Logistic Regression for Massive Data with Rare Events

HaiYing Wang [1]

Abstract

This paper studies binary logistic regression for rare events data, or imbalanced data, where the number of events (observations in one class, often called cases) is significantly smaller than the number of nonevents (observations in the other class, often called controls). We first 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, indicating that the available information in rare events data is at the scale of the number of events instead of the full data sample size. Furthermore, we 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. This demonstrates the advantage of under-sampling nonevents for rare events data, because this procedure may significantly reduce the computation and/or data collection costs. Another common practice in analyzing rare events data is to over-sample (replicate) the events, which has a higher computational cost. We show that this procedure may even result in efficiency loss in terms of parameter estimation.

1. Introduction

Big data with rare events in binary responses, also called imbalanced data, are data in which the number of events (observations for one class of the binary response) is much smaller than the number of non-events (observations for the other class of the binary response). In this paper we also call the events "cases" and call the nonevents "controls". Rare events data are common in many scientific fields and applications. However, several important questions remain unanswered that are essential for valid data analysis and appropriate decision-making. For example, should we consider the amount of information contained in the data to be at the scale of the full-data sample size (very large) or the number of cases (relatively small)? Rare events data provide unique challenges and opportunities for sampling. On the one hand, sampling will not work without looking at responses because the probability of not selecting a rare case is high. On the other hand, since the rare cases are more informative than the controls, is it possible to use a small proportion of the full data to preserve most or all of the relevant information in the data about unknown parameters? A common practice when analyzing rare events data is to under-sample the controls and/or over-sample (replicate) the cases. Is there any information loss when using this approach? This paper provides a rigorous theoretical analysis on the aforementioned questions in the context of parameter estimation, which is critical in practical applications such as classification and statistical inference. Some answers may be counter-intuitive. For example, keeping all the cases, there may be no efficiency loss at all for under-sampling controls; on the other hand, using all the controls and over-sampling cases may reduce estimation efficiency.

Rare events data, or imbalanced data, have attracted a lot of attentions in machine learning and other quantitative fields, such as (Japkowicz, 2000; King & Zeng, 2001; Chawla et al., 2004; Estabrooks et al., 2004; Owen, 2007; Sun et al., 2007; Chawla, 2009; Rahman & Davis, 2013; Fithian & Hastie, 2014; Lemaître et al., 2017). A commonly implemented approach in practice is to try balancing the data by under-sampling controls (Drummond et al., 2003; Liu et al., 2009) and/or over-sampling cases (Chawla et al., 2002; Han et al., 2005; Mathew et al., 2017; Douzas & Bacao, 2017). However, most existing investigations focus on algorithms and methodologies for classification. Theoretical analyses of the effects of under-sampling and over-sampling in terms of parameter estimation are still rare.

(King & Zeng, 2001) considered logistic regression in rare events data and focused on correcting the biases in estimating the regression coefficients and probabilities. (Fithian & Hastie, 2014) utilized the special structure of logistic regression models to design a novel local case-control sampling method. These investigations obtained theoretical results

[1] Department of Statistics. Correspondence to: HaiYing Wang <haiying.wang@uconn.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

based on the the regular assumption that the probability of event occurring is fixed and does not go to zero. This assumption rules out the scenario of extremely imbalanced data, because for extremely imbalanced data, it is more appropriate to assume that the event probability goes to zero. (Owen, 2007)'s investigation did not require this fixed-probability assumption. He assumed that the number of rare cases is fixed, and derived the non-trivial point limit of the slope parameter estimator in logistic regression. However, the convergence rate and distributional properties of this estimator were not investigated. 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, and they grow large in rates that the number of cases divided by the number of controls decays to zero. This is the first study that provides distributional results for rare events data with a decaying event rate, and it gives the following indications.

- The convergence rate of the maximum likelihood estimator (MLE) is at the inverse of the number of cases instead of the total number of observations. This means that the amount of available information about unknown parameters in the data may be limited even the full data volume is massive.
- There maybe no efficiency loss at all in parameter estimation if one removes most of the controls in the data, because the control under-sampled estimators may have an asymptotic distribution that is identical to that of the full data MLE.
- Besides higher computational cost, over-sampling cases may result in estimation efficiency loss, because the asymptotic variances of the resulting estimators may be larger than that of the full data MLE.

The rest of the paper is organized as follows. We introduce the model setup and related assumptions in Section 2, and derive the asymptotic distribution for the full data MLE. We investigate under-sampled estimators in Section 3 and study over-sampled estimators in Section 4. Section 5 presents some numerical experiments, and Section 6 concludes the paper and points out some necessary future research. All the proofs of theoretical findings in this paper are presented in the supplementary material.

2. Model Setups and Assumptions

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

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

Here $\mathbf{x} \in \mathbb{R}^d$ is the covariate, $y \in \{0, 1\}$ is the binary class label, α is the intercept parameter, and $\boldsymbol{\beta}$ is the slope parameter vector. For ease of presentation, denote $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ as the full vector of regression coefficient, and define $\mathbf{z} = (1, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$ accordingly. This paper focuses on estimating the unknown $\boldsymbol{\theta}$.

If $\boldsymbol{\theta}$ is fixed (does not change with n changing), then model (1) is just the regular logistic regression model, and classical likelihood theory shows that the MLE based on the full data $\mathcal{D}_n$ converges at a rate of $n^{-1/2}$. A fixed $\boldsymbol{\theta}$ implies that $\mathbb{P}(y = 1) = \mathbb{E}\{\mathbb{P}(y = 1|\mathbf{x})\}$ is also a fixed constant bounded away from zero. However, for rare events data, because the event rate is so low in the data, it is more appropriate to assume that $\mathbb{P}(y = 1)$ approaches zero in some way. We discuss how to model this scenario in the following.

Let n_1 and n_0 be the numbers cases (observations with $y_i = 1$) and controls (observations with $y_i = 0$), respectively, in $\mathcal{D}_n$. 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. 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}, goes to negative infinity in a certain rate. Specifically, we assume $\alpha_{nt} \to -\infty$ as $n \to \infty$ in a rate such that

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

where $o_P(1)$ means a term that converges to zero in probability, i.e., a term that is arbitrarily small with probability approaching one. The assumption of a diverging α_{nt} with a fixed $\boldsymbol{\beta}_t$ means that the baseline probability of a rare event is low, and variation of covariate values does not change the order of the probability for a rare event to occur. This is a very reasonable assumption for many practical problems. For example, although making phone calls when driving may increase the probability of car accidents, it may not make car accidents a high-probability event.

2.1. How Much Information Do We Have in Rare Events Data

To demonstrate how much information is really available in rare events data, we derive the asymptotic distribution of the MLE for model (1) in the scenario described in (2) and (3). The MLE based on the full data $\mathcal{D}_n$, say $\hat{\boldsymbol{\beta}}$, is the maximizer of

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

which is also the solution to the following equation,

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

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

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

Theorem 1. *If* $\mathbb{E}(e^{t\|\mathbf{x}\|}) < \infty$ *for any* $t > 0$ *and* $\mathbb{E}(e^{\boldsymbol{\beta}_t^{\mathrm{T}} \mathbf{x}} \mathbf{z}\mathbf{z}^{\mathrm{T}})$ *is a positive-definite matrix, then under the conditions in (2) and (3), as* $n \to \infty$,

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

in distribution, where

$$\mathbf{V}_f = \mathbb{E}\big(e^{\boldsymbol{\beta}_t^{\mathrm{T}} \mathbf{x}}\big) \mathbf{M}_f^{-1}, \qquad \textit{and} \tag{7}$$

$$\mathbf{M}_f = \mathbb{E}\big(e^{\boldsymbol{\beta}_t^{\mathrm{T}} \mathbf{x}} \mathbf{z}\mathbf{z}^{\mathrm{T}}\big) = \mathbb{E}\left\{ e^{\boldsymbol{\beta}_t^{\mathrm{T}} \mathbf{x}} \begin{pmatrix} 1 & \mathbf{x}^{\mathrm{T}} \\ \mathbf{x} & \mathbf{x}\mathbf{x}^{\mathrm{T}} \end{pmatrix} \right\}. \tag{8}$$

Remark 1. The result in (6) shows that the convergence rate of the full-data MLE is at the order of $n_1^{-1/2}$, i.e, $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{nt} = O_P(n_1^{-1/2})$. This is different from the classical result of $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{nt} = O_P(n^{-1/2})$ for the case that $\mathbb{P}(y = 1)$ is a fixed constant. Theorem 1 indicates that for rare events data, the real amount of available information is actually at the scale of n_1 instead of n. A large volume of data does not mean that we have a large amount of information.

3. Efficiency of Under-sampled Estimators

Theorem 1 in the previous section shows that the full-data MLE has a convergence rate of $n_1^{-1/2}$. If we under-sample controls to reduce the number of controls to the same level of n_1, whether the resulting estimator has the full-data estimator convergence rate of $n_1^{-1/2}$? If so, one can significantly improve the computational efficiency and reduce the storage requirement for massive data. Furthermore, will under-sampling controls causes any estimation efficiency loss (an enlarged asymptotic variance)? This section answers the aforementioned questions.

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

$$\delta_i = y_i + (1 - y_i) I(u_i \le \pi_0), \quad i = 1, ..., n, \tag{9}$$

where $u_i \sim \mathbb{U}(0, 1)$, $i = 1, ..., n$, are independent and identically distributed (i.i.d.) random variables with the standard uniform distribution. This is a mixture of deterministic selection and random sampling. The resulting control under-sampled data include all rare cases (with $y_i = 1$) and the number of controls (with $y_i = 0$) is on average at the order of $n_0 \pi_0$. The average sample size for the under-sampled data given the full-data is $\sum_{i=1}^{n} \mathbb{E}(\delta_i | \mathcal{D}_n) = n_1 + n_0 \pi_0$, which is $o_p(n)$ if $\pi_0 \to 0$. The average sample size reduction is $n_0(1 - \pi_0)$ which is at the same order of n if $\pi_0 \nrightarrow 1$, and $n_0(1 - \pi_0)/n \to 1$ if $\pi_0 \to 0$.

Note that the under-sampled data taken according to δ_i in (9) is a biased sample, so we need to maximize a weighted objective function to obtain an asymptotically unbiased estimator. Alternatively, we can maximize an unweighted objective function and then correct the bias for the resulting estimator in logistic regression.

3.1. Under-sampled Weighted Estimator

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

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

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

$$\ell_{\mathrm{under}}^{\mathrm{w}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \big\{ y_i \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta} - \log(1 + e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta}}) \big\}. \tag{10}$$

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

Theorem 2. *If* $\mathbb{E}(e^{t\|\mathbf{x}\|}) < \infty$ *for any* $t > 0$, $\mathbb{E}\big(e^{\boldsymbol{\theta}_{nt}^{\mathrm{T}} \mathbf{x}} \mathbf{z}\mathbf{z}^{\mathrm{T}}\big)$ *is a positive-definite 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* $n \to \infty$,

$$\sqrt{n_1} (\hat{\boldsymbol{\theta}}_{\mathrm{under}}^{\mathrm{w}} - \boldsymbol{\theta}_{nt}) \longrightarrow \mathbb{N}(\mathbf{0},\ \mathbf{V}_{\mathrm{under}}^{\mathrm{w}}), \tag{11}$$

in distribution, where

$$\mathbf{V}_{\mathrm{under}}^{\mathrm{w}} = \mathbb{E}(e^{\boldsymbol{\beta}_t^{\mathrm{T}} \mathbf{x}}) \mathbf{M}_f^{-1} \mathbf{M}_{\mathrm{under}}^{\mathrm{w}} \mathbf{M}_f^{-1}, \quad \textit{and} \tag{12}$$

$$\mathbf{M}_{\mathrm{under}}^{\mathrm{w}} = \mathbb{E}\big\{ e^{\boldsymbol{\beta}_t^{\mathrm{T}} \mathbf{x}} (1 + c e^{\boldsymbol{\beta}_t^{\mathrm{T}} \mathbf{x}}) \mathbf{z}\mathbf{z}^{\mathrm{T}} \big\}. \tag{13}$$

Remark 2. 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

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

Since $n_0\pi_0$ is the average number of the controls in the under-sampled data, $c\mathbb{E}(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}})$ 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 controls than cases.

Theorem 2 requires that $0 \leq c < \infty$. This means that the number of controls to take should not be significantly smaller than the number of cases, which is a very reasonable assumption.

Remark 3. Theorem 2 shows that as long as π_0 does not make the number of controls in the under-sampled data much smaller than the number of cases n_1, then the under-sampled estimator $\hat{\boldsymbol{\theta}}^{\mathrm{w}}_{\mathrm{under}}$ preserves the convergence rate of the full-data estimator. Furthermore, if $c = 0$ then $\mathbf{M}^{\mathrm{w}}_{\mathrm{under}} = \mathbf{M}_f$, which implies that $\mathbf{V}^{\mathrm{w}}_{\mathrm{under}} = \mathbf{V}_f$. This means that if one takes much more controls than cases, then asymptotically there is no estimation efficiency loss at all. Here, the number of controls to take can still be significantly smaller than n_0 so that the computational burden is significantly reduced. If $c > 0$, since $\mathbf{M}^{\mathrm{w}}_{\mathrm{under}} > \mathbf{M}_f$, we know that $\mathbf{V}^{\mathrm{w}}_{\mathrm{under}} > \mathbf{V}_f$, in the Loewner order[1]. Thus reducing the number of controls to the same order of the number of cases may reduce the estimation efficiency, although the convergence rate is the same as that of the full-data estimator.

3.2. Under-sampled Unweighted Estimator with Bias Correction

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

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

then in $\tilde{\boldsymbol{\theta}}^{\mathrm{u}}_{\mathrm{under}} = (\hat{\alpha}^{\mathrm{u}}_{\mathrm{under}}, \hat{\boldsymbol{\beta}}^{\mathrm{u}}_{\mathrm{under}}{}^{\mathrm{T}})^{\mathrm{T}}$, the intercept estimator $\hat{\alpha}^{\mathrm{u}}_{\mathrm{under}}$ is asymptotically biased while the slope estimator $\hat{\boldsymbol{\beta}}^{\mathrm{u}}_{\mathrm{under}}$ is still asymptotically unbiased. The bias in the intercept estimator can be corrected using $\log(\pi_0)$ (Fithian & Hastie, 2014; Wang, 2019). We use this bias correction term and define the under-sampled **u**nweighted estimator with **b**ias **c**orrection $\hat{\boldsymbol{\theta}}^{\mathrm{ubc}}_{\mathrm{under}}$ as

$$\hat{\boldsymbol{\theta}}^{\mathrm{ubc}}_{\mathrm{under}} = \tilde{\boldsymbol{\theta}}^{\mathrm{u}}_{\mathrm{under}} + \mathbf{b}, \tag{14}$$

where

$$\mathbf{b} = \{\log(\pi_0), 0, ..., 0\}^{\mathrm{T}}. \tag{15}$$

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

Theorem 3. *If $\mathbb{E}(e^{t\|\mathbf{x}\|}) < \infty$ for any $t > 0$, $\mathbb{E}\big(e^{\boldsymbol{\theta}_{nt}^{\mathrm{T}}\mathbf{x}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\big)$ is a positive-definite 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 $n \to \infty$,*

$$\sqrt{n_1}(\hat{\boldsymbol{\theta}}^{\mathrm{ubc}}_{\mathrm{under}} - \boldsymbol{\theta}_{nt}) \longrightarrow \mathbb{N}(\mathbf{0},\ \mathbf{V}^{\mathrm{ubc}}_{\mathrm{under}}), \tag{16}$$

in distribution, where

$$\mathbf{V}^{\mathrm{ubc}}_{\mathrm{under}} = \mathbb{E}(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}})(\mathbf{M}^{\mathrm{ubc}}_{\mathrm{under}})^{-1}, \quad \textit{and} \tag{17}$$

$$\mathbf{M}^{\mathrm{ubc}}_{\mathrm{under}} = \mathbb{E}\bigg(\frac{e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}}{1 + ce^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\bigg). \tag{18}$$

Remark 4. Similarly to the case of under-sampled weighted estimator, Theorem 3 shows that the estimator $\hat{\boldsymbol{\theta}}^{\mathrm{ubc}}_{\mathrm{under}}$ preserves the same convergence rate of the full-data estimator if $c < \infty$. Furthermore, if $c > 0$ then $\mathbf{V}^{\mathrm{ubc}}_{\mathrm{under}} > \mathbf{V}_f$; if $c = 0$, then $\mathbf{V}^{\mathrm{ubc}}_{\mathrm{under}} = \mathbf{V}_f$.

The following proposition is useful to compare the asymptotic variances of the weighted and the unweighted estimators.

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 $\mathbb{E}(h^{-1}\mathbf{v}\mathbf{v}^{\mathrm{T}})$ are all finite and positive-definite matrices. The following inequality holds in the Loewner order.*

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

Remark 5. If we let $\mathbf{v} = e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}/2}\mathbf{z}$ and $h = 1 + ce^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}$ in Proposition 1, then we know that $\mathbf{V}^{\mathrm{ubc}}_{\mathrm{under}} \leq \mathbf{V}^{\mathrm{w}}_{\mathrm{under}}$ in the Loewner order. This indicates that with the same control under-sampled data, the unweighted estimator with bias correction, $\hat{\boldsymbol{\theta}}^{\mathrm{ubc}}_{\mathrm{under}}$, has a higher estimation efficiency than the weighted estimator, $\hat{\boldsymbol{\theta}}^{\mathrm{w}}_{\mathrm{under}}$.

4. Efficiency Loss Due to Over-sampling

Another common practice to analyze rare events data is to use all the controls and over-sample the cases. To investigate

[1] For two Hermitian matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ of the same dimension, $\mathbf{A}_1 \geq \mathbf{A}_2$ if $\mathbf{A}_1 - \mathbf{A}_2$ is positive semi-definite and $\mathbf{A}_1 > \mathbf{A}_2$ if $\mathbf{A}_1 - \mathbf{A}_2$ is positive definite.

the effect of this approach, let τ_i denote the number of times that a data point is used, and define

$$\tau_i = y_i v_i + 1, \quad i = 1, ..., n, \tag{19}$$

where $v_i \sim \mathbb{POI}(\lambda_n)$, $i = 1, ..., n$, are i.i.d. Poisson random variables with parameter λ_n. 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 $\mathbb{E}(\tau_i|\mathcal{D}_n, y_i = 1) = 1 + \lambda_n$ times. Here, λ_n can be interpreted as the average over-sampling rate for cases. We let v_i be Poisson random variables because this is commonly used, and the over-sampling plan is called Poisson sampling. However, the number of times that the cases are over-sampled do not have to follow a Poisson distribution. Any distribution works under some moment conditions if its support is the set of non-negative integers. For a different distribution, the leading term on the right hand side of equation (22) will be different, but it is always larger than or equal to one.

Again, the case over-sampled data according to (19) is a biased sample, and we need to use a weighted objective function or to correct the bias of the estimator form an unweighted objective function.

4.1. Over-sampled Weighted Estimator

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

$$\ell^{\mathrm{w}}_{\mathrm{over}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\tau_i}{w_i} \big\{ y_i \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta} - \log(1 + e^{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta}}) \big\}. \tag{20}$$

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

Theorem 4. *If $\mathbb{E}(e^{t\|\mathbf{x}\|}) < \infty$ for any $t > 0$, $\mathbb{E}\big(e^{\boldsymbol{\theta}_{nt}^{\mathrm{T}}\mathbf{x}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\big)$ is positive-definite, and $\lambda_n \to \lambda \geq 0$, then under the conditions in (2) and (3), as $n \to \infty$,*

$$\sqrt{n_1}(\hat{\boldsymbol{\theta}}^{\mathrm{w}}_{\mathrm{over}} - \boldsymbol{\theta}_{nt}) \longrightarrow \mathbb{N}(\mathbf{0}, \mathbf{V}^{\mathrm{w}}_{\mathrm{over}}), \tag{21}$$

in distribution, where

$$\mathbf{V}^{\mathrm{w}}_{\mathrm{over}} = \frac{(1+\lambda)^2 + \lambda}{(1+\lambda)^2} \mathbb{E}(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}) \mathbf{M}_f^{-1}. \tag{22}$$

Remark 6. *Note that in (22), $\frac{(1+\lambda)^2+\lambda}{(1+\lambda)^2} \geq 1$ and the equality holds only if $\lambda = 0$ or $\lambda = \infty$. Thus, $\mathbf{V}^{\mathrm{w}}_{\mathrm{over}} \geq \mathbf{V}_f$, meaning that over-sampling the cases may result in estimation efficiency loss unless the number of over-sampled cases is small enough to be negligible ($\lambda = 0$) or it is very large ($\lambda = \infty$). Considering that over-sampling incurs additional computational cost with potential estimation efficiency loss, this procedure is not recommended if the primary goal is parameter estimation.*

4.2. Over-sampled Unweighted Estimator with Bias Correction

For completeness, we derive the asymptotic distribution of the over-sampled unweighted estimator with **b**ias **c**orrection, $\hat{\boldsymbol{\theta}}^{\mathrm{ubc}}_{\mathrm{over}}$, defined as $\hat{\boldsymbol{\theta}}^{\mathrm{ubc}}_{\mathrm{over}} = \tilde{\boldsymbol{\theta}}^{\mathrm{u}}_{\mathrm{over}} - \mathbf{b}_o$, where

$$\begin{aligned} \tilde{\boldsymbol{\theta}}^{\mathrm{u}}_{\mathrm{over}} &= \arg\max_{\boldsymbol{\theta}} \ \ell^{\mathrm{u}}_{\mathrm{over}}(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \tau_i \big[y_i \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta} - \log\{1 + e^{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta}}\} \big], \end{aligned} \tag{23}$$

and

$$\mathbf{b}_o = (b_{o0}, 0, ..., 0)^{\mathrm{T}} = \{\log(1+\lambda_n), 0, ..., 0\}^{\mathrm{T}}. \tag{24}$$

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

Theorem 5. *If $\mathbb{E}(e^{t\|\mathbf{x}\|}) < \infty$ for any $t > 0$, $\mathbb{E}\big(e^{\boldsymbol{\theta}_{nt}^{\mathrm{T}}\mathbf{x}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\big)$ is positive-definite, $\lambda_n \to \lambda \geq 0$, and $\lambda_n e^{\alpha_{nt}} \to c_o$ for a constant $c_o \in [0, \infty)$, then under the conditions in (2) and (3), as $n \to \infty$,*

$$\sqrt{n_1}(\hat{\boldsymbol{\theta}}^{\mathrm{ubc}}_{\mathrm{over}} - \boldsymbol{\theta}_{nt}) \longrightarrow \mathbb{N}(\mathbf{0}, \ \mathbf{V}^{\mathrm{ubc}}_{\mathrm{over}}), \tag{25}$$

in distribution, where

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

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

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

Remark 7. Unlike the case of under-sampled estimators, for over-sampled estimators, the unweighted estimator with bias correction $\hat{\boldsymbol{\theta}}^{\mathrm{ubc}}_{\mathrm{over}}$ has a lower estimation efficiency than the weighted estimator $\hat{\boldsymbol{\theta}}^{\mathrm{w}}_{\mathrm{over}}$. To see this, letting $h = (1 + c_o e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}})^{-1}$ and $\mathbf{v} = e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}/2}(1 + c_o e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}})^{-1/2}\mathbf{z}$ in Proposition 1, we know that $\mathbf{V}^{\mathrm{ubc}}_{\mathrm{over}} \geq \mathbf{V}^{\mathrm{w}}_{\mathrm{over}}$, and the equality holds if $c_o = 0$. Here, since $\lambda_n e^{\alpha_{nt}} \mathbb{E}(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}}) = \frac{n_1 \lambda_n}{n_0}\{1 + o_P(1)\}$, we can intuitively interpret $c_o \mathbb{E}(e^{\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}})$ as the ratio of the average times of over-sampled cases to the number of controls. If in addition $\lambda = 0$, then $\mathbf{V}^{\mathrm{ubc}}_{\mathrm{over}} = \mathbf{V}^{\mathrm{w}}_{\mathrm{over}} = \mathbf{V}_f$; but in general, $\mathbf{V}^{\mathrm{ubc}}_{\mathrm{over}} \geq \mathbf{V}^{\mathrm{w}}_{\mathrm{over}} \geq \mathbf{V}_f$.

Remark 8. Compared with Theorem 4 for $\hat{\boldsymbol{\theta}}^{\mathrm{w}}_{\mathrm{under}}$, Theorem 5 for $\hat{\boldsymbol{\theta}}^{\mathrm{ubc}}_{\mathrm{over}}$ requires an extra condition that $\lambda_n e^{\alpha_{nt}} \to c_o \in [0, \infty)$. In addition, $\mathbf{V}^{\mathrm{ubc}}_{\mathrm{over}} \geq \mathbf{V}^{\mathrm{w}}_{\mathrm{over}}$. Thus, if over-sampling has to be implemented, then we recommend using the weighted estimator $\hat{\boldsymbol{\theta}}^{\mathrm{w}}_{\mathrm{over}}$.

The relative efficiency between the under-sampled and over-sampled estimators is complicate as it depends on the parameter $\boldsymbol{\beta}$, the values of c and λ, and the distribution of the covariate. Thus a general explicit expressions is hard to obtain.

However, we still have some interesting results for some special situation. For example, if $c(1+\lambda)^2\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})<\lambda$ almost surely, then the unweighted under-sampled estimator is more efficient than the weighted over-sampled estimator, and *vice versa*.

5. Numerical Experiments

5.1. Full Data Estimator $\hat{\boldsymbol{\theta}}$

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 and 0.00016, and generate corresponding full data of sizes $n=10^3$, 10^4, 10^5 and 10^6, respectively. As a result, the average numbers of cases ($y_i=1$) in the resulting data are $\mathbb{E}(n_1)=20$, 40, 80 and 160. The above value configuration aims to mimic the scenario that $n\to\infty$, $\mathbb{P}(y=1)\to 0$, and $\mathbb{E}(n_1)\to\infty$. 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, the true value of β is fixed $\beta_t=1$, and the true values of α are $\alpha_{nt}=-4.39$, -6.02, -7.63 and -9.24, respectively for the four different values of n. We repeat the simulation for $S=1,000$ times and calculate empirical MSEs as $\mathrm{eMSE}(\hat{\theta}_j)=S^{-1}\sum_{s=1}^{S}(\hat{\theta}_j^{(s)}-\theta_{tj})^2$, $j=0,1$, where $\hat{\theta}_0=\hat{\alpha}$, $\hat{\theta}_1=\hat{\beta}$, and $\hat{\theta}_j^{(s)}$ is the estimate from the s-th repetition.

Table 1 presents empirical MSEs (eMSEs) multiplied by $\mathbb{E}(n_1)$ and n, respectively. We see that $\mathbb{E}(n_1)\times\mathrm{eMSE}(\hat{\theta}_j)$ does not diverge as n increases for both $\hat{\alpha}$ and $\hat{\beta}$. This confirms the conclusion in Theorem 1 that $\hat{\boldsymbol{\theta}}$ converges at a rate of $n_1^{-1/2}$ (It implies that $n_1\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{nt}\|^2=O_P(1)$). On the other hand, values of $n\times\mathrm{eMSE}(\hat{\theta}_j)$ are large, and they increase fast as n increases, indicating that $n\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{nt}\|^2$ diverges to infinity. Table 1 confirms that although the values of the full data sample sizes n are very large, it is the values of n_1 that reflect the real amount of available information about regression parameters, and they are actually much smaller.

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

n	$\mathbb{E}(n_1)$	$\mathbb{E}(n_1)\times\mathrm{eMSE}(\hat{\theta}_j)$		$n\times\mathrm{eMSE}(\hat{\theta}_j)$	
		$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
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

5.2. Sampling-based Estimators

Now we provide numerical results about under-sampled and over-sampled estimators. Consider model (1) with $n=10^5$, $x\sim\mathbb{N}(0,1)$, and $\boldsymbol{\theta}_{nt}=(-6,1)^{\mathrm{T}}$, so that $\mathbb{P}(y=1)\approx 0.004$. For under-sampling, consider $\pi_0=0.005$, 0.01, 0.05, 0.1, 0.2, 0.5, 0.8, and 1.0; for over-sampling, consider $\lambda_n=0, 0.22, 0.49, 1.23, 3.48, 6.39, 11.18$ and 53.6, which corresponds to $\log(1+\lambda_n)=0$, 0.2, 0.4, 0.8, 1.5, 2.0, 2.5 and 4.0, respectively. We repeat the simulation for $S=1,000$ times and calculate empirical MSEs as

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

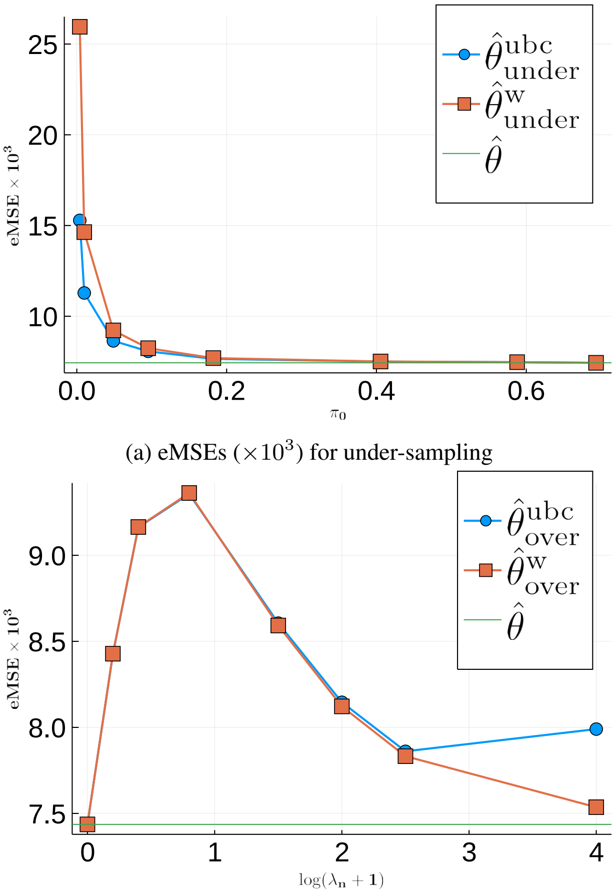
where $\hat{\boldsymbol{\theta}}_g^{(s)}$ is the estimate from the s-th repetition for some estimator $\hat{\boldsymbol{\theta}}_g$. We consider $\hat{\boldsymbol{\theta}}_g=\hat{\boldsymbol{\theta}}_{\text{under}}^{\text{w}}$, $\hat{\boldsymbol{\theta}}_{\text{under}}^{\text{ubc}}$, $\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{w}}$, and $\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{ubc}}$. Note that if $\pi_0=1$ then the under-sampled estimators become the full data estimator, i.e., $\hat{\boldsymbol{\theta}}_{\text{under}}^{\text{w}}=\hat{\boldsymbol{\theta}}_{\text{under}}^{\text{ubc}}=\hat{\boldsymbol{\theta}}$; if $\lambda_n=0$, then the over-sampled estimators become the full data estimator, i.e., $\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{w}}=\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{ubc}}=\hat{\boldsymbol{\theta}}$.

Figure 1 presents the simulation results. Figure 1 (a) plots eMSEs ($\times 10^3$) against π_0. When π_0 is small, the number of controls in under-sampled data is small, and the resulting estimators are not as efficient as the full-data estimator. For example, when $\pi_0=0.005$, the numbers of cases and the numbers of controls are roughly the same, and we do see significant information loss in this case. However, when π_0 gets larger, under-sampled estimators becomes more efficient, and when $\pi_0>0.1$, the performances of the under-sampled estimators are almost as good as the full-data estimator. In addition, the unweighted estimator $\hat{\boldsymbol{\theta}}_{\text{under}}^{\text{ubc}}$ is more efficient than the weighted estimator $\hat{\boldsymbol{\theta}}_{\text{under}}^{\text{w}}$ for smaller π_0's, and they both perform more similarly to the full data estimator $\hat{\boldsymbol{\theta}}$ as π_0 grows. These observations are consistent with the conclusions in Theorems 2 and 3, and the discussions in the relevant remarks.

Figure 1 (b) plots eMSEs ($\times 10^3$) against $\log(\lambda_n+1)$. We see that the case over-sampled estimators are less efficient than the full data estimator unless the average number of over-sampled cases λ_n is very small or very large. For small λ_n, $\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{w}}$ and $\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{ubc}}$ perform similarly, but $\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{w}}$ is more efficient than $\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{ubc}}$ for large λ_n. The reason of this phenomenon is that if λ_n is large, then the required condition of $\lambda_n e^{\alpha_{nt}}\to c_o\in[0,\infty)$ in Theorem 5 for $\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{ubc}}$ may not be valid. This confirms our recommendation that the weighted estimator $\hat{\boldsymbol{\theta}}_{\text{over}}^{\text{w}}$ is preferable if over-sampling has to be used.

6. Discussion and Future Research

In this paper, we have obtained distributional results showing that the amount of information contained in massive data with rare events is at the scale of the relatively small total number of cases rather than the large total number of observations. We have further demonstrated that aggressively under-sampling the controls may not sacrifice the estimation



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

(b) eMSE for over-sampling

Figure 1. Empirical MSEs ($\times 10^3$) of under-sampled and over-sampled estimators. The eMSE ($\times 10^3$) for the full data estimator $\hat{\boldsymbol{\theta}}$ (the horizontal line) is also plotted for comparison. A smaller eMSE means that the corresponding estimator has a higher estimation efficiency.

efficiency at all while over-sampling the cases may reduce the estimation efficiency.

We end the paper by pointing out some possible future research topics: 1) Although the current paper focuses on the logistic regression model, we conjecture that our conclusions are generally true for rare events data and will investigate more complicated and general models in future research projects. 2) As another direction, more comprehensive numerical experiments are helpful to gain further understandings on parameter estimation with imbalanced data. This paper has focused on point estimation. How to make valid and more accurate statistical inference with rare events data still need further research. 3) There is a long standing literature investigating the effects of under-sampling and over-sampling in classification. However, most investigations adopted an empirical approach, so theoretical investigations on the effects of sampling are still needed for classification. 4) In this paper, we use the same probability to sample the controls, it is interesting to analyze a setting where the probabilities are different for different data points in the controls. It is also interesting to analyze what happens if both under-sampling and over-sampling are performed concurrently.

References

Chawla, N. V. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pp. 875–886. Springer, 2009.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Chawla, N. V., Japkowicz, N., and Kotcz, A. Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004.

Douzas, G. and Bacao, F. Self-organizing map oversampling (somo) for imbalanced data set learning. *Expert systems with Applications*, 82:40–52, 2017.

Drummond, C., Holte, R. C., et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pp. 1–8. Citeseer, 2003.

Estabrooks, A., Jo, T., and Japkowicz, N. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36, 2004.

Fithian, W. and Hastie, T. Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of statistics*, 42(5):1693, 2014.

Han, H., Wang, W.-Y., and Mao, B.-H. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In Huang, D.-S., Zhang, X.-P., and Huang, G.-B. (eds.), *Advances in Intelligent Computing*, pp. 878–887, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

Japkowicz, N. Learning from imbalanced data sets: Papers from the AAAI workshop, AAAI, 2000. Technical Report WS-00-05, 2000. URL https://aaai.org/Library/Workshops/ws00-05.php.

King, G. and Zeng, L. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

Lemaître, G., Nogueira, F., and Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.

Liu, X., Wu, J., and Zhou, Z. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

Mathew, J., Pang, C. K., Luo, M., and Leong, W. H. Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE transactions on neural networks and learning systems*, 29(9):4065–4076, 2017.

Owen, A. B. Infinitely imbalanced logistic regression. *The Journal of Machine Learning Research*, 8:761–773, 2007.

Rahman, M. M. and Davis, D. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224, 2013.

Sun, Y., Kamel, M. S., Wong, A. K., and Wang, Y. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007.

Wang, H. More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research*, 20(132):1–59, 2019.