{V}_{\text{under}}^{\text{ubc}}\right), \tag{14}$$

in distribution, where

$$\mathbf{V}_{\text{under}}^{\text{ubc}} = \mathbb{E}\left(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\right)\left(\mathbf{M}_{\text{under}}^{\text{ubc}}\right)^{-1}, \quad \text{and}$$
$$\mathbf{M}_{\text{under}}^{\text{ubc}} = \mathbb{E}\left(\frac{e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}}{1 + ce^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\mathbf{z}\mathbf{z}^{\mathrm{T}}}\right). \tag{15}$$

Under-sampled Unweighted Estimator

Proposition 1. Let $\mathbf{v}$ be a random vector and h be a positive scalar random variable. Assume that $\mathbb{E}(\mathbf{v}\mathbf{v}^{\mathrm{T}})$, $\mathbb{E}(h\mathbf{v}\mathbf{v}^{\mathrm{T}})$ and $h^{-1}\mathbb{E}(\mathbf{v}\mathbf{v}^{\mathrm{T}})$ are all finite and positive-define matrices. The following inequality holds in the Loewner order.

$$\left\{\mathbb{E}\left(h^{-1}\mathbf{v}\mathbf{v}^{\mathrm{T}}\right)\right\}^{-1} \leq \left\{\mathbb{E}\left(\mathbf{v}\mathbf{v}^{\mathrm{T}}\right)\right\}^{-1} \mathbb{E}\left(h\mathbf{v}\mathbf{v}^{\mathrm{T}}\right) \left\{\mathbb{E}\left(\mathbf{v}\mathbf{v}^{\mathrm{T}}\right)\right\}^{-1}.$$

If we let $\mathbf{v} = e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}/2}\mathbf{z}$ and $h = 1 + ce^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}$, then we can know that $\mathbf{V}_{\text{under}}^{\text{ubc}} \leq \mathbf{V}_{\text{under}}^{\text{w}}$ in the Loewner order.

Over-sampled Estimator

Let τ_i denote the number of times that a data point is used, and define

$$\tau_i = y_i v_i + 1, \quad i = 1, \ldots, n, \tag{16}$$

where $v_i \sim \mathbb{POI}(\lambda_n), i = 1, \ldots, n$, are i.i.d. For this over-sampling plan, a data point with $y_0 = 0$ will be used only one time, while a data point with $y_i = 1$ will be on average used in the over-sampled data for $\mathbb{E}(\tau_i \mid \mathcal{D}_n, y_i = 1) = 1 + \lambda_n$ times. Here λ_n can be interpreted as the average over-sampling rate for cases.

Over-sampled Weighted Estimator

Let $\omega_i = \mathbb{E}(\tau_i \mid \mathcal{D}_n) = 1 + \lambda_n y_i$. The case over-sampled weighted estimator, $\hat{\boldsymbol{\theta}}^w_{over}$, is the maximizer of

$$\ell^{\mathrm{w}}_{\mathrm{over}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\tau_i}{w_i} \left\{ y_i \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta} - \log\left(1 + e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta}}\right) \right\}. \tag{17}$$

The following theorem gives the asymptotic distribution of $\hat{\boldsymbol{\theta}}^w_{over}$.

Over-sampled Weighted Estimator

Theorem 4

If $\mathbb{E}(e^{t\|\mathbf{x}\|}) < \infty$ for any $t > 0, \mathbb{E}(e^{\boldsymbol{\theta}_{nt}^{\mathrm{T}}\mathbf{x}}\mathbf{z}\mathbf{z}^{\mathrm{T}})$ is a positive-define matrix, and $\lambda_n \to \lambda \geq 0$, then under the condition in (2) and (3), as $n \to \infty$,

$$\sqrt{n_1}\left(\hat{\boldsymbol{\theta}}_{\text{over}}^{\mathrm{w}} - \boldsymbol{\theta}_{nt}\right) \longrightarrow \mathbb{N}\left(\mathbf{0}, \mathbf{V}_{\text{over}}^{\mathrm{w}}\right), \tag{18}$$

in distribution, where

$$\mathbf{V}_{\text{over}}^{\mathrm{w}} = \frac{(1+\lambda)^2 + \lambda}{(1+\lambda)^2}\mathbb{E}\left(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\right)\mathbf{M}_f^{-1}. \tag{19}$$

Over-sampled Unweighted Estimator

Define $\hat{\boldsymbol{\theta}}^{ubc}_{over} = \tilde{\boldsymbol{\theta}}^{u}_{over} - \mathbf{b}_o$,where

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}^{\mathrm{u}}_{\mathrm{over}} &= \arg\max_{\boldsymbol{\theta}} \ell^{\mathrm{u}}_{\mathrm{over}}(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \tau_i \left[y_i \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta} - \log\left\{ 1 + e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta}} \right\} \right],
\end{aligned} \tag{20}
$$

and

$$
\mathbf{b}_o = (b_{o0}, 0, \ldots, 0)^{\mathrm{T}} = \{\log(1+\lambda_n), 0, \ldots, 0\}^{\mathrm{T}}. \tag{21}
$$

The following theorem is about the asymptotic distribution of $\hat{\boldsymbol{\theta}}^{ubc}_{over}$.

Over-sampled Unweighted Estimator

Theorem 5

If $\mathbb{E}(e^{t\|\mathbf{x}\|}) < \infty$ for any $t > 0, \mathbb{E}(e^{\boldsymbol{\theta}_{nt}^{\mathrm{T}}\mathbf{x}}\mathbf{z}\mathbf{z}^{\mathrm{T}})$ is a positive-define matrix, and $\lambda_n \to \lambda \geq 0$, and $\lambda_n e^{\alpha_{nt}} \to c_o$ for a constant $c_o \in [0, \infty)$, then under the condition in (2) and (3), as $n \to \infty$,

$$\sqrt{n_1}\left(\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{ubc}} - \boldsymbol{\theta}_{nt}\right) \longrightarrow \mathbb{N}\left(\mathbf{0}, \mathbf{V}_{\text{over}}^{\text{ubc}}\right), \tag{22}$$

$$\mathbf{V}_{\textsf{over}}^{\textsf{ubc}} = \frac{(1+\lambda)^2+\lambda}{(1+\lambda)^2}\mathbb{E}\left(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\right)\mathbf{M}_{obc2}^{-1}\mathbf{M}_{obc1}\mathbf{M}_{obc2}^{-1}$$

$$\mathbf{M}_{obc1} = \mathbb{E}\left\{\frac{e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}}{\left(1+c_o e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}\right)^2}\mathbf{z}\mathbf{z}^{\mathrm{T}}\right\}, \quad \text{and}$$

$$\mathbf{M}_{obc2} = \mathbb{E}\left(\frac{e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}}{1+c_o e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\right).$$

Over-sampled Unweighted Estimator

Let $h = (1 + c_o e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}})^{-1}$ and $\mathbf{v} = e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}/2}(1 + c_o e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}})^{-1/2}\mathbf{z}$. Then in Proposition 1, we know that $\mathbf{V}_{\text{over}}^{\text{ubc}} \geq \mathbf{V}_{\text{over}}^{\text{w}}$.

If sampling has to be implemented, then we recommend using the weighted estimator $\hat{\boldsymbol{\theta}}_{\text{over}}^{\mathrm{w}}$.

Simulation: Full Data Estimator

Consider model (1) with one covariate x and $\boldsymbol{\theta} = (\alpha, \beta)^{\mathrm{T}}$. We set $\mathbb{P}(y = 1) = 0.02, 0.004, 0.0008, 0.00016$, and generate corresponding full data of size $n = 10^3, 10^4, 10^5, 10^6$. The covariates x_i's are generated from $\mathbb{N}(1, 1)$ for cases $(y_i = 1)$ and from $\mathbb{N}(0, 1)$ for controls $(y_i = 0)$. For the above setup,

- $\beta_t = 1$,
- $\alpha_{nt} = -4.39, -6.02, -7.63, -9.24$.

And the simulation for S = 1000 times and calculate empirical MSEs as eMSE($\hat{\theta}_j$) $= S^{-1} \sum_{s=1}^{S} (\hat{\theta}_j^{(s)} - \theta_{tj})^2, j = 0, 1$.

Simulation: Full Data Estimator

Table 1. Empirical MSE (eMSE) multiplied by $\mathbb{E}(n_1)$ and n.

n	$\mathbb{E}(n_1)$	$\mathbb{E}(n_1) \times \text{eMSE}\left(\hat{\theta}_j\right)$		$n \times \text{eMSE}\left(\hat{\theta}_j\right)$	
		$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
10^3	20	2.51	1.21	125.7	60.6
10^4	40	2.06	1.09	515.5	271.9
10^5	80	2.22	1.00	2774.4	1248.8
10^6	160	2.16	1.08	13474.9	6731.6

Sampling-based Estimators

Consider model (1) with $n = 10^5$, $x \sim \mathbb{N}(0,1)$ and $\boldsymbol{\theta}_{nt} = (-6,1)^{\mathrm{T}}$, so that $\mathbb{P}(y=1) \approx 0.004$.

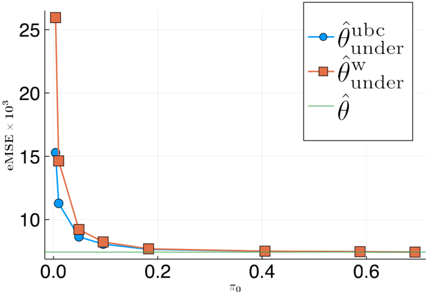
- Under-sampling: $\pi_0 = 0.05, 0.01, 0.05, 0.1, 0.2, 0.5, 0.8, 1.0$;
- Over-sampling: $\log(1+\lambda_n) = 0, 0.2, 0.4, 0.8, 1.5, 2.0, 2.5, 4.0$.

We repeat the simulation for S = 1000 times and calculate empirical MSEs as

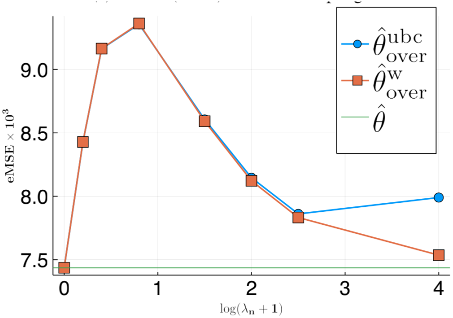
$$\mathrm{eMSE}\left(\hat{\boldsymbol{\theta}}_g\right) = \frac{1}{S}\sum_{s=1}^{S}\left\|\hat{\boldsymbol{\theta}}_g^{(s)} - \boldsymbol{\theta}_{nt}\right\|^2,$$

Note that if $\pi_0 = 1$ then the under-sampled estimators become the full data estimator, i.e., $\hat{\boldsymbol{\theta}}_{\mathrm{under}}^{\mathrm{w}} = \hat{\boldsymbol{\theta}}_{\mathrm{under}}^{\mathrm{ubc}} = \hat{\boldsymbol{\theta}}$;if $\lambda_n = 0$, then the over-sampled estimators become the full data estimator, i.e., $\hat{\boldsymbol{\theta}}_{\mathrm{over}}^{\mathrm{w}} = \hat{\boldsymbol{\theta}}_{\mathrm{over}}^{\mathrm{ubc}} = \hat{\boldsymbol{\theta}}$.

Sampling-based Estimators



(a) eMSEs ($\times 10^3$) for under-sampling



(b) eMSE for over-sampling

Figure: Empirical MSEs ($\times 10^3$) of under-sampled and over-sampled estimators. A smaller eMSE means that the corresponding estimator has a higher estimation efficiency

Future Work

- Multinomial logit models with rare events.
- Model averaging for logit models with rare events.