Finding Needles in Haystacks: A Comparative Study of Algorithms for Handling Imbalanced Data in Logistic Regression Classification

Master's Thesis presented to the
Department of Economics at the
Rheinische Friedrich-Wilhelms-Universität Bonn

In Partial Fulfillment of the Requirements for the Degree of
Master of Science (M.Sc.)

Supervisor: Prof. Dr. Dominik Liebl

Submitted in June 2023 by:
Carolina Alvarez
Matriculation Number: 3288942

Contents

1 Introduction **1**

2 Models for binary data **3**

2.1 Problem setting 3

2.2 Logistic regression 3

2.3 Estimation of $\widehat{\theta}_{MLE}$ 4

2.4 How large samples affect the computation of $\widehat{\theta}_{MLE}$ 6

3 Subsampling methods **7**

3.1 The imbalance problem 7

3.2 Case-control subsampling 9

3.2.1 Intercept adjustment 9

3.2.2 Consistency of the CC estimator 12

3.3 Weighted case-control subsampling 13

3.4 Local case-control subsampling 15

3.4.1 Preliminaries for asymptotic results 17

3.4.2 Consistency of $\hat{\theta}_{LCC}$ 18

3.4.3 Asymptotic distribution 19

4 Evaluation metrics **22**

4.1 The bias-variance decomposition 22

4.2 Obtaining the true intercept value 22

5 Simulation study **23**

5.1 General setup 23

5.2 Simulation 1 24

5.3 Simulation 2 26

5.4 Simulation 3 27

5.5 Simulation 4 29

6 Data application **34**

6.1 Adult income dataset 34

6.2 Methodological considerations 35

6.3 Results 37

7 Conclusion 39

8 Bibliography 41

9 Appendix i

9.1 Formal derivations i

9.1.1 Formal derivation of intercept adjustment in Section 3.2 i

9.2 Additional tables ii

9.2.1 Summary statistics for LCC subsample sizes across simulations . . . ii

9.2.2 Additional tables from simulation section iv

9.2.3 Additional tables from data application viii

9.3 Additional R packages ix

List of Figures

1 Graphical example of the between-class imbalance problem. 2

2 PCA plot (DGP example) 24

3 Simulation 1 - Distribution of Monte Carlo realizations $\hat{\theta}_{CC}$, $\hat{\theta}_{WCC}$ and $\hat{\theta}_{LCC}$ 26

4 Simulation 2 - Distribution of Monte Carlo realizations $\hat{\theta}_{logit}$, $\hat{\theta}_{LCC}$ and theoretical $\hat{\theta}_{LCC}$ 28

5 Simulation 4 - Relationship between empirical squared bias and subsample size by method 32

6 Simulation 4 - Relationship between empirical variance and subsample size by method 33

7 Data application - PCA plot Income dataset 36

8 Data application - Relative variance of coefficients 38

List of Tables

1 Simulation 1 - Baseline numerical exercise 25

2 Simulation 2 - Approximation of asymptotic behavior of LCC against logit 27

3 Simulation 3 - Increasing marginal imbalance for $N = 10^5$ and $k = 30$ . . . 30

4 Simulation 3 - Increasing marginal imbalance for $N = 10^5$ and $k = 10$ . . . 31

5 Summary Statistics - Income dataset (missing values removed) 35

6 Data application - Empirical variance results for the Income dataset 37

7 Simulation 1 - Summary statistics for LCC's subsample size. ii

8 Simulation 3 - Summary statistics for LCC's subsample size. iii

9 Simulation 4 - Summary statistics for LCC's subsample size. iv

10 Simulation 3 - Increasing marginal imbalance for $N = 10^5$ and $k = 20$ . . . v

11 Simulation 4 - Results from variation in full population size N, holding $k = 10$ and $P(Y = 0) = 0.7$ constant v

12 Simulation 4 - Results from variation in full population size N, holding $k = 10$ and $P(Y = 0) = 0.8$ constant vi

13 Simulation 4 - Results from variation in full population size N, holding $k = 10$ and $P(Y = 0) = 0.9$ constant vii

14 Simulation 4 - Results from variation in full population size N, holding $k = 10$ and $P(Y = 0) = 0.95$ constant viii

15 Data application - Summary statistics for individuals with an annual income of more than 50K viii

16 Data application - Summary statistics for individuals with an annual income of 50K or less. ix

17 Data application - Fixed subsample sizes ix

1 Introduction

Over the past recent years, technological advances have led to a remarkable increase in data collection capabilities and database sizes (Wang, Yang, and Stufken 2019). Datasets with hundreds of features or thousands to millions of instances are now common across many research fields and industries (R. Li, Lin, and B. Li 2013). From an economic research perspective, Einav and Levin (2014) emphasize the use of large administrative datasets to create innovative research designs and predictive models that may be challenging or infeasible to develop with smaller sample sizes or larger data aggregation.

However, even though massive data brings new opportunities for understanding different phenomena through rich datasets, its enormous volume also imposes large computational and statistical challenges (Fan, F. Han, and Liu 2014). For instance, conducting optimal estimation of parameters and inference could become cumbersome when dealing with a massive amount of information (Ng 2017). The computational cost could be too high, or it might even become computationally infeasible to fit standard statistical models when the datasets are huge (L. Han et al. 2020).

One common approach is to downsize the data volume by subsampling data points and fitting the methods in a much smaller sample, thus reducing computational burden while still being able to make statistical inferences about the full sample parameters (J. Yu, Ai, and Ye 2023). Since the variance of the estimates will be larger in smaller samples, there is an inevitable trade-off between statistical efficiency and computational gains, and designing an effective subsampling scheme that reduces this trade-off becomes crucial (Lee and Ng 2020, L. Han et al. 2020). However, for imbalanced datasets, where the target variable shows a significantly unequal distribution between the classes, subsampling becomes more of a challenge (He and Garcia 2009). The most naive approach, uniform subsampling, will most likely fail to construct meaningful samples, as it assigns the same acceptance probability to each data point (Fithian and Hastie 2014, L. Han et al. 2020, Wang 2020, Yao and Wang 2021, Cheng, Wang, and Yang 2020).

First originated in epidemiology, case-control is a well-known subsampling design mostly used in medical research to assess risk factors related to rare diseases (Rose and Laan 2008). It improves upon uniform subsampling by having different acceptance probabilities for "cases" (minority class) and "controls" (majority class). A logistic regression is then fitted to the subsample, and estimates of the population parameters are retrieved after making an adjustment to the intercept. The case-control estimate is asymptotically consistent and unbiased in scenarios when there is high marginal imbalance, and the model is correctly specified. Additionally, weighted case-control estimate has been proven to be efficient under model misspecification and mild marginal imbalance (Prentice and Pyke 1979, King and Zeng 2001, A. Scott and Wild 1986, Shen, Chen, and W. Yu 2021). More recently, Fithian and Hastie (2014) proposed a local-case control subsampling design that aims to improve upon standard case-control methods by balancing the classes locally in the feature space via an accept–reject subsampling scheme. Their method is consistent even under misspecification, and it is argued to exploit the conditional imbalance in the data.

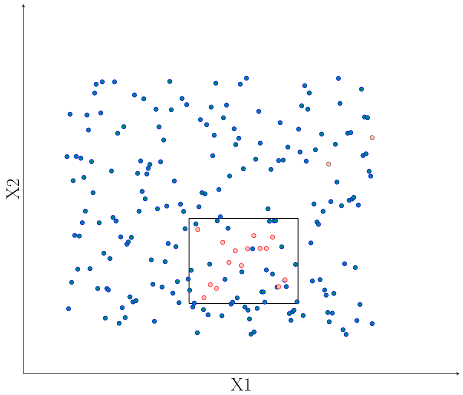

Figure 1: A two-dimensional graphical representation of a between-class imbalance problem based on He and Garcia 2009. The blue instances represent the majority class, while the pink observations the minority class.

The aim of this thesis is to study the statistical performance of the three above-mentioned case-control methods in approximating the true population parameters. Section 2 shows the estimation of the Maximum Likelihood logistic regression estimate and discusses how large samples could affect its computation. Section 3 introduces the subsampling algorithms under study and shows their underlying statistical assumptions for consistency and unbiasedness. Section 4 presents the metrics for evaluating the methods. Section 5 contains the numerical exercises conducted to compare the performance of the methods across different levels of marginal imbalance and sample sizes. It also shows an empirical analysis of the asymptotic properties of the local-case control, focusing on the assumption of a data-independent pilot for approximating its asymptotic distribution. Section 6 illustrates the performance of the methods in a real dataset, and Section 7 concludes.

2 Models for binary data

2.1 Problem setting

Assume N i.i.d samples where $i = 1, 2, \ldots, N$, each related to a covariate vector's values $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})' \in X$ and an outcome variable $y_i \in [0, 1]$ (Fithian and Hastie 2014). In the binary classification problem, where the number of classes C is reduced to only 2, the outcome variable follows a Bernoulli distribution $Y \sim \text{Bernoulli}(p(x_i))$ that takes on a value of 1 or "*success*" with probability $p(x_i)$, and a value of 0 or "*failure*" with probability $(1 - p(x_i))$ (King and Zeng 2001). Following McCullagh and Nelder (1989), the goal is to examine the relationship between the probability of an event occurring, $p(x_i)$, and the covariate vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$. Usually, the model is assumed to be a linear combination of the form:

$$\eta_i = \sum_{j=1}^{k} x_{ij}\beta_j \tag{1}$$

with unknown $(k \times 1)$ vector of coefficients $\beta = \beta_1, \beta_2, \ldots, \beta_k$. If no additional restrictions are set on β, η_i could span from $-\infty < \eta < +\infty$. This problem can be avoided by the transformation of η_i through $g(p(x_i))$, which leads to a generalized linear model of the form:

$$g(p(x_i)) = \eta_i = \sum_{j=1}^{k} x_{ij}\beta_j \tag{2}$$

(McCullagh and Nelder 1989).

2.2 Logistic regression

The choice of the logistic function to model $g(p(x_i))$ leads to a linear logistic model or logistic regression. It is a common tool in predictive classification and, as mentioned in the previous section, it allows to estimate the posterior probabilities of each class C via a linear model, while ensuring that the probabilities sum up to 1 and remain in $[0, 1]$ (Hastie et al. 2009). Assume the model has a constant term α and the number of covariates is k, such that both β and x_i are $(k \times 1)$ vectors. Then, let the probability of success of the ith individual $p(x_i)$ be defined as:

$$\mathbb{P}(Y=1|X=x)=p(x_i)=\frac{e^{\alpha+x_i'\beta}}{1+e^{\alpha+x_i'\beta}}=\left[1+e^{-(\alpha+x_i'\beta)}\right]^{-1} \tag{3}$$

(O. Montesinos, A. Montesinos, and Crossa 2022). An equivalent version of the model its representation in terms of the odds of a positive case, given by

$$\frac{p(x_i)}{1-p(x_i)}=e^{\alpha+x_i'\beta}, \tag{4}$$

which leads to its specification in terms of the log-odds or the logit transformation, providing the probability of a positive event occurring

$$\log\left(\frac{p(x_i)}{1-p(x_i)}\right)=f_\theta(x_i)=\alpha+x_i'\beta, \tag{5}$$

where $\theta=(\alpha,\beta)$ is a $(k+1)\times 1$ vector (McCullagh and Nelder 1989, Fithian and Hastie 2014). In a multi-class scenario with $C>2$ classes, the model turns into a system of $C-1$ log-odds equations with the last class used as the denominator in the odds ratios. The following methods focus on the binomial class problem $C=2$.

2.3 Estimation of $\widehat{\theta}_{MLE}$

Logistic regression is usually fit by Maximum Likelihood Estimation (MLE) (Hastie et al. 2009). The MLE estimator $\widehat{\theta}_{MLE}=(\widehat{\alpha},\widehat{\beta})$ is obtained through the likelihood function given by

$$L(\theta)=\prod_{i}^{N}p(x_i)^{y_i}\left(1-p(x_i)\right)^{1-y_i}, \tag{6}$$

with log-likelihood function

$$\begin{aligned}\ell(\theta)=\log[L(\theta)]&=\sum_{i=1}^{N}\left\{y_i\log p\left(x_i\right)+\left(1-y_i\right)\log\left(1-p\left(x_i\right)\right)\right\} &(7)\\ &=\sum_{i=1}^{N}y_i\left(\alpha+x_i'\beta\right)-\sum_{i=1}^{N}\log\left(1+e^{\alpha+x_i'\beta}\right) &(8)\end{aligned}$$

(O. Montesinos, A. Montesinos, and Crossa 2022). The objective is then to find the parameters' value that maximizes the log-likelihood

$$\widehat{\theta}_{MLE}=\arg\max\nolimits_\theta \ell(\theta).$$

To solve the maximization problem, one needs to set first the gradient of the likelihood function to zero:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{N} x_i(y_i - p(x_i)) = 0, \tag{9}$$

which turn out to be $k+1$ equations (Hastie et al. 2009). However, the maximization problem for logistic regression has no closed-form solution, and numerical optimizing methods are often used to find $\widehat{\beta}_{MLE}$ (Cheng, Wang, and Yang 2020). Thus, the gradients in 9 are solved by the Newton–Raphson Algorithm, an iterative optimization technique that uses a local-quadratic approximation to Equation 8 (O. Montesinos, A. Montesinos, and Crossa 2022). The Newton–Raphson method requires the second partial derivatives or the Hessian matrix of Equation 9:

$$\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'} = -\sum_{i=1}^{N} x_i x_i' p\left(x_i\right)\left(1 - p\left(x_i\right)\right) \tag{10}$$

(Hastie et al. 2009). One then sets an initial guess of the parameter values, $\theta^{(t)}$, and computes Equations 9 and 10 at this initial point. These estimations are then used to update the algorithm in what is called the Newton step $\theta^{(t+1)}$, until the convergence criteria are met:

$$\theta^{(t+1)} = \theta^{(t)} - \left(\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'}\right)^{-1} \frac{\partial \ell(\theta)}{\partial \theta} \tag{11}$$

(Hastie et al. 2009). The Newton step can also be re-expressed as a weighted least squares step. For this, consider Equations 9 and 10 in matrix notation, where $\mathbf{y}$ refers to the vector of y_i, $\mathbf{X}$ to the $N \times (k+1)$ matrix of x_i values, $\mathbf{p}$ to the vector of fitted probabilities $p(x_i)$ and $\mathbf{W}$ to a $N \times N$ diagonal matrix of weights, where the ith element takes up the value $p(x_i)(1-p(x_i))$ (Hastie et al. 2009). Then we have,

$$\frac{\partial \ell(\theta)}{\partial \theta} = \mathbf{X}'\left(\mathbf{y} - \mathbf{p}\right), \tag{12}$$

$$\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'} = -\mathbf{X}'\mathbf{W}\mathbf{X} \tag{13}$$

which can be used to express the Newton step as

$$\begin{aligned}
\theta^{(t+1)} &= \theta^{(t)} + \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y}-\mathbf{p}\right) \\
&= \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\theta^{(t)}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right) + \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y}-\mathbf{p}\right) \\
&= \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\theta^{t} + \mathbf{X}'\left(\mathbf{y}-\mathbf{p}\right)\right) \\
&= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\underbrace{\left(\mathbf{X}\theta^{(t)} + \mathbf{W}^{-1}\left(\mathbf{y}-\mathbf{p}\right)\right)}_{\mathbf{z}} \\
&= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z} \qquad (14)
\end{aligned}$$

(Hastie et al. 2009). This algorithm, developed from the Newton–Raphson algorithm and shown in Equation 14, is called Iteratively Reweighted Least Squares (IRLS) because, in each iteration, it solves a weighted least squares problem of the form:

$$\theta^{(t+1)} \leftarrow \arg\min_{\theta}(\mathbf{z}-\mathbf{X}\theta)'\mathbf{W}(\mathbf{z}-\mathbf{X}\theta)$$

(Hastie et al. 2009). Again, the estimation procedure stops once the convergence criteria are met.

2.4 How large samples affect the computation of $\widehat{\theta}_{MLE}$

Fitting a logistic regression model through a method like IRLS can have a high computational cost since it relies on an iterative procedure to find the algorithm's convergence (Hastie et al. 2009). This thesis will only focus on the computational burdens arising from very large N, i.e., $N >> k$; however, the problem can also come from high-dimensional data, namely when $k >> N$ (see Koh, Kim, and Boyd (2007) for a comprehensive review on regularized logistic regression). When dealing with a large dataset, solving the likelihood function optimization problem might become infeasible (L. Han et al. 2020). The problem arises when trying to calculate the update $\theta^{(t+1)}$, shown in Equation 11 as the Newton step. The term $\left(\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\right)$ is the observed information matrix, also known as the negative of the Hessian matrix, which has to be calculated and inverted in each iteration of the Newton–Raphson method (J. Yu, Ai, and Ye 2023).

According to Wang, Zhu, and Ma 2018, each iteration of Equation 11 has an order of Nd^2 time complexity or $O(Nd^2)$, where d refers to the dimensionality of the data (k in the notation of this thesis). This implies that the computation time for logistic regression, as shown in Section 2.3, increases linearly with an increase in N. Moreover, the whole optimization procedure to find $\widehat{\theta}_{MLE}$ would take $O(\zeta N d^2)$ time, where ζ is the number of iterations required until the optimization algorithm has converged. For extremely large sample problems, the computation time of $O(Nd^2)$ for just one run could already be too costly, not to mention to calculate it iteratively (Wang, Zhu, and Ma 2018, Wang 2019, Cheng, Wang, and Yang 2020).

3 Subsampling methods

To alleviate the computational burden arising from massive data sets, several methods have surfaced with the strategy of approximating $L(\theta)$ by a smaller subsample in which a logistic regression can be fitted more easily (J. Yu, Ai, and Ye 2023). In general, subsampling works by assigning an acceptance probability to each data point in the full sample and then selecting a smaller sample based on the assigned probabilities (L. Han et al. 2020). It is important to consider that estimates from the subsample may suffer from a loss in statistical accuracy; namely, the estimator's variance may become very large as the subsample size gets smaller. Thus, it is crucial to design an effective subsampling scheme that reduces the loss in statistical accuracy while assuring computational gains from the use of a smaller set of data (L. Han et al. 2020). However, as shown next, the challenge of finding an informative subsample becomes even larger when the data set at hand is imbalanced.

3.1 The imbalance problem

The imbalance problem comes from a significant or extremely unequal distribution between the classes C in the data set, usually with the class $Y = 1$ being under-represented (He and Garcia 2009). The degree of imbalance is usually measured by an imbalance ratio (IR), which for the purposes of this thesis, I define as $IR = n_0/N$, where n_0 denotes the cardinality of the majority class set $Y = 0$. If the $IR \geq 0.7$, the data set is considered to have a mild to severe imbalance. Furthermore, following Fithian and Hastie (2014) and O'Brien and Ishwaran (2019), two types of imbalance can be distinguished:

Definition 3.1 (Marginal imbalance)
The probability of the minority class is low or close to zero throughout the feature space, $\mathbb{P}\{Y = 1|X = x\} << \frac{1}{2}\ \forall\ x \in X$. Severe marginal imbalance is the case where $\mathbb{P}\{Y = 1|X = x\} \approx 0\ \forall\ x \in X$ (O'Brien and Ishwaran 2019).

Definition 3.2 (Conditional imbalance)
There exists a set $A \subset X$ with nonzero probability, $\mathbb{P}\{X \in A\}$ such that $\mathbb{P}\{Y = 1|X \in A\} \approx 1$ and $\mathbb{P}\{Y = 1|X \notin A\} \approx 0$ (O'Brien and Ishwaran 2019).

The first definition indicates a low probability of observing a case throughout the feature space. In contrast, the second one implies that, for certain feature values, the probability of observing a case is close to 1 (O'Brien and Ishwaran 2019). Marginal imbalance is common in data regarding fraud detection or rare disease diagnosis. In contrast, conditional imbalance is more likely to be observed in applications such as web spam filtering and image recognition (Fithian and Hastie 2014, L. Han et al. 2020).

A significant amount of research has been conducted in the machine learning field to find solutions for the imbalance problem: i) data-level approaches (i.e., random over and undersampling, synthetic sampling with data generation and cluster-based sampling); ii) cost-sensitive methods (i.e., cost-sensitive boosting methods, decision trees, and neural networks); and iii) kernel-based methods (i.e., integration with sampling methods and kernel modification). For a detailed description of each, see survey papers He and Garcia (2009) and Chawla, Japkowicz, and Kotcz (2014). For large-scale data, however, where the usual approach is to uniformly subsample the data for solving the computational problem, an imbalance in the data set generates an additional challenge that cannot be addressed with the traditional approaches mentioned before. Uniform subsampling will most likely fail in creating meaningful subsamples for the imbalanced data, as it assigns the same acceptance probability to every data point (L. Han et al. 2020, Wang 2020, Yao and Wang 2021, Fithian and Hastie 2014).

To show why this might be a problem, consider the example in Cheng, Wang, and Yang (2020) and let $z = (z_1, z_2, ..., z_N)$ and $\sum_{i=1}^{N} z_i = N_s$, with $i = 1, 2, \ldots, N$ denote a random subsample without replacement of size N_s taken from the full sample, where z_i takes the value of 1 if the ith data point has been included in the subsample. Furthermore, let π_i be the corresponding subsample probabilities such that $\sum_{i=1}^{N} \pi_i = 1$. Then, a subsampling-based logistic regression estimator has the general form:

$$\widehat{\theta}^z = \arg\max_{\beta} \sum_{i=1}^{N} \frac{z_i}{\pi_i} \left(y_i \log p\left(x_i\right) + \left(1 - y_i\right) \log\left(1 - p\left(x_i\right)\right)\right). \tag{15}$$

As mentioned, uniform subsampling assigns the same sampling probability to each data point, i.e., $\pi_i = \frac{1}{N}$. For imbalanced data sets, where the number of controls greatly exceeds the number of cases, having the same sampling probability for each data point means that the probability of sampling a case is very low, and the estimator in Equation 15 will be inefficient, as uniform subsampling fails to exploit the unequal importance of the data points (L. Han et al. 2020, Wang 2020, Yao and Wang 2021, Fithian and Hastie 2014). Therefore, traditional case-control subsampling methods have been proposed as a more accurate solution to the problem, as they create informative subsamples by adjusting the mixture of the classes.

3.2 Case-control subsampling

Case-control sampling is an important statistical tool mostly used in epidemiological studies, where the goal is to identify factors related to disease incidence and risk (Prentice and Pyke 1979). When the disease under study is rare, that is, the number of disease-free units exceeds the number of positive cases, case-control allows for the evaluation of such factors by over-sampling cases and under-sampling controls from the population (Borgan et al. 2018). Normally, the same number of cases and controls are sampled, creating a new subsample with no marginal imbalance (Fithian and Hastie 2014). Although originally developed in epidemiology, this method can be applied to other contexts, including economic research where large imbalanced data sets are common. The main idea is that adjusting the intercept of the logistic regression model fitted to the subsample can create a valid model for the original population (Anderson 1972, Prentice and Pyke 1979, A. Scott and Wild 1986, King and Zeng 2001, Fithian and Hastie 2014).

3.2.1 Intercept adjustment

Consider the derivation for binary models presented in King and Zeng (2001) (pg. 159-160). Suppose X, Y (binary) are random variables with full sample density $P(X, Y)$ and x,y are random variables with density $P(x, y)$, defined by a subsampling scheme, which samples all cases and a random selection of controls from X, Y. Let D and d be random samples of size N_s taken from $P(X, Y)$ and $P(x, y)$, respectively. Additionally, let $A_y = P(Y \mid D)/P(y \mid D)$ be a factor correction function and $B = P(x \mid d)/P(X \mid D) = \left[\sum_{\text{all } y} P(y \mid x, d)A_y\right]^{-1}$ a constant normalization factor (King and Zeng 2001).

Let $Pr(Y = 1) = \tau$ be the fraction of cases in the population and $Pr(y = 1) = \bar{y}$

the fraction of cases in the sample. The correction factor equations for the binary case are then $A_1 = \tau/\bar{y}$ and $A_0 = (1-\tau)/(1-\bar{y})$, with $B^{-1} = Pr(Y = 1 \mid x, d)\tau/\bar{y} + Pr(Y = 0 \mid x, d)(1-\tau)/(1-\bar{y})$ (King and Zeng 2001). Then we have

$$P(y = 1 \mid x, d)A_1B = \frac{P(y = 1 \mid x, d)\tau/\bar{y}}{P(y = 1 \mid x, d)(\tau/\bar{y}) + P(y = 0 \mid x, d)(1-\tau)/(1-\bar{y})} \tag{16}$$

$$= \left[1 + \left(\frac{1}{P(y = 1 \mid x, d)} - 1\right)\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]^{-1} \tag{17}$$

For the logit model, where $Pr(Y = 1 \mid x, d) = 1/1 + e^{-(\alpha + x_i'\beta)}$, Equation 17 becomes

$$P(y = 1 \mid x, d)A_1B = \left[1 + e^{-(\alpha + x_i'\beta) + \log\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]}\right]^{-1},$$

which is a corrected version of Equation 3. Thus, we can further derive it to get the odds and the log-odds functions for the corrected subsample distribution:

$$\frac{P(y = 1 \mid x, d)A_1B}{1 - P(y = 1 \mid x, d)A_1B} = e^{\alpha + x_i'\beta - \log\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]}$$

$$\log\left(\frac{P(y = 1 \mid x, d)A_1B}{1 - P(y = 1 \mid x, d)A_1B}\right) = \alpha - \log\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right] + x_i'\beta$$

$$= \alpha - b + x_i'\beta = \alpha^* + x_i'\beta. \tag{18}$$

Equation 18 shows that the intercept α needs to be corrected by the factor $b = \log\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$, while all the β slope coefficients remain unchanged (King and Zeng 2001, A. Scott and Wild 1986). The full derivation of this result can be found in the Appendix section 9.1.1.

The algorithmic approach shown in Fithian and Hastie (2014) provides a systematic framework for applying case-control subsampling. The algorithm generates an independent Bernoulli distribution and an acceptance probability function that together randomly samples data points from the population. Let $a(y)$ be the acceptance probability function

$$a(y) = \begin{cases} a(1), & \text{if } y = 1. \\ a(0), & \text{otherwise,} \end{cases} \tag{19}$$

and $b = \log\left(\frac{a(1)}{a(0)}\right)$ the correction factor for the intercept, refer by the authors as a log-selection bias. The pseudo-code in Algorithm 1 describes Fithian and Hastie's full case-control subsampling algorithm (CC).

The definition of b provided by the authors slightly differs from the correction shown by King and Zeng (2001), mainly because the methods in Fithian and Hastie (2014) imply a 50-50 split between the two classes. To ensure that the classes proportions are equal in the subsample, the authors assume that $\bar{y} = (1 - \bar{y})$, $(1 - \tau) = a(1)$ and $\tau = a(0)$, which then leads to $b = \log \frac{a(1)}{a(0)}$. It is important to note that there is no restriction that states that $a(y) \leq 1$, and, for a 50-50 split, the acceptance probabilities could always be defined differently, as we will see in Section 6. Additionally, the marginal probability of $Z = 1$ (i.e., the size of the subsample as a fraction of the full sample) can be estimated by:

$$\bar{a} = a(1)P(Y = 1) + a(0)P(Y = 0)$$

Algorithm 1 *CC subsampling*

```
1. Generate independent $z_i \sim \text{Bernoulli}\left(a\left(y_i\right)\right)$, where $z_i$ is generated by:
    1.1 Generate $u_i \sim U(0,1)$, which is independent of the data, the pilot, and each $i$
    1.2 Create $z_i = \mathbf{1}_{u_i \leq a(y_i)}$
    1.3 Generate the subsample $S = \{(x_i, y_i) : z_i = 1\}$
2. Fit a logistic regression to the subsample $S$ and obtain unadjusted estimates $\hat{\theta}_S = (\hat{\alpha}_S, \hat{\beta}_S)$
3. Get adjusted $\widehat{\theta}_{CC}$ estimates for the population by:
    3.1 $\hat{\alpha}_{CC} \leftarrow \hat{\alpha}_S - b$
    3.2 $\hat{\beta}_{CC} \leftarrow \hat{\beta}_S$
```

In general, the algorithm provides a framework to apply the CC method easily and correct the intercept for getting adjusted estimates for the population. Next, I will use the derivation in King and Zeng (2001) to show the consistency of the CC estimator and why, under certain assumptions, it allows us to make inferences about the population parameters.

3.2.2 Consistency of the CC estimator

The CC estimate is consistent and asymptotically efficient under the assumption that both the functional form and the regressors are correctly specified. To see this, recall that $P(X, Y)$ and $P(x, y)$ are the full sample and the subsample distributions, respectively. Under CC, the assumption is that the sampling scheme allows us to have $P(x, y) = P(X, Y)$. However, the marginal distributions are not necessarily the same between the population and the subsample, i.e., $P(x) \neq P(X)$, $P(y) \neq P(Y)$, and $P(y \mid x) \neq P(Y \mid X)$. The objective of CC is then to make inferences about the population $P(X, Y)$ through $P(x, y)$ (King and Zeng 2001). By Bayes theorem, we can write:

$$P(Y|X) = P(X \mid Y)\frac{P(Y)}{P(X)} = P(x \mid y)\frac{P(Y)}{P(X)} = P(y \mid x)\frac{P(Y)}{P(y)}\frac{P(x)}{P(X)} \tag{20}$$

Recall also that N_s is the size of the random samples taken from the population. So, as $N_s \to \infty$, it follows that

$$P(Y \mid X, D) = P(X \mid Y, D)\frac{P(Y \mid D)}{P(X \mid D)} \overset{d}{\to} P(X \mid Y)\frac{P(Y)}{P(X)} = P(Y \mid X), \tag{21}$$

as well as

$$P(x \mid y, d) \overset{d}{\to} P(x \mid y) = P(X \mid Y)$$

$$P(Y \mid D) \overset{d}{\to} P(Y) \quad \text{and} \quad P(X \mid D) \overset{d}{\to} P(X)$$

where $\overset{d}{\to}$ stands for convergence in distribution. However,

$$P(y \mid x, d) = P(x \mid y, d)\frac{P(y \mid d)}{P(x \mid d)} \overset{d}{\nrightarrow} P(Y \mid X), \tag{22}$$

meaning that the unadjusted subsample distribution does not converge to the full sample distribution. Nonetheless, by correcting the subsampling distribution once again using A_y and B functions and applying the Bayes theorem, we get:

$$P(y \mid x, d) A_y B = P(x \mid y, d) \frac{P(y \mid d)}{P(x \mid d)} A_y B = P(x \mid y, d) \frac{P(y \mid d)}{P(x \mid d)} \frac{P(Y \mid D)}{P(y \mid D)} \frac{P(x \mid d)}{P(X \mid d)}$$

$$= P(x \mid y, d) \frac{P(Y \mid D)}{P(X \mid D)} \xrightarrow{d} P(X \mid Y) \frac{P(Y)}{P(X)} = P(Y \mid X), \tag{23}$$

which indicates that the corrected subsample distribution does converge and is consistent with the full sample distribution. (King and Zeng 2001). Despite its success in correcting the bias in the intercept, the CC estimator still has some drawbacks, as Fithian and Hastie (2014) point out, specifically when the model is not correctly specified. Under misspecification, the CC estimates will be biased and inconsistent since the CC algorithm will yield different parameters for every choice of b. The authors argue that, in the limit, CC will yield a different estimate if we sample an equal number of cases and controls or if we sample twice as many cases as controls, and so on. This inconsistency is harmful to inferences about θ, as the subsample parameters will differ greatly from the population ones. Additionally, Fithian and Hastie (2014) claim that CC performs well in cases where there is only marginal imbalance present in the dataset and fails to exploit the conditional imbalance problem.

3.3 Weighted case-control subsampling

An alternative sampling procedure to CC is the weighted case-control subsampling algorithm (WCC), based on the weighted exogenous sampling maximum-likelihood estimator first introduced by Manski and Lerman (1977). The authors proved that the WCC estimator is consistent and asymptotically normal if the population probabilities τ and $(1-\tau)$ are available to the researcher (Maalouf 2011). The main idea is to weight each subsample data point by the inverse of their probability of being sampled, $a(y)^{-1}$ (Fithian and Hastie 2014). The intuition behind weighting is that if the sampling probability for some event is large, then $a(y)^{-1}$ will be small, and the data points with larger sampling probabilities are given less weight in the parameter's estimation (Maalouf 2011). Therefore, instead of maximizing 8, the method maximizes:

$$\ell(\theta) = \left(\frac{1}{a(1)} \right) \sum_{i=1}^{N} y_i \log p\left(x_i\right) - \left(\frac{1}{a(0)} \right) \sum_{i=1}^{N} \left(1 - y_i\right) \log\left(1 - p\left(x_i\right)\right). \tag{24}$$

Algorithm 2 *WCC subsampling*

```
1. Generate independent $z_i \sim \text{Bernoulli}\left(a\left(y_i\right)\right)$, where $z_i$ is generated by:
    1.1 Generate $u_i \sim U(0,1)$, which is independent of the data, the pilot, and each $i$
    1.2 Create $z_i = \mathbf{1}_{u_i \leq a(y_i)}$
    1.3 Generate the subsample $S = \{(x_i, y_i) : z_i = 1\}$
2. Fit a weighted logistic regression to the subsample $S$ as specified in Equation 24 and
   obtain unadjusted estimates $\hat{\theta}_S = (\hat{\alpha}_S, \hat{\beta}_S)$
3. Get unadjusted $\widehat{\theta}_{WCC}$ estimates for the population by:
    3.1 $\hat{\alpha}_{WCC} \leftarrow \hat{\alpha}_S$
    3.2 $\hat{\beta}_{WCC} \leftarrow \hat{\beta}_S$
```

Since $a(y_i) > 0$ for every $i \in N$, the WCC estimator is considered a Horvitz-Thompson estimator, and thus, it is an unbiased, $\sqrt{N}$-consistent and asymptotically normal estimator for the population parameter θ under unequal selection probabilities. (Manski and Lerman 1977, A. Scott and Wild 2002, Fithian and Hastie 2014, Overton and Stehman 1995). In A. Scott and Wild (2002), it is mentioned that the appeal of the WCC estimator is its robustness in the case of misspecification of the model. When the linear model in 24 does not hold, the WCC estimator can still yield a consistent solution for θ (A. Scott and Wild 2002).

To show this, consider the formulation in A. Scott and Wild (1986) and A. Scott and Wild (2002). If a random sample of size N_s is taken from the population and we assume the logistic form is valid, then the maximum likelihood estimator $\hat{\theta}_{MLE}$ would satisfy:

$$\sum_{i:y=1} x_i p_0(x_i, \hat{\theta}_{MLE}) = \sum_{i:y=0} x_i p_1(x_i, \hat{\theta}_{MLE}) \tag{25}$$

As $N_s \to \infty$, then $\hat{\theta}_{MLE}$ converges to β:

$$\tau E_1\{\mathbf{X} p_0(\mathbf{X}, \beta)\} = (1-\tau) E_0\{\mathbf{X} p_1(\mathbf{X}, \beta)\} \tag{26}$$

where E_1 and E_0 denote the expectation of the conditional distribution of the $(N \times k)$ regressors matrix $\mathbf{X}$ given $Y = 1$ and $Y = 0$, respectively (A. Scott and Wild 1986). The same approach can be written for the weighted estimator using the inverse of the selection probabilities, and thus $\hat{\theta}_{WCC}$ is the solution of:

$$\sum_{j=1}^{n_1} \frac{x_{1j} p_0(x_{1j}, \hat{\theta}_{WCC})}{a(1)} = \sum_{j=1}^{n_0} \frac{x_{0j} p_1(x_{0j}, \hat{\theta}_{WCC})}{a(0)}. \tag{27}$$

where n_1 is the number of cases, and n_0 is the number of controls in the subsample. According to A. Scott and Wild (1986), using the weak law of large numbers, one can show that $\hat{\theta}_{WCC}$ also converges to θ in probability as $n_1, n_0 \to \infty$. The main message of this result is that even when the linear logistic model is not valid, the $\hat{\theta}_{WCC}$ estimator can still be interpreted as the best-fitting logistic model solution for the full sample. In other words, no matter the form of the true model, $\hat{\theta}_{WCC}$ will converge in probability to θ (A. Scott and Wild 1986).

Despite the method's clear robustness when the model is misspecified, its main drawback is that the variation in the selection probabilities between cases and controls generates a large variance in the parameters' estimators (A. Scott and Wild 2002, Y. Li, Graubard, and DiGaetano 2011). Moreover, the larger the imbalance in the data, the greater the efficiency loss since high disproportions in the data will induce large weights in the sample design (Elliott and Little 2000). Furthermore, the increase in the variance can overpower the potential bias reduction, in general incrementing the Mean Squared Error (MSE) and harming the estimator's prediction performance (Elliott and Little 2000). For a humorous example, one can refer to Basu's elephant tale (DasGupta 2011 pg. 176-177), where a circus statistician faces the dramatic consequences of working with severely noisy sampling probabilities.

3.4 Local case-control subsampling

Local-case control subsampling (LCC) was proposed by Fithian and Hastie (2014) as a generalization of the standard case-control subsampling. As opposed to the methods discussed above, LCC allows the acceptance probability function to depend not only on y but also on the covariates x. To do so, it uses a pilot estimator of $P(Y = 1 \mid X = x)$, which is fitted with an independent dataset from the original sample. Denote the pilot estimate as:

$$\tilde{p}(x_i) = \frac{e^{\tilde{\alpha} + x_i' \tilde{\beta}}}{1 + e^{\tilde{\alpha} + x_i' \tilde{\beta}}} \tag{28}$$

where $(\tilde{\alpha}, \tilde{\beta}) = \tilde{\theta}$. In some cases, there could already be an available pilot fit; for example, when a model is fitted every day with incoming data, then yesterday's fit is a good pilot for today's model (Fithian and Hastie 2014). However, when no pilot is available, the authors propose a first pass of the WCC algorithm using a fraction of the full sample to estimate $\tilde{p}(x_i)$.

Since the WCC estimate has been proved to be $\sqrt{N}$ consistent and asymptotically unbiased, the LCC estimate in the second stage will subsequently be consistent and unbiased, as explained later in this chapter. The authors recommend using a pilot data sample of the same size as the sample used to fit the LCC algorithm. However, it is important to note that this may not always provide a sufficient sample size for the pilot model, especially for small N. This will be further discussed in Section 5, where finite sample guarantees will also be explored.

The acceptance probability function $a(x_i, y_i)$ for the LCC is then defined as follows:

$$a(x_i, y_i) = |y - \tilde{p}(x_i)| = \begin{cases} 1 - \tilde{p}(x_i), & \text{if } y = 1. \\ \tilde{p}(x_i), & \text{otherwise.} \end{cases} \tag{29}$$

As before, the adjustment of the estimates can be justified using the derivation of Section 3.2; however, this time, both the intercept and the slope coefficients need to be adjusted. For LCC, the correction factor b is a function of the data, such that $b(x) = \log\left(\frac{a(x_i,1)}{a(x_i,0)}\right)$. Then, from Equation 18 we have:

$$\begin{aligned}
\frac{P(y = 1 \mid x, d)A_1 B}{1 - P(y = 1 \mid x, d)A_1 B} &= \alpha - b(x) + x_i'\beta \\
&= \alpha - \log\left(\frac{a(x_i, 1)}{a(x_i, 0)}\right) + x_i'\beta \\
&= \alpha - \log\left(\frac{1 - \tilde{p}(x_i)}{\tilde{p}(x_i)}\right) + x_i'\beta \\
&= \alpha - \log\left(\frac{1}{e^{\tilde{\alpha} + x_i'\tilde{\beta}}}\right) + x_i'\beta \\
&= \alpha - (-\tilde{\alpha} - x_i'\tilde{\beta}) + x_i'\beta \\
&= (\alpha + \tilde{\alpha}) + x_i'(\beta + \tilde{\beta}) = \alpha^* + x_i'\beta^* \qquad (30)
\end{aligned}$$

Algorithm 3 *LCC subsampling*

```
1. Generate independent $z_i \sim \text{Bernoulli}\left(a\left(x_i, y_i\right)\right)$, where $z_i$ is generated by:
    1.1 Generate $u_i \sim U(0,1)$, which is independent of the data, the pilot, and each $i$
    1.2 Create $z_i = \mathbf{1}_{u_i \leq a(y_i)}$
    1.3 Generate the subsample $S = \{(x_i, y_i) : z_i = 1\}$
2. Fit a logistic regression to the subsample $S$ and obtain unadjusted estimates $\hat{\theta}_S = (\hat{\alpha}_S, \hat{\beta}_S)$
3. Get adjusted estimates for the population by:
    3.1 $\hat{\alpha} \leftarrow \hat{\alpha}_S + \tilde{\alpha}$
    3.2 $\hat{\beta} \leftarrow \hat{\beta}_S + \tilde{\beta}$
```

The intuition behind LCC is that it measures the degree of "surprisingness" of each data point through the difference between the observed y_i and the predicted $\tilde{p}(x_i)$, under the assumption that the pilot estimate accurately approximates $\tilde{p}(x_i)$. The larger the difference, the more likely the data point will be selected for inclusion into S. Thus, each subsampled y_i is more informative than a traditional case-control subsampled data point, and valid estimates for the full sample can be obtained by adjusting log-odds using Equation 30.

Furthermore, Fithian and Hastie (2014) argue that one of the main advantages of LCC is that it can fully exploit conditional imbalance, whereas CC and WCC cannot. Thus, we could expect that in scenarios with high levels of conditional imbalance, LCC outperforms the methods in terms of both bias and efficiency. The performance of LCC under marginal imbalance, however, is not discussed by the authors. From the theory, it is still unknown whether LCC will outperform the other methods in cases where there is a high marginal imbalance but a low conditional imbalance. This theoretical gap will later be explored in the numerical exercises of Section 5.

3.4.1 Preliminaries for asymptotic results

For this part of the thesis, I will follow the notation used by Fithian and Hastie (2014) in the asymptotic section of their paper. I will not present formal proofs but rather the intuition and the underlying assumptions behind the Lemmas, Propositions, and Theorems. For simplicity, let λ now denote the pilot estimates $\tilde{\theta}$. Furthermore, the constant term is absorbed to x, so that $f_\lambda(x_i) = \lambda' x_i$. The LCC acceptance probability is now a function of λ, such that $a_\lambda(x_i, y_i) = \left| y_i - \frac{e^{x_i'\lambda}}{1+e^{x_i'\lambda}} \right| \in (0,1)$ and its expected value is $\bar{a}(\lambda) = \mathbb{E}a_\lambda(X,Y) \in (0,1)$. The acceptance probability of x_i into the subsample S is:

$$\hat{a}_\lambda(x_i) = \tilde{p}(x_i)(1 - p(x_i)) + (1 - \tilde{p}(x_i))p(x_i) \in (0,1), \tag{31}$$

where $\hat{a}_\lambda(x_i)$ is gonna be small if $\tilde{p}(x_i)$ is well approximated but large for data points where the true $p(x_i)$ differs greatly from $\tilde{p}(x_i)$. Additionally, a key assumption throughout this section is the non-separability of classes, as stated in Lemma 1:

Lemma 1 as stated in Fithian and Hastie (2014): Assume there is no $v \in \mathbb{R}^p$ for which $P(Y = 0, v\prime X > 0) = P(Y = 1, v\prime X < 0) = 0$. Hence, it follows that there is non-separability of classes.

This assumption means that there exists some degree of class overlap that does not allow the classes to be perfectly separable by a hyperplane in the feature space. Therefore, some misclassification can always be expected, and the objective is to minimize it. Let then $Q_\lambda(\theta)$ be the population risk function, such that $\hat{\theta}_{LCC} \approx \arg\min_\theta Q_\lambda(\theta)$. Assume further that the model is correctly specified and that the best linear predictor for the full sample is $\theta^* = \bar{\theta}(0)$, which is the large-sample limit of the LCC estimator with pilot fixed at $\lambda = 0$. Then, the authors present the following Proposition:

Proposition 2 as in Fithian and Hastie (2014): Assume $\mathbb{E}\|X\| < \infty$, the classes are non-separable and that $\theta^* = \bar{\theta}(0)$. Then, $\theta^* = \arg\min_\theta Q_{\theta^*}(\theta) = \bar{\theta}\left(\theta^*\right)$.

The main takeaway of Proposition 2 is that it suggests that if the pilot comes near to θ^*, then the LCC estimator will also converge eventually to θ^*. Furthermore, as a WCC estimator, the pilot will be consistent even under model misspecification, and so will the LCC estimate. Note that for this to hold, it must be that both $N \to \infty$ and the pilot $N_s \to \infty$, which could be difficult to hold, especially in small samples.

3.4.2 Consistency of $\hat{\theta}_{LCC}$

For the consistency of the LCC estimator, assume that there are infinite data point pairs (x_i, y_i), a sequence of i.i.d uniform variables $\{u_1, u_2, ..., u_n\}$ and a sequence of pilot estimates $\{\lambda_1, \lambda_2, ..., \lambda_n\}$. In this section, the assumption of independence of the pilot and the data is not strict, and it actually allows a dependency between the two. However, the sequence of uniform variables for performing the accept-reject decisions are assumed to be independent of the data, the pilot, and themselves.

The consistency of $\hat{\theta}_{LCC}$ is built on the asymptotic results of the pilot, specifically on the idea that if the pilot converges in probability to the optimal population parameter, then $\widehat{Q}_{\lambda_n} \approx Q_{\theta^*}$. $\widehat{Q}_{\lambda_n}$ denotes the empirical risk function minimized by LCC, whereas Q_{θ^*} is the analogous function for the true population, minimized by θ^*. In Proposition 3 of the paper, the authors define and prove pointwise convergence of the sequence of pilot estimates to their asymptotic limit $\widehat{Q}_{\lambda_n}(\theta) \to Q_{\lambda_\infty}(\theta)$, which in turns also implies that uniform convergence on compacts also holds (see Proposition 4 in Fithian and Hastie (2014)). With these two results, it follows:

Theorem 5 as in Fithian and Hastie (2014) : Assume $\mathbb{E}\|X\| < \infty$ and that the classes are non-separable. Then, if the pilot estimate is consistent such that $\lambda_n \xrightarrow{p} \theta^*$, then the local case-control estimate will be consistent $\hat{\theta}_n \xrightarrow{p} \theta^*$ as well.

3.4.3 Asymptotic distribution

In contrast to the previous section, the authors assume strict independence between the pilot and data to present the results for the theoretical asymptotic distribution of the LCC estimate. To derive the asymptotic variance of the estimate, denote the gradient of $Q_\lambda(\theta)$ as:

$$G(\theta, \lambda) \triangleq -\bar{a}(\lambda)\nabla_\theta Q_\lambda(\theta), \tag{32}$$

with variance-covariance matrix:

$$J(\theta, \lambda) \triangleq \mathrm{Var}_\lambda \left[\left(Y - \frac{e^{X'(\theta-\lambda)}}{1 + e^{X'(\theta-\lambda)}} \right) X \right]. \tag{33}$$

One can retrieve the Hessian from Equation 32 by differentiating again with respect to θ, obtaining:

$$H(\theta, \lambda) \triangleq -\bar{a}(\lambda)\nabla_\theta^2 Q_\lambda(\theta). \tag{34}$$

According to Fithian and Hastie (2014), following Maximum Likelihood estimation theory, by fixing the pilot estimates λ and approximating the sample size to the limit, $N \to \infty$, the coefficients of a logistic regression fitted on a sample sized $n\bar{a}(\lambda)$ from the original population would be asymptotically normal, with covariance matrix:

$$\frac{1}{n\bar{a}(\lambda)} H(\bar{\theta}(\lambda), \lambda)^{-1} J(\bar{\theta}(\lambda), \lambda) H(\bar{\theta}(\lambda), \lambda)^{-1}. \tag{35}$$

From these results, it follows:

Theorem 6 as in Fithian and Hastie (2014): Assume $\mathbb{E}\|X\|^2 < \infty$. If the pilot is independent of the data and its estimates are consistent, such that $\lambda_n \xrightarrow{p} \theta^*$, then

$$\sqrt{n}\left(\hat{\theta}_n - \bar{\theta}\left(\lambda_n\right)\right) \xrightarrow{d} N\left(0, \bar{a}\left(\theta^*\right)^{-1} \Sigma\right) \tag{36}$$

with

$$\Sigma = H\left(\theta^*, \theta^*\right)^{-1} J\left(\theta^*, \theta^*\right) H\left(\theta^*, \theta^*\right)^{-1}. \tag{37}$$

The proof of Theorem 6 can be found in the Appendix section of Fithian and Hastie (2014). Furthermore, from the theorem's derivations, the authors obtain some reassuring facts that emphasize the close asymptotic relation between the pilot and the LCC estimates. These are presented in Corollary 7 of the original paper and briefly summarized below:

1. If λ_n is $\sqrt{N}$-consistent as a sequence of Horvitz-Thompson estimators, then so is $\hat{\theta}_n$.

2. If λ_n is asymptotically unbiased, then so is $\hat{\theta}_n$.

3. If $\sqrt{n}\left(\lambda_n - \theta^*\right) \xrightarrow{d} N(0, V)$ then $\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{d} N(0, \Sigma)$ with,

$$\Sigma = H^{-1}\left(CVC' + \bar{a}^{-1}J\right)H^{-1} \tag{38}$$

   where C is a matrix of cross-sectional partial derivatives (for simplicity, the whole expression suppresses the argument of θ^* as in the original paper).

Equation 38 highlights the fact that the variance of the LCC estimator incorporates the variance of the pilot, denoted as V. Thus, we can expect the variance of $\hat{\theta}_{LCC}$ to increase or decrease as the pilot becomes less or more efficient, respectively. Furthermore, under the assumption that the logistic model is correctly specified, the following important theorem holds:

Theorem 8 as in Fithian and Hastie (2014): Assume that the logistic model is correctly specified and let $\frac{1}{n}\Sigma_{full}$ be the asymptotic variance of the θ_{MLE} for the full population. If $\mathbb{E}\|X\|^2 < \infty$, the pilot is independent of the data, and its estimates are consistent such that $\lambda_n \xrightarrow{p} \theta^*$, then:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{D}} N\left(0, a\left(\theta_0\right)^{-1} \Sigma\right) = N\left(0, 2\Sigma_{\text{full}}\right) \tag{39}$$

Assuming that the model is accurately specified, Theorem 8 offers a robust and important theoretical result for the LCC estimate. The intuition behind it is that even if the LCC takes only $n\bar{a}(\lambda)$ of the full sample data points, its variance is only twice as large as the variance of the logistic regression estimates on the full sample. According to the authors, this means that its variance is the same as the variance of a logistic regression estimate fitted on a random uniform subsample of size $\frac{N}{2}$ from the full population. That is, each point sampled by the LCC method is worth $\frac{1}{2\bar{a}(\lambda)}$ data points taken by uniform sampling.

This implies that the smaller the expected value of the true sampling probability $a(\theta_0)$ is, i.e., $\bar{a}\left(\theta_0\right) = \mathbb{E}(|Y - \tilde{p}(X)|)$, the most advantageous the LCC method will be in terms of reducing computational costs. In other words, the more the outcome variable Y is easy to predict throughout the feature space, the smaller the LCC subsample will be, reducing the burden in computational time of estimating θ_{MLE}, yet still only having twice its variance. Again, the assumptions of correct specification of the model and large N must hold.

As we will see in the numerical exercises of Section 5, there are some drawbacks to the LCC method, especially related to its effective subsample size and the consistency of the pilot. In corner cases, even when there is high conditional imbalance, a severe marginal imbalance, i.e., $P(Y = 1) > 0.9$, will reduce the effective sample size of LCC drastically. Consequently, the pilot estimate will also suffer from a reduced sample size, failing to reach consistency, and thus, the LCC estimate itself could become extremely inefficient.

4 Evaluation metrics

4.1 The bias-variance decomposition

This section introduces the metrics for evaluating the performance of the subsampling methods in Section 5. Specifically, we would want to see the behavior of the bias and variance of the estimators as the underlying assumptions and structure of the data change. For this, recall Equation 5, which defines the true model. Now, we add an error term, such that $f_\theta(x_i) = \alpha + x_i^\intercal \beta + \varepsilon$, with $\mathbb{E}(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma_\varepsilon^2$. Consider the expected prediction error derivation as in Hastie et al. (2009):

$$\begin{aligned} \operatorname{Err}(x_i) &= \mathbb{E}\left[\left(f_\theta(x_i) - \hat{f}_\theta(x_i)\right)^2 \mid X = x_i\right] \\ &= \sigma_\varepsilon^2 + \left[\mathbb{E}\hat{f}(x_i) - f(x_i)\right]^2 + \mathbb{E}\left[\hat{f}(x_i) - \mathbb{E}\hat{f}(x_i)\right]^2 \\ &= \sigma_\varepsilon^2 + \operatorname{Bias}^2\left(\hat{f}(x_i)\right) + \operatorname{Var}\left(\hat{f}(x_i)\right) \\ &= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance} \end{aligned} \tag{40}$$

When considering the accuracy of a model, it's important to evaluate the last two factors in Equation 40. The first factor is what's known as the irreducible error, which is the variance of $f_\theta(x_i)$ around its true mean and is inevitable. The second factor is the squared bias, which measures how much the estimate's mean differs from the true mean. Finally, the third factor is the model's variance, which is the expected squared deviation of the estimated model. Following Fithian and Hastie (2014), I evaluate the CC, WCC and LCC estimators by calculating these two terms empirically as $\widehat{\text{Bias}}^2 = \|\mathbb{E}\hat{\theta} - \theta\|^2$ and $\widehat{\text{Var}} = \operatorname{Var}(\hat{\alpha}) + \sum_{j=1}^{k} \operatorname{Var}(\hat{\beta}_j)$, where $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ denotes the Monte Carlo realizations of the estimates.

4.2 Obtaining the true intercept value

To calculate the squared bias and variance of the model, it is crucial to know the true parameters' value. The value of β is easy to know; in the simulation study, it will be equal to the value of μ from the Gaussian distribution it comes from. As for the intercept, it can be retrieved by fixing the log-odds function at a point $x = x_0$ and solving for α:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \alpha + x_0\beta; \quad \text{where} \quad \alpha = \log(p(x)) - \log(1 - p(x)) - x_0\beta.$$

Throughout the simulations, I fix $x_0 = 0.5$ as in Wang (2020).

5 Simulation study

In this section, I compare the performance of CC, WCC, and LCC methods as in Fithian and Hastie (2014). Furthermore, I conduct numerical exercises to show the main statistical properties of the LCC method, particularly its asymptotic variance behavior as a scalar of Σ_{full}. Additionally, I expand on the author's analysis by varying the data generation process (DGP) in the last two simulation exercises to examine the methods' performance on different population sample sizes N and various levels of marginal imbalance.

To the best of my knowledge, there are currently no available R packages to implement the algorithms as presented in Section 2. Thus, I developed and wrote the functions for their application using R Statistical Software, version 4.2.1 (R Core Team 2021). Additional packages used for the core analysis of the results are listed in Appendix Section 9.3. All functions, simulation exercises, analyses, plots, and figures shown here can be accessed for review and reuse at https://github.com/carolinalvarez/code-MA-thesis.

5.1 General setup

Following mainly Fithian and Hastie (2014) but also Wang (2020) and L. Han et al. (2020), the regressors in the general DGP follow a different Gaussian distribution depending on the value of Y. That is, $X \mid Y = y \sim N(\mu_y, \Sigma_y)$. If $Y = 1$, then $X \sim N(\mu_1, \Sigma_1)$, and if $Y = 0$, then $X \sim N(\mu_0, \Sigma_0)$. The imbalanced ratio or degree of marginal imbalance is set by $P(Y = 0) = r$, implying $P(Y = 1) = 1 - r$. Furthermore, assume the data also has conditional imbalance, generated by creating a specific set in μ_1 where $P(Y = 1|X = \mu_1) \approx 1$. Thus, let μ_0 and μ_1 be defined as:

$$\mu_0 = [\underbrace{0, 0, \dots, 0}_{k}]', \mu_1 = [\underbrace{0, 0, \dots, 0}_{\frac{k}{2}}, \underbrace{1, 1, \dots, 1}_{\frac{k}{2}}]',$$

where it is expected that the conditional imbalance increases as k gets larger. For all simulations, the model is correctly specified. Thus, the variance-covariance matrices are set to be identical between the classes, that is, $\Sigma_0 = \Sigma_1$.

$$\Sigma_1 = \Sigma_0 = Cov(X) = \begin{bmatrix} Var(x_1) & \cdots & Cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ Cov(x_n, x_1) & \cdots & Var(x_n) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

Any population size can be drawn from the above setup by setting the desired values of $N, k,$ and $P(Y = 1)$. Additionally, Lemma 1 holds in this setup, indicating a non-separability of classes within the general DGP. To demonstrate this, let's consider a data realization from Simulation 5.2. Figure 2 shows a Principal Components Analysis (PCA) plot of the generated population, where a separation between classes is observed, primarily due to the conditional imbalance present in the data. However, the classes cannot be perfectly separated by a hyperplane, as there is some overlap between the two data clusters.

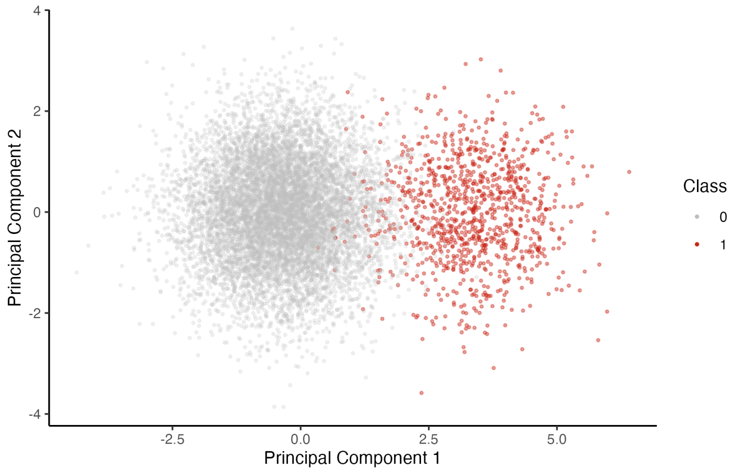

Figure 2: PCA plot of example population in Simulation 5.2 showing non-separability of classes.

5.2 Simulation 1

The first simulation closely follows the setup in simulation 2 of Fithian and Hastie (2014). The main idea is to "replicate" as closely as possible the findings in the original paper and check whether the R functions for applying the algorithms work as intended. Since the linear model is correctly specified, all methods are asymptotically unbiased; therefore, the exercise is meant to show the superiority of the LCC method in terms of the variance of the model.

Let the size of the population be $N = 10^5$, the marginal imbalance in the population

$P(Y = 1) = 0.1$, the number of regressors $k = 30$, $\mu_1 = (1_{15}, 0_{15}), \mu_0 = 0_{30}$, and $\Sigma_0 = \Sigma_1 = I_{30}$ as in Section 5.1. To fit the methods, I set the acceptance probability to $a(1) = 0.9$ for CC, WCC, and the pilot, so there is a 50-50 split between the classes in the subsample. Since the pilot is taken from a subsample of the full data, it is a data-dependent pilot, as an observation i can be observed by the pilot and LCC model. However, we can expect the LCC estimate still be consistent, as the independence of the pilot is not required for consistency of the estimate.

For comparison between the algorithms, the authors fix the subsample sizes, letting CC and WCC have exactly twice as many data points as LCC alone (in this sense, LCC must "pay" for its pilot, where the latter gets the same amount of data points as LCC). Since the paper does not explicitly show how the authors fix N_s, I first run 100 simulations, fit the methods, store the subsample size N_s for LCC, and approximate the LCC's fixed N_s by roughly taking the median from the N_s realizations. In order to closely follow the authors, I round the LCC N_s to its nearest thousand, and then I fix the CC and WCC subsample sizes as twice the N_S size for LCC. The tables with the median LCC's subsample size used as a reference can be found in Appendix Section 9.2.1 for all simulations.

Table 1 and Figure 3 illustrate the simulation study's findings, which support the original paper's results. The LCC algorithm exhibits the lowest bias and variance among all three algorithms. This result is not surprising given the large population sample and high levels of marginal and conditional imbalance in this setup. According to the authors, LCC outperforms CC in terms of variance because traditional case-control methods have no way to exploit conditional imbalance. On the other hand, WCC exhibits the largest bias and variance. As for the bias, it could be that the fixed size of N_s may be too small to allow the method to reach its asymptotic bias behavior fully. Additionally, recall that WCC's main drawback is its high variance as a result of variation in the weights; in this setup, as the imbalance is quite high, a large efficiency loss is observed.

	N_s	$\widehat{bias^2}$	$\widehat{var}$
CC	4000	0.25	0.68
WCC	4000	1.02	1.27
LCC	2000	0.006	0.18

Table 1: Simulation 1 - Baseline numerical exercise (results for 1,000 runs).

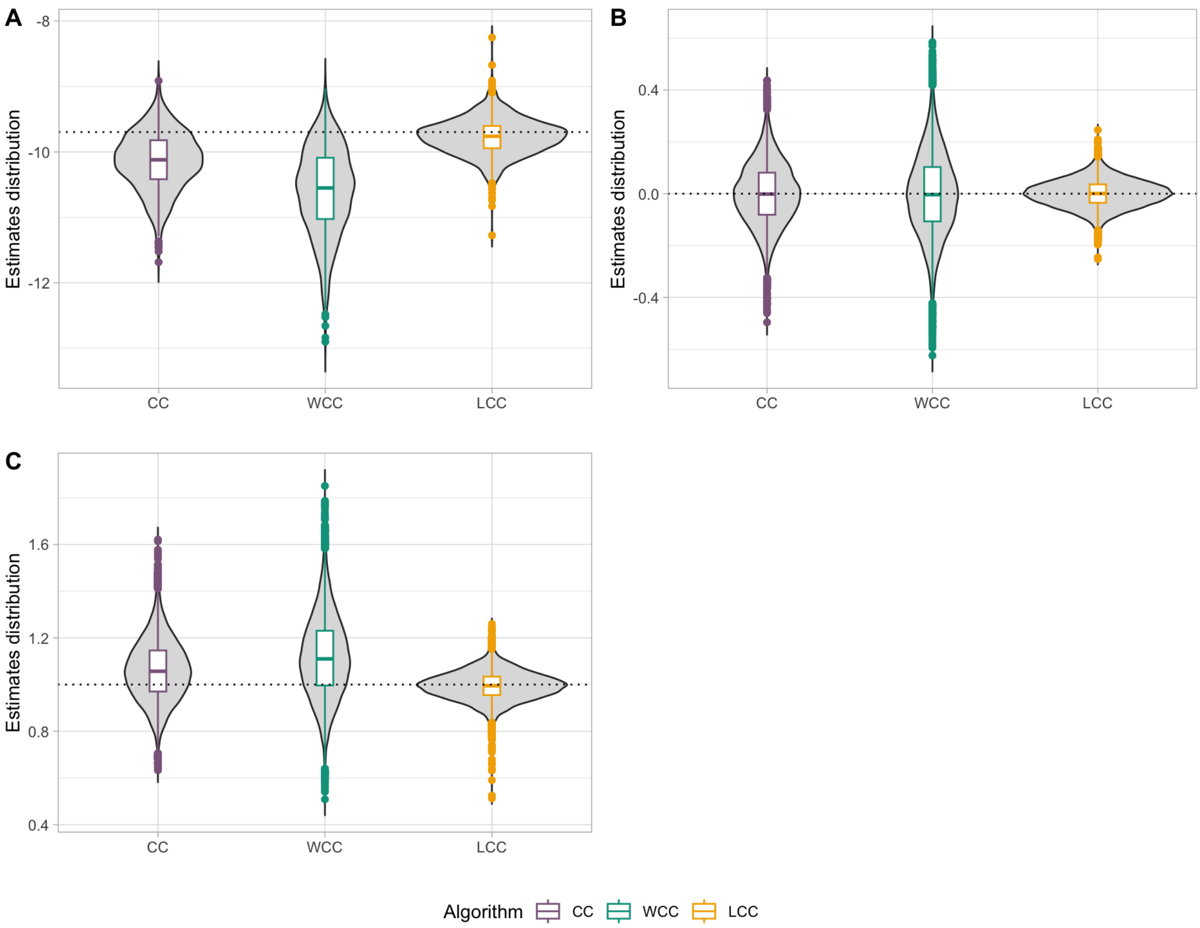

Figure 3: Simulation 1 - Distribution of $\hat{\theta}$ estimates across subsampling algorithms, where the dotted line corresponds to the true parameters' values. **Panel A:** Distribution of intercept coefficients with true $\alpha = -9.7$. **Panel B:** Distribution of slope coefficients with true $\beta = 0$, namely $\beta_1, \ldots, \beta_{\frac{k}{2}} = 0$. **Panel C:** Distribution of slope coefficients with true $\beta = 1$, namely $\beta_{\frac{k}{2}+1}, \ldots, \beta_k = 1$.

5.3 Simulation 2

This second numerical exercise focuses on the asymptotic behavior of the LCC estimate. The focal point is Theorem 8 in Section 3.4.3, which states that if the model is correctly specified and the pilot model is consistent and independent of the data, the asymptotic variance of the local case-control estimate will be precisely twice the asymptotic variance of the logistic regression estimate on the full sample. As stated before, the simulation study uses a data-dependent pilot since, according to the authors, an independent one would require, for example, fitting the model in data from an earlier period.

Consider the simulation setup in 5.1, this time with $k = 30$ and full population sizes of $N = \{10^5, 10^6\}$ that aim to approximate the asymptotic limit behavior. For this part of the analysis, there are no fixed subsample sizes N_s neither for the pilot nor the

LCC algorithm. Rather, the pilot estimates $\tilde{\theta}$ are taken from an initial pass on the whole data set, and the LCC subsample should be roughly $n\bar{a}(\tilde{\theta})$. The results are shown in Table 2. Both methods are asymptotically unbiased, and so it is observed in the table results, where the squared bias of both logistic regression and LCC are approaching zero.

As for the variance's relation between LCC and the logit model, the results show that for $N = 10^5$, the LCC's variance is approximately 2.85 times the variance of the logistic regression. Letting the full population size increase to $N = 10^6$ reduces this factor to 2.66. There are two possible reasons why the LCC's variance might not be exactly twice as of the logit for this exercise. First, since it is an asymptotic result, much larger population size might be needed to approximate the asymptotic behavior of the LCC estimate. And secondly, the theoretical results assume a data-independent pilot, which in this case, does not hold. A data-dependent pilot could be inflating the variance of the LCC in finite samples.

	$N = 10^5$				$N = 10^6$			
	$\bar{N}_s$	$\widehat{bias}^2$	$\widehat{var}$	$\bar{a}(\tilde{\theta})$	$\bar{N}_s$	$\widehat{bias}^2$	$\widehat{var}$	$\bar{a}(\tilde{\theta})$
Logit	10^5	0.0012	0.0279	-	10^6	8.47×10^{-6}	0.0027	-
LCC	2140	0.0003	0.0796	0.0214	21394	3.71×10^{-6}	0.0072	0.0214

Table 2: Simulation 2 - Approximation of asymptotic behavior of LCC against logit (results for 1,000 runs).

Although the simulation findings do not match the theoretical predictions exactly, they are still very close. Figure 4 shows the different distributions of the Monte Carlo simulations, distinguishing between $\hat{\alpha}$'s distribution, one $\hat{\beta}$ distribution for $\beta_1, \ldots, \beta_{\frac{k}{2}} = 0$ and one $\hat{\beta}$ distribution for $\beta_{\frac{k}{2}+1}, \ldots, \beta_k = 1$. The grey distribution represents the theoretical distribution for each set of estimates. It is shown that despite not fitting it perfectly, the LCC distribution closely follows the theoretical one for all sets of estimates. It is worth noting that $\bar{a}(\tilde{\theta}) = 0.0214$, which means that despite only using approximately 2% of the complete sample, LCC's variance is still no larger than 3 times the logit variance for large samples.

5.4 Simulation 3

The third numerical exercise aims to analyze how the methods perform when faced with varying levels of marginal imbalance. It is worth noting that changing $P(Y = 1)$ would not

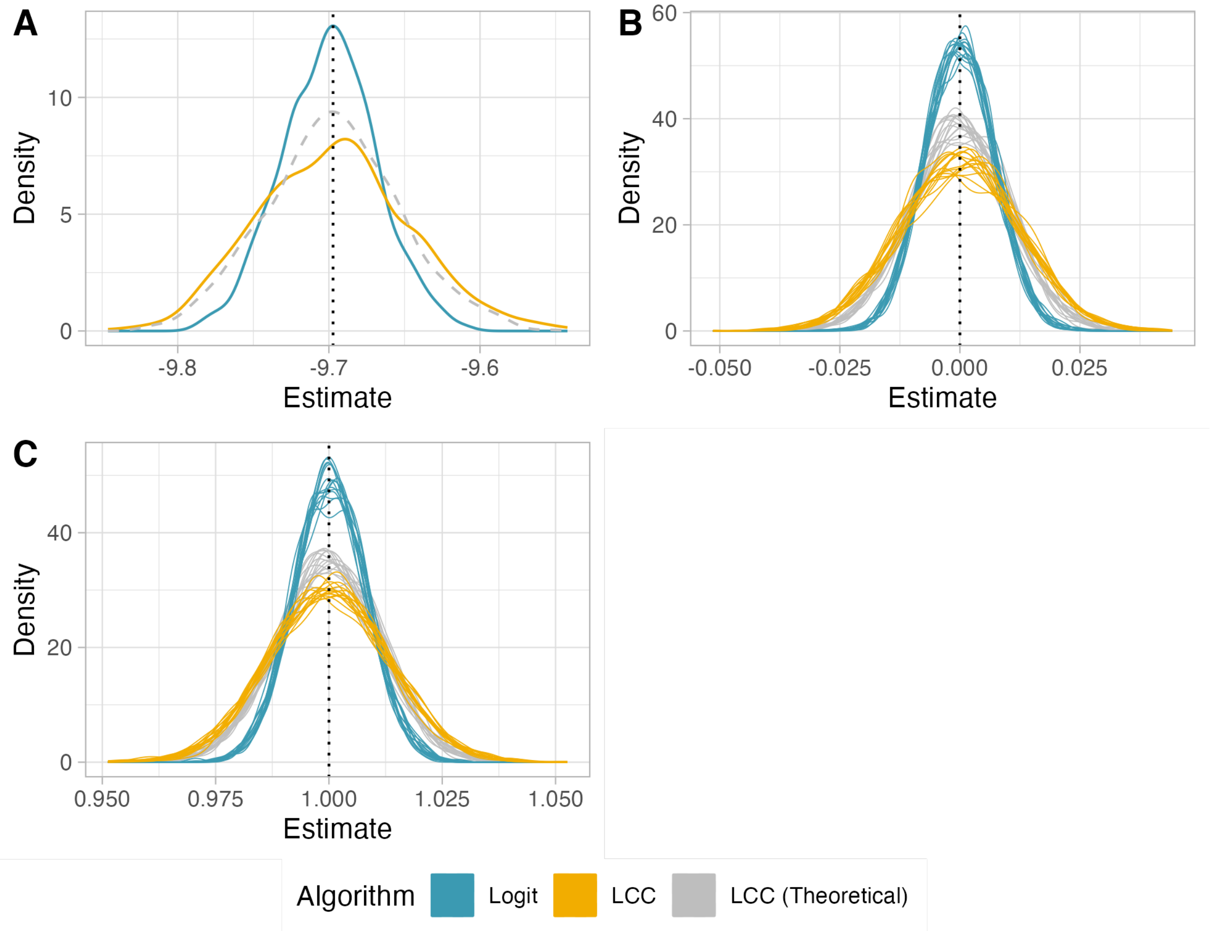

Figure 4: Simulation 2 - Approximation of $\hat{\theta}$ asymptotic distributions across methods for $N = 10^6$, where the grey line corresponds to the theoretical distribution of the LCC estimate (results from $1,000$ runs). **Panel A**: True $\alpha = -9.7$ **Panel B**: True $\beta = 0$, namely $\beta_1, \ldots, \beta_{\frac{k}{2}} = 0$. **Panel C**: True $\beta = 1$, namely $\beta_{\frac{k}{2}+1}, \ldots, \beta_k = 1$.

necessarily impact the level of conditional imbalance in the dataset. This is because conditional imbalance is defined by the subset within the feature space where $P(Y = 1) \approx 1$, which consists of $x \in X$ with $x_{\frac{k}{2}}, \ldots, x_k \approx 1$. Thus, the level of conditional imbalance in this DGP setup is dependent on the number of regressors; the larger the value of k, the more conditional imbalance there will be. Nevertheless, since increasing $P(Y = 1)$ will result in a smaller LCC's N_s size and consequently lead to a dimensionality problem for large k, I will also be varying the number of regressors in subsequent simulations. This will create conditions for more or less conditional imbalance, which will also be analyzed.

I will further elaborate on the analysis of Fithian and Hastie (2014) and estimate the logit squared bias and variance as well. The idea is not to compare the methods' performance against logit; the logistic regression estimates will still be more unbiased and consistent than the others because the model uses the full sample information. Nevertheless, I will use the logit results as a reference point to assess the method's effectiveness as an alternative to the logit model on large sample sizes to reduce computational costs.

Let $P(Y = 0) = \{0.7, 0.8, 0.9, 0.95, 0.99\}$ be the marginal imbalance variation from mild to severe. Table 3 shows the simulation's results for the baseline setup with $N = 10^5$ and $k = 30$. This scenario reflects a setup with large sample size and high conditional imbalance. As expected, LCC outperforms CC and WCC in terms of bias and variance for $P(Y = 0) \leq 0.9$, the setup of Simulation 5.2 and Fithian and Hastie (2014) baseline study. It even outperforms the logit model in squared bias, and the empirical variance factor of 3 discussed in Simulation 5.3 partially holds. However, the method's performance decays drastically for $P(Y = 0) = 0.95$ and $P(Y = 0) = 0.99$, leading to an extremely inflated bias and variance for the latter one. All the methods seem more or less affected by the severe imbalance, but LCC shows the most extreme results. These findings can be explained by the reduction of the subsample size combined with a large number of regressors, which increases the complexity of the problem.

To control for this, Table 4 shows the results for a simulation that decreases the number of regressors to $k = 10$. This, in contrast, will decrease the degree of conditional imbalance in the sample, which will likely affect the performance of LCC but improve the performance of CC and, in the cases of mild imbalance, WCC as well. The findings displayed in the table support this intuition. For mild marginal imbalance levels, CC, WCC, and LCC perform very similarly in bias and variance, and WCC shows a lower loss of efficiency. As the imbalance gets larger, WCC gets relegated, but CC and LCC still behave very similarly. For severe levels of imbalance, CC outperforms the other two subsampling methods in terms of bias and variance, showing the superiority of CC in highly marginal imbalanced datasets. As an additional robustness check, Table 10 in the Appendix shows the result for a mild level of conditional imbalance with $k = 20$, where it shows that LCC still outperforms the methods expected in the corner case of $P(Y = 0) = 0.99$, where the method struggles with a reduced sample size with high dimensionality.

5.5 Simulation 4

The purpose of the final section in the simulation study is to observe how the methods behave when dealing with different population sample sizes N. Once again, the challenge

N	k	$P(Y=0)$	α	Method	N_s	$\widehat{bias}^2$	$\widehat{var}$	$\bar{a}(\tilde{\theta})$
10^5	30	0.7	-8.35	CC	7,000	0.0868	0.3600	-
				WCC	7,000	0.1022	0.3801	-
				LCC	3,500	6×10^{-5}	0.0550	0.0360
				Logit	10^5	0.0003	0.0171	-
10^5	30	0.8	-8.89	CC	6000	0.1196	0.4172	-
				WCC	6,000	0.1858	0.5318	-
				LCC	3,000	0.0001	0.0729	0.0308
				Logit	10^5	0.0006	0.0192	-
10^5	30	0.9	-9.70	CC	4,000	0.2539	0.6761	-
				WCC	4,000	1.0239	1.2715	-
				LCC	2,000	0.0059	0.1825	0.0226
				Logit	10^5	0.0011	0.0273	-
10^5	30	0.95	-10.44	CC	2,800	0.6462	1.1605	-
				WCC	2,800	5.5285	3.3779	-
				LCC	1,400	4.6931	18457.86	0.0181
				Logit	10^5	0.0025	0.0370	-
10^5	30	0.99	-12.10	CC	1,440	3.3686	3.5402	-
				WCC	1,440	100.58	23.756	-
				LCC	720	8×10^{28}	1×10^{30}	0.0263
				Logit	10^5	0.0135	0.0909	-

Table 3: Simulation 3 - Results from variation in class marginal imbalance, holding population size $N = 10^5$ and $k = 30$ constant (results from $1,000$ runs).

here is that a high degree of conditional imbalance may lead to an additional dimensionality problem as the sample size gets smaller. To avoid this issue from affecting the analysis of the methods' performance, I will keep the number of regressors at a constant $k = 10$ throughout the simulation. Although this means that the exercise is limited to scenarios where only the marginal imbalance varies, Simulation 5.4 has already covered cases with very small subsample sizes and high conditional imbalance that have shown the methods to perform poorly. Therefore, working with a smaller k could shed light on how the methods perform on different sample sizes without the interference of dimensionality issues.

As all techniques primarily rely on asymptotics for consistency, particularly the LCC method, it is anticipated that the bias and variance will be inflated when N and, consequently, N_s decrease. However, the performance of the methods on small sample sizes is not solely determined by the size of the entire population but also relies on a combination of the sample size N and the amount of marginal imbalance present in the sample. The sample sizes utilized in this analysis may not correspond to massive dataset setups because

N	k	$P(Y=0)$	α	Method	N_s	$\widehat{bias}^2$	$\widehat{var}$	$\bar{a}(\tilde{\theta})$
10^5	10	0.7	-3.35	CC	33,400	1×10^{-5}	0.0029	-
				WCC	33,400	1×10^{-5}	0.0032	-
				LCC	16,700	3×10^{-6}	0.0021	0.1678
				Logit	10^5	4×10^{-7}	0.0003	-
10^5	10	0.8	-3.89	CC	27,600	4×10^{-6}	0.0033	-
				WCC	27,600	4×10^{-6}	0.0043	-
				LCC	13,800	3×10^{-6}	0.0027	0.1387
				Logit	10^5	2×10^{-7}	0.0004	-
10^5	10	0.9	-4.69	CC	18,000	8×10^{-6}	0.0044	-
				WCC	18,000	2×10^{-5}	0.0090	-
				LCC	9,000	3×10^{-6}	0.0042	0.0915
				Logit	10^5	5×10^{-7}	0.0006	-
10^5	10	0.95	-5.44	CC	11,200	4×10^{-5}	0.0075	-
				WCC	11,200	5×10^{-4}	0.0244	-
				LCC	5,600	4×10^{-5}	0.0070	0.0564
				Logit	10^5	5×10^{-6}	0.0009	-
10^5	10	0.99	-7.09	CC	3,100	0.0003	0.0391	-
				WCC	3,100	0.0855	0.2832	-
				LCC	1,550	0.0039	0.0442	0.0163
				Logit	10^5	9×10^{-6}	0.0027	-

Table 4: Simulation 3 - Results from variation in class marginal imbalance, holding population size $N = 10^5$ and $k = 10$ constant (results from $1,000$ runs).

of computational limitations. However, they serve the purpose of providing a scaled-up understanding of the techniques' behavior as N changes.

Let $N = \{10^5, 10^4, 5000, 2000, 1500\}$ be the possible full population sizes to be used. As mentioned before, I will also be varying the marginal imbalance levels $P(Y = 0) = \{0.7, 0.8, 0.9, 0.95\}$ for each of the sample sizes in order to get a more comprehensive analysis for distinct DGPs. Table 9 in the Appendix shows the summary statistics of LCC subsample sizes used to pick the fixed subsamples for this exercise, whereas Tables 11, 12, 13, and 14 contain all the numerical results for all the simulations in a similar manner as the tables presented in Simulation 5.4.

Figure 5 displays a visual representation of the empirical squared bias outcomes for all simulation setups. Individually, the x-axis in each graph shows the squared root of the squared bias, and the y-axis displays the log transformation of the fixed subsample size N_s used for each setup. As expected, the squared bias decreases as N_s increases as a result of a larger population sample size. Panel A and B depict the simulation findings for a mild marginal imbalance of $P(Y=0)=0.7$ and $P(Y=0)=0.8$, respectively. In both cases, the LCC outperforms the other two subsampling methods in terms of squared bias for all N_s values, even for very small subsample sizes. CC and WCC seem to perform quite similarly for the lowest level of imbalance, but, as expected, WCC starts to perform worse as the imbalance level increases. The most dramatic deterioration, however, is seen by LCC in scenarios with high to severe marginal imbalance and small sample sizes, as shown in Panels C and D. Specifically, the problem arises in cases with a full sample size of $1,500$ for $P(Y=0)=0.9$ and $2,000$ or lower for $P(Y=0)=0.95$. The explosion in the bias is so severe that it actually reaches 10^{25} for the last corner case.

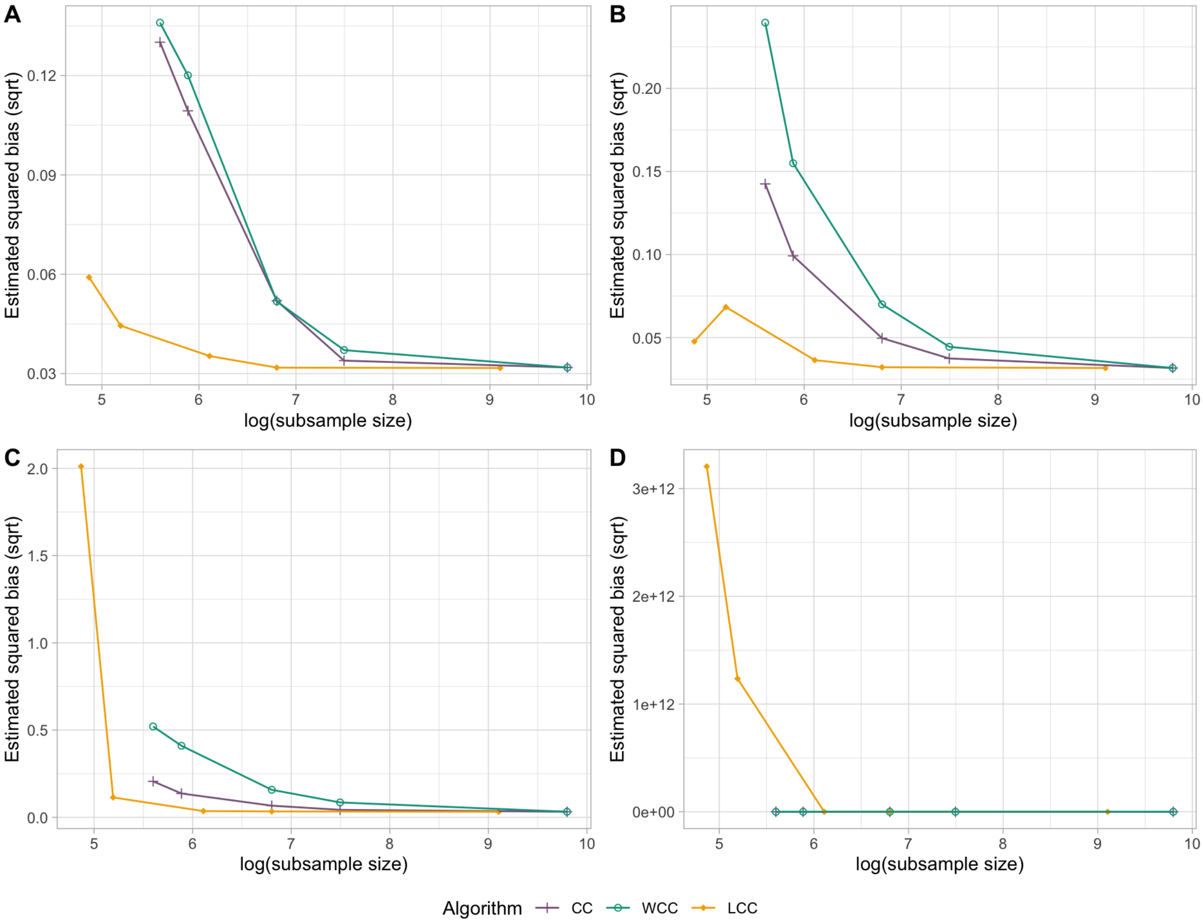

Figure 5: Simulation 4 - Relationship between estimated squared bias and subsample size by Algorithm under different levels of marginal imbalance. **Panel A:** Mild imbalance $P(Y=0)=0.7$. **Panel B:** Mild-high imbalance $P(Y=0)=0.8$. **Panel C:** High imbalance $P(Y=0)=0.9$. **Panel C:** Severe imbalance $P(Y=0)=0.95$.

The performance of the methods in terms of variance is shown in Figure 6. Here, for mild levels of marginal imbalance illustrated in Panels A and B, the algorithms perform very similarly, with LCC doing as well as the other two. This similarity can be attributed to the fact that when there is low conditional imbalance, LCC does not have any significant advantage over the other two methods, resulting in a similar performance between the three. Once again, however, problems arise as N gets smaller and $P(Y = 0)$ becomes larger, as in the last two panels. As with the squared bias, the variance of LCC also gets extremely inflated for sample sizes of $2,000$ or lower and large to severe marginal imbalance.

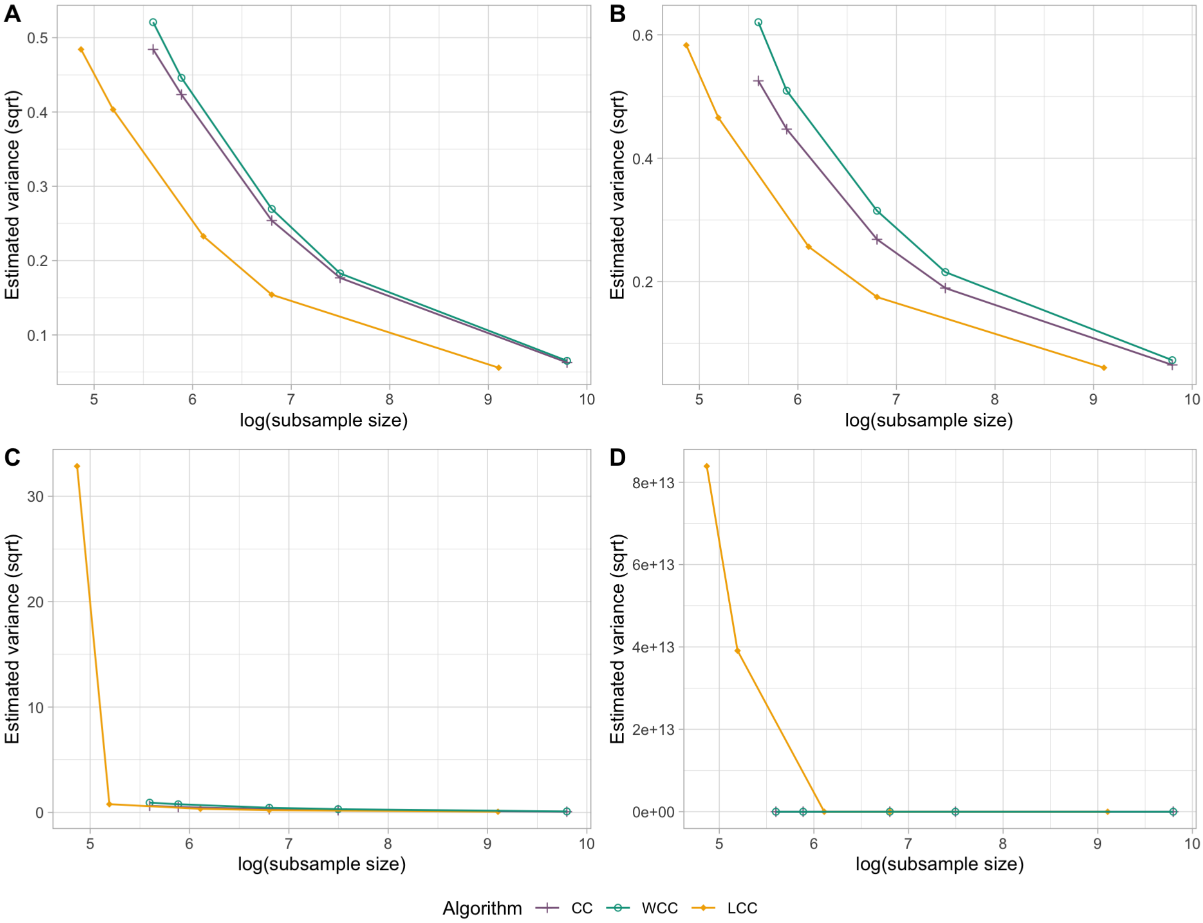

Figure 6: Simulation 4 - Relationship between estimated variance and subsample size by Algorithm under different levels of marginal imbalance. **Panel A:** Mild imbalance $P(Y = 0) = 0.7$. **Panel B:** Mild-high imbalance $P(Y = 0) = 0.8$. **Panel C:** High imbalance $P(Y = 0) = 0.9$. **Panel C:** Severe imbalance $P(Y = 0) = 0.95$.

Thus, a reasonable conclusion is that in "corner-case" scenarios, LCC will usually be the worst method among the three. Looking at Table 14, a more appropriate choice among the subsampling methods will be the CC algorithm since WCC also shows poor performance due to the variance induced from the disproportional weights in high imbalance setups. CC not only seems pretty stable in these scenarios, but it is the one method that gets closer to the bias and variance empirical benchmark of logistic regression. Once again, it is important to note that under such a small N, the logistic regression will always be preferred as it is still the method that shows the lowest squared bias and variance, and it is still a case under which it would be feasible to compute the full sample $\hat{\theta}_{MLE}$. This exercise serves as a demonstration, however, that LCC could fail in instances where the imbalance is extremely severe, even for large N.

Wang, Zhu, and Ma (2018) exhibit similar conclusions. The authors conducted simulation experiments and showcased the performance of various two-step subsampling techniques in terms of empirical MSE under different subsample size ratios. LCC was used as a comparison estimator together with uniform subsampling. Their findings indicate that LCC does not perform well on small subsample sizes, mainly because the effective sample size is smaller than the other methods under study. They conclude that LCC fails to approximate the full sample parameters under such conditions.

6 Data application

6.1 Adult income dataset

The Census Income dataset, sometimes referred to as the Adult dataset, was extracted by Becker and Kohavi (1996) from the USA 1994 Census database. It has been widely used as a benchmark dataset in many machine learning research papers such as Poulos and Valle (2018), Chawla, Bowyer, et al. (2002), Menardi and Torelli (2014), Yao and Wang (2021), among others. It contains $48,842$ observations in total and includes 14 regressors such as age, type of work (workclass), survey final weight (fnlwgt), level of education (education), number of years of education (education-num), marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country, and income (class). The last one is the target value, where the purpose of the classification task is to predict whether the person will make more than 50K per year.

This dataset is an ideal choice for a real-world dataset application within the frame of this thesis for two main reasons. Firstly, it contains a large number of observations, making it a popular subject in the research area of optimal subsampling methods for handling massive data, as demonstrated by studies such as Wang, Zhu, and Ma (2018) and Yao and Wang (2021). Secondly, the data contains a moderate to high marginal imbalance, with $P(Y = 1) \approx 0.25$. For the regression analysis, I follow Yao and Wang (2021) and use 5 out of the 14 features as relevant regressors for predicting y. After removing missing data, the sample consists of $N = 45,222$ instances, where $11,208$ of them belong to class $Y = 1$, i.e., adults who earned more than 50K per year in 1994. Table 5 provides summary statistics of the full sample used for the analysis, and additional summary statistics by class can be found in Section 9.2.

Statistic	N	Mean	St. Dev.	Min	Max
age	45,222	38.548	13.218	17	90
fnlwgt	45,222	189,734.700	105,639.200	13,492	1,490,400
education_num	45,222	10.118	2.553	1	16
capital_gain	45,222	1,101.430	7,506.430	0	99,999
capital_loss	45,222	88.595	404.956	0	4,356
hours_per_week	45,222	40.938	12.008	1	99
class_1	11,208	1.000	0.000	1	1
class_0	34,014	1.000	0.000	1	1

Table 5: Summary Statistics - Income dataset (missing values removed)

One of the challenges is to determine whether or not there is conditional imbalance in the data. To address this, I visually analyze the PCA components as in Section 5.1. This can help identify whether the data clusters of each class overlap or not, which in turn could be an indicator of the presence of conditional imbalance. For instance, when the clusters cannot be fully separated by a hyperplane but are still distinguishable between them, it is possible that there exists conditional imbalance, which means that the probabilities of $P(Y = 1)$ for certain feature values can be easily predicted. Figure 7 illustrates the PCA components for each class, which appear dense due to the high number of observations. However, from the visualization, it is clear that there is a significant amount of class overlap, suggesting that this may be a case of mild-to-high marginal imbalance but low conditional imbalance, similar to the second numerical exercise of Simulation 5.5.

6.2 Methodological considerations

In order to estimate the empirical variance of the estimators similarly to a Monte Carlo study, 100 artificial samples of size $N \times m$ were created with the help of stratified sampling without replacement. Bickel and Sakov (2008) refer to this approach as the m-bootstrap, where m refers to the block size or the percentage of data points from the original sample to be used to construct the new sample. Stratified subsampling works such that each

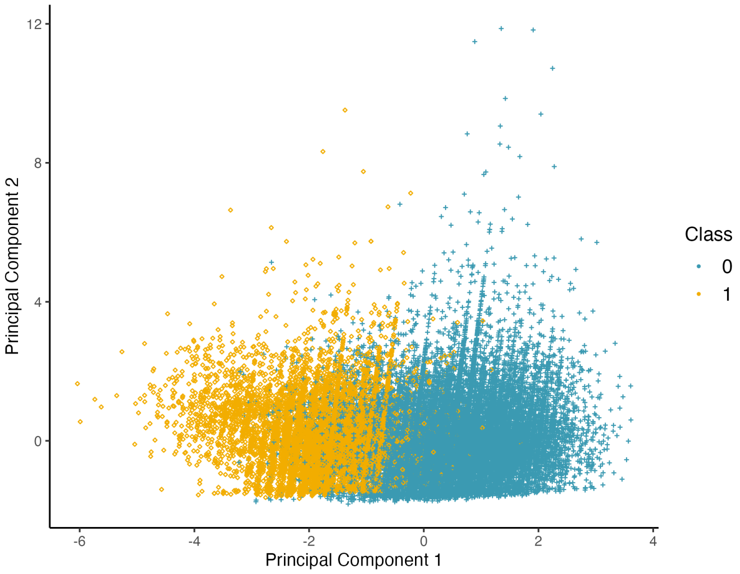

Figure 7: PCA plot Income dataset (Becker and Kohavi 1996).

time, random data points are taken only once from the full sample in order to create a subsample of smaller size while maintaining the original proportions of cases and controls.

The main practical problem while constructing the artificial samples is the choice of the block size m (Politis, Romano, and Wolf 1999). The selection of this block size is crucial since a size that is too small could hinder the subsampling estimates by further reducing their effective subsampling size. To ensure accuracy, I tested different values of block size $m = \{0.95, 0.9, 0.85, \ldots, 0.45, 0.4\}$. In this way, for each value of m, a set of 100 independent artificial samples was created.

An LCC pass was fitted to each data in a sample set, and its average subsample size was then used to obtain the LCC's fixed N_s, as in the Simulation section. However, a key difference here is that the effective subsample size for CC and WCC was not enough to ensure that they would have roughly twice as many data points as LCC. Thus, their acceptance probability was set such that $a(1) = 1$ and $a(0) = (P(Y = 1)a(1))/P(Y = 0)$; in other words, it was necessary to subsample all cases and the same amount of controls. This approach was still not enough to provide a CC and WCC subsample size twice as large as the one for LCC, but it did help to approximate their desired dimension (see Appendix Section 9.2.3 for the exact fixes N_s used in the analysis). The algorithms were then fitted to each artificial sample, thus obtaining 100 realizations of each subsampling scheme and m value. The variance of the estimates was calculated as outlined in Section 4.

6.3 Results

The results for the empirical variance of the estimates are shown in Table 6. The block size $m = 0.85$ shows the lowest empirical variances for all methods, and thus it is chosen as a robust block size for the analysis. Recall that this dataset refers to a scenario of large N, small k, mild-to-large marginal imbalance, and what appears to suggest low conditional imbalance. In this regard, we could expect that the results resemble the findings presented in panels 1 and 2 from Table4 of Section 5.4, where it was shown that for mild marginal imbalance levels, CC, WCC, and LCC perform similar in terms of variance.

m	$\widehat{var}_{logit}$	$\widehat{var}_{CC}$	$\widehat{var}_{WCC}$	$\widehat{var}_{LCC}$
0.95	0.0004	0.0071	0.0084	0.0077
0.90	0.0009	0.0100	0.0117	0.0084
0.85	0.0014	0.0097	0.0106	0.0084
0.80	0.0027	0.0114	0.0136	0.0103
0.75	0.0026	0.0110	0.0131	0.0125
0.70	0.0034	0.0127	0.0150	0.0148
0.65	0.0057	0.0121	0.0129	0.0200
0.60	0.0062	0.0195	0.0230	0.0161
0.55	0.0085	0.0197	0.0232	0.0207
0.50	0.0089	0.0267	0.0320	0.0304
0.45	0.0119	0.0337	0.0373	0.0305
0.40	0.0125	0.0301	0.0349	0.0294

Table 6: Results for the Income dataset - Empirical variance of the $\hat{\theta}$ estimates for the three subsampling methods under different levels of subsampling m for creating the synthetic samples.

As expected, the results for the Income dataset seem to agree with both the theory of the methods and the results from the numerical exercises. For $m = 0.85$, LCC shows the lowest empirical variance among all case-control methods. However, unlike scenarios with high conditional imbalance, LCC does not fully dominate the other methods in terms of efficiency. In fact, its performance is not very different from the CC algorithm, which is not surprising given that CC has shown to perform well under high marginal imbalance. On the other hand, despite having the worst variance performance, WCC does not show to be underperforming so severely, mostly because the variation in the weights is not severe when the imbalance is mild.

Moreover, the LCC estimate displays an empirical variance 6 times higher than the empirical variance of the logistic regression on the full sample. One explanation could be that the effective sample size for the LCC and its pilot is too small to approximate an asymptotic behavior ($N_s = 11,242$, see Appendix). However, in Simulation 5.3, LCC had access to a much smaller sample size, and yet the estimate showed to be very close to its asymptotic behavior. Thus, it seems more likely that the loss in efficiency is primarily driven by the data dependency of the pilot and the low presence of conditional imbalance.

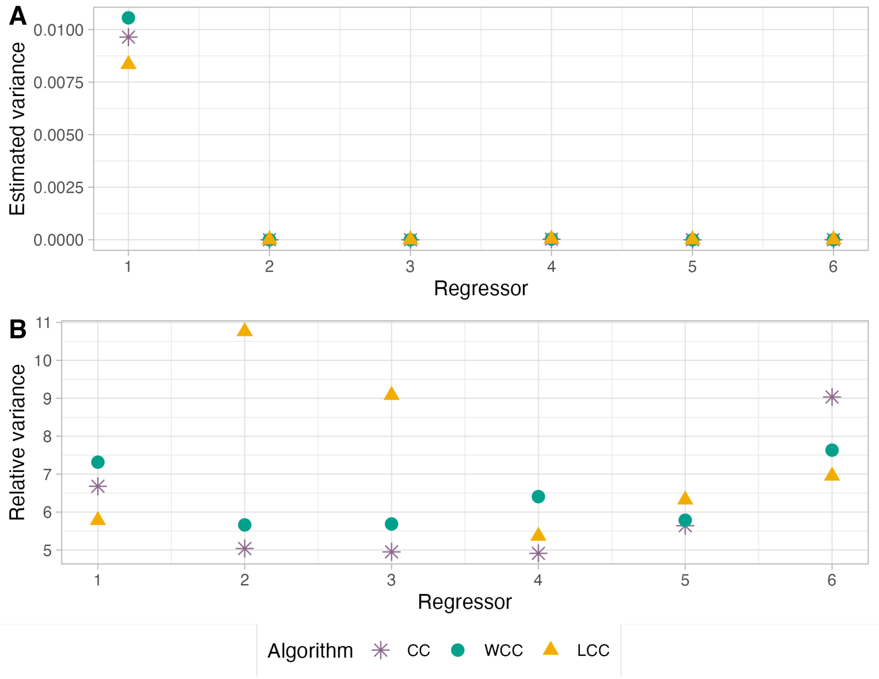

Figure 8: Empirical variance of coefficients, where: 1=intercept, 2=age, 3=fnlwgt, 4=years of education, 5=capital loss, and 6=hours per week. **Panel A.** Variance ratio of coefficients for the 3 different subsampling algorithms relative to full sample logistic regression variance. **Panel B.** Variance of coefficients for the 3 different subsampling algorithms.

Figure 8 Panel A shows the empirical variance of the coefficients by algorithm. The intercept term shows the largest variance of all, with LCC's coefficient displaying the lowest. For the other regressors, the variances are small and similar across methods. However, Panel B further decomposes the analysis and shows the variance of each coefficient relative to the variance of the full sample logistic regression. LCC shows the lowest ratio against the logistic regression estimate for the intercept but exhibits the largest relative variance for two slope coefficients, even reaching a ratio of 11 times the full sample logistic regression for age's estimated coefficient. Overall, the analysis of the individual variance suggests that some of the LCC estimates have large individual noise, potentially caused by the utilization of a data-dependent pilot with its own variance, which may be impacting the general efficiency of LCC. Additionally, it is worth noting that logistic regression yields the most efficient coefficient among all methods and it is clear that, given the characteristics of the dataset, the trade-off between efficiency and computational gains is too costly.

7 Conclusion

In the context of large datasets, this thesis addressed the challenge of designing effective subsampling schemes for reducing computational burdens while still approximating the parameter estimates of the full sample. It has been shown that this task is particularly challenging when the data exhibits marginal or conditional imbalance, making it harder to create subsamples that balance statistical efficiency with computational gains. The study focused on case-control subsampling methods such as standard case-control (CC), weighted case-control (WCC), and local case-control (LCC), which aim to alleviate this problem by assigning different acceptance probabilities to cases and controls. It is important to note that the sample sizes used in this analysis might not correspond to massive datasets due to limitations in computational power. However, it serves to showcase a scaled-down exercise of the effectiveness of the methods under different scenarios.

The simulation study and data application section revealed that the performance of the methods is heavily influenced by the nature of the DGP. In this regard, their performance varies depending on factors such as the degree of marginal and conditional imbalance, the full sample size, and the extent to which their theoretical assumptions are valid. For instance, while the CC estimate seems to be the most efficient even for levels of high-to-severe marginal imbalance, it fails to exploit the conditional imbalance present in the data. The WCC estimate, on the other hand, is asymptotically unbiased and consistent and shows good efficiency for mild levels of marginal imbalance. Moreover, it serves as a pilot estimate in the first stage of the LCC procedure. Yet, one strong drawback of the method is that for large-to-severe levels of imbalance, the variance of the estimator quickly escalates, as the variation in its computational weights adds additional noise to the estimate.

LCC was proposed by Fithian and Hastie (2014) as a good alternative to CC and WCC for accounting conditional imbalance in the data and model misspecification. Although this thesis did not study LCC's behavior within misspecification, it aimed to extend the original paper's analysis and showcase LCC's performance for different DGPs. Again, the numerical exercises and real-data implementation showed that, although LCC outperforms CC and WCC for large N and high levels of conditional imbalance, it still shows some drawbacks in marginal imbalanced datasets and small samples. In particular, for high-to-severe marginal imbalanced data and moderate sample sizes, LCC shows a dramatic efficiency loss, with severe inflation for both squared bias and variance. It was also suggested that the consistency of the LCC estimator could be hindered by the inconsistency of the pilot estimate in small samples. In contrast, the estimate's capability of approximating its asymptotic distribution could be affected by the use of a data-dependent pilot in finite samples, which is a direct violation of one of the estimate's asymptotic assumptions.

It is important to be aware of the limitations of the findings presented in this thesis. One such limitation is the interpretation of results that used fixed subsample sizes N_s for comparing the methods across settings, as it was suggested in Fithian and Hastie (2014). For this specific study, that entailed estimating the median subsample size for LCC and then allowing CC and WCC to have just twice as many data points as LCC. For most of the cases in the simulation study, this translated into trimming CC and WCC subsample sizes, possibly affecting their statistical performance, especially as all methods rely heavily upon large N for unbiasedness and consistency. Thus, although this procedure allows for a better comparison of the estimators' performance, it should be taken into consideration when interpreting the results.

Nonetheless, further work has been undertaken since Fithian and Hastie (2014)'s publication on LCC. For example, L. Han et al. (2020) extended the LCC framework for solving large-scale multiclass logistic regression problems and proposed a new subsampling scheme called Local Uncertainty Sampling (LUC). They proved that their method always has a lower variance than uniform subsampling, and their empirical studies show that LUC improves upon CC and LCC under several scenarios. In addition, Shen, Chen, and W. Yu 2021 developed a new approach based on LCC, called Surprised Sampling Design. This version not only accepts a logistic model but rather supports several loss functions, making it compatible not only with the Logit but also with a wider range of machine learning models. Finally, Wang (2020) proved that heavily undersampling controls while maintaining the cases do not sacrifice efficiency, and their estimator has shown to have an identical asymptotic distribution as the full sample θ_{MLE}. Thus, an interesting avenue for future research could be to compare LCC's performance against the above-mentioned novel estimators that aim to generalize LCC's framework and overcome its limitations.

8 Bibliography

Anderson, James A (1972). "Separate sample logistic discrimination". In: *Biometrika* 59.1, pp. 19–35.

Becker, Barry and Ronny Kohavi (1996). *Adult.* UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Bickel, Peter J and Anat Sakov (2008). "On the choice of m in the m out of n bootstrap and confidence bounds for extrema". In: *Statistica Sinica*, pp. 967–985.

Borgan, Ørnulf, Norman Breslow, Nilanjan Chatterjee, Mitchell H Gail, Alastair Scott, and Chris J Wild (2018). *Handbook of statistical methods for case-control studies.* CRC Press.

Chawla, Nitesh, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Chawla, Nitesh, N Japkowicz, and A Kotcz (2014). *Editorial: Special Issue on Learning from Imbalanced Data Sets, SIGKDD Explor.*

Cheng, Qianshun, HaiYing Wang, and Min Yang (2020). "Information-based optimal subdata selection for big data logistic regression". In: *Journal of Statistical Planning and Inference* 209, pp. 112–122.

Dahl, David B, David Scott, Charles Roosen, Arni Magnusson, Jonathan Swinton, Ajay Shah, Arne Henningsen, Benno Puetz, Bernhard Pfaff, Claudio Agostinelli, et al. (2019). *Package 'xtable'.*

DasGupta, Anirban (2011). *Selected Works of Debabrata Basu.* Springer Science & Business Media.

Einav, Liran and Jonathan Levin (2014). "The data revolution and economic analysis". In: *Innovation Policy and the Economy* 14.1, pp. 1–24.

Elliott, Michael R and Roderick JA Little (2000). "Model-based alternatives to trimming survey weights". In: *Journal of Official Statistics Stockholm* 16.3, pp. 191–210.

Fan, Jianqing, Fang Han, and Han Liu (2014). "Challenges of big data analysis". In: *National science review* 1.2, pp. 293–314.

Fithian, William and Trevor Hastie (2014). "Local case-control sampling: Efficient subsampling in imbalanced data sets". In: *Annals of statistics* 42.5, p. 1693.

Han, Lei, Kean Ming Tan, Ting Yang, and Tong Zhang (2020). "Local uncertainty sampling for large-scale multiclass logistic regression". In: *The Annals of Statistics* 48.3, pp. 1770–1788.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.* Vol. 2. Springer.

He, Haibo and Edwardo A Garcia (2009). "Learning from imbalanced data". In: *IEEE Transactions on knowledge and data engineering* 21.9, pp. 1263–1284.

Hlavac, Marek (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables.* R package version 5.2.3. Social Policy Institute. Bratislava, Slovakia. URL: `https://CRAN.R-project.org/package=stargazer`.

Kassambara, Alboukadel et al. (2020). "ggpubr:"ggplot2" based publication ready plots". In: *R package version 0.4. 0* 438.

King, Gary and Langche Zeng (2001). "Logistic regression in rare events data". In: *Political analysis* 9.2, pp. 137–163.

Koh, Kwangmoo, Seung-Jean Kim, and Stephen Boyd (2007). "An interior-point method for large-scale l1-regularized logistic regression". In: *Journal of Machine learning research* 8.Jul, pp. 1519–1555.

Lee, Sokbae and Serena Ng (2020). "An econometric perspective on algorithmic subsampling". In: *Annual Review of Economics* 12, pp. 45–80.

Li, Runze, Dennis KJ Lin, and Bing Li (2013). "Statistical inference in massive data sets". In: *Applied Stochastic Models in Business and Industry* 29.5, pp. 399–409.

Li, Yan, Barry I Graubard, and Ralph DiGaetano (2011). "Weighting methods for population-based case–control studies with complex sampling". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60.2, pp. 165–185.

Maalouf, Maher (2011). "Logistic regression in data analysis: an overview". In: *International Journal of Data Analysis Techniques and Strategies* 3.3, pp. 281–299.

Manski, Charles F and Steven R Lerman (1977). "The estimation of choice probabilities from choice based samples". In: *Econometrica: Journal of the Econometric Society*, pp. 1977–1988.

McCullagh, Peter and John A Nelder (1989). *Generalized linear models.* Chapman and Hall.

Menardi, Giovanna and Nicola Torelli (2014). "Training and assessing classification rules with imbalanced data". In: *Data mining and knowledge discovery* 28, pp. 92–122.

Montesinos, Osval, Abelardo Montesinos, and José Crossa (2022). *Multivariate statistical machine learning methods for genomic prediction*. Springer Nature.

Ng, Serena (2017). *Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data*. Working Paper 23673. National Bureau of Economic Research.

O'Brien, Robert and Hemant Ishwaran (2019). "A random forests quantile classifier for class imbalanced data". In: *Pattern recognition* 90, pp. 232–249.

Overton, W Scott and Stephen V Stehman (1995). "The Horvitz-Thompson theorem as a unifying perspective for probability sampling: with examples from natural resource sampling". In: *The American Statistician* 49.3, pp. 261–268.

Politis, Dimitris N, Joseph P Romano, and Michael Wolf (1999). *Subsampling*. Springer Science & Business Media.

Poulos, Jason and Rafael Valle (2018). "Missing data imputation for supervised learning". In: *Applied Artificial Intelligence* 32.2, pp. 186–196.

Prentice, Ross L and Ronald Pyke (1979). "Logistic disease incidence models and case-control studies". In: *Biometrika* 66.3, pp. 403–411.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Ram, Karthik and Hadley Wickham (2018). *wesanderson: A Wes Anderson Palette Generator*. R package version 0.3.6. URL: `https://CRAN.R-project.org/package=wesanderson`.

Rose, Sherri and Mark J van der Laan (2008). "Simple optimal weighting of cases and controls in case-control studies". In: *The International Journal of Biostatistics* 4.1.

Scott, Alastair and Chris Wild (1986). "Fitting logistic models under case-control or choice based sampling". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 48.2, pp. 170–182.

— (2002). "On the robustness of weighted methods for fitting models to case–control data". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.2, pp. 207–219.

Shen, Xinwei, Kani Chen, and Wen Yu (2021). "Surprise sampling: Improving and extending the local case-control sampling". In: *Electronic Journal of Statistics* 15.1.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: `https://www.stats.ox.ac.uk/pub/MASS4/`.

Wang, HaiYing (2019). "More efficient estimation for logistic regression with optimal subsamples". In: *Journal of machine learning research* 20.

— (2020). "Logistic regression for massive data with rare events". In: *International Conference on Machine Learning*. PMLR, pp. 9829–9836.

Wang, HaiYing, Min Yang, and John Stufken (2019). "Information-based optimal subdata selection for big data linear regression". In: *Journal of the American Statistical Association* 114.525, pp. 393–405.

Wang, HaiYing, Rong Zhu, and Ping Ma (2018). "Optimal subsampling for large sample logistic regression". In: *Journal of the American Statistical Association* 113.522, pp. 829–844.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: `https://ggplot2.tidyverse.org`.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan (2023). *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org, https://github.com/tidyverse

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich (2023). *tidyr: Tidy Messy Data*. https://tidyr.tidyverse.org, https://github.com/tidyverse/tidyr.

Yao, Yaqiong and HaiYing Wang (2021). "A review on optimal subsampling methods for massive datasets". In: *Journal of Data Science* 19.1, pp. 151–172.

Yu, Jun, Mingyao Ai, and Zhiqiang Ye (2023). "A review on design inspired subsampling for big data". In: *Statistical Papers*, pp. 1–44.

9 Appendix

9.1 Formal derivations

9.1.1 Formal derivation of intercept adjustment in Section 3.2

The following is based on the Appendix B in King and Zeng 2001, although own derivations are presented. We can get Equation 17 from Equation 16 as follow:

$$P(y = 1 \mid x, d)A_1B = \frac{P(y = 1 \mid x, d)\tau/\bar{y}}{P(y = 1 \mid x, d)(\tau/\bar{y}) + [1 - P(y = 1 \mid x, d)](1 - \tau)/(1 - \bar{y})}$$

$$= \frac{P(y = 1 \mid x, d)\tau/\bar{y}}{[P(y = 1 \mid x, d)\tau(1 - \bar{y} + \bar{y}(1 - \tau) - \bar{y}(1 - \tau)P(y = 1 \mid x, d)]/\bar{y}(1 - \bar{y})}$$

$$= \left[\frac{P(y = 1 \mid x, d)\tau\bar{y}(1 - \bar{y}) + \bar{y}^2(1 - \tau) - \bar{y}^2(1 - \tau)P(y = 1 \mid x, d)}{P(y = 1 \mid x, d)\tau\bar{y}(1 - \bar{y})}\right]^{-1}$$

$$= \left[1 + \frac{1}{P(y = 1 \mid x, d)}\left(\frac{1 - \tau}{\tau}\right)\left(\frac{\bar{y}}{1 - \bar{y}}\right) - \left(\frac{\bar{y}}{1 - \bar{y}}\right)\left(\frac{1 - \tau}{\tau}\right)\right]^{-1}$$

$$= \left[1 + \left(\frac{1}{P(y = 1 \mid x, d)} - 1\right)\left(\frac{1 - \tau}{\tau}\right)\left(\frac{\bar{y}}{1 - \bar{y}}\right)\right]^{-1}$$

Then, with the assumption that the model follows a logistic function with $Pr(Y = 1 \mid x, d) = 1/1 + e^{-(\alpha + x_i^\intercal \beta)}$, we have:

$$P(y = 1 \mid x, d)A_1B = \left[1 + \left(\frac{1}{1/1/1 + e^{-(\alpha + x_i^\intercal \beta)}} - 1\right)\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]^{-1}$$

$$= \left[1 + e^{-(\alpha + x_i^\intercal \beta)}\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]^{-1}$$

$$= \left[1 + e^{-(\alpha + x_i^\intercal \beta)} e^{\log\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]}\right]^{-1}$$

$$= \left[1 + e^{-(\alpha + x_i^\intercal \beta) + \log\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]}\right]^{-1}$$

9.2 Additional tables

9.2.1 Summary statistics for LCC subsample sizes across simulations

Simulation 1

	N	Min.	X1st.Qu.	Median	Mean	X3rd.Qu.	Max.
1	10^5	2013	2109.50	2145.50	2142.33	2181.50	2263

Table 7: Simulation 1 - Summary statistics for LCC's subsample size (results from 100 runs).

Simulation 3

	$P(Y=0)$	Min.	X1st.Qu.	Median	Mean	X3rd.Qu.	Max.
				$k = 30$			
1	0.70	3440	3526.00	3564.50	3570.87	3615.00	3727
2	0.80	2879	3006.25	3034.00	3037.39	3077.50	3184
3	0.90	2017	2102.00	2143.50	2141.51	2174.25	2306
4	0.95	1320	1414.00	1440.50	1439.35	1468.25	1532
5	0.99	562	727.00	866.50	872.37	972.25	1296
				$k = 10$			
1	0.70	16489	16686.75	16789.50	16775.96	16842.00	17102
2	0.80	13542	13799.75	13865.00	13864.25	13909.50	14211
3	0.90	8959	9104.00	9154.50	9152.26	9210.00	9354

4	0.95	5389	5583.50	5642.00	5630.47	5684.00	5861
5	0.99	1471	1551.00	1584.50	1600.48	1636.00	1858
				$k = 20$			
1	0.70	7362	7505.00	7560.50	7556.70	7603.25	7774
2	0.80	6134	6318.50	6371.50	6368.14	6423.25	6586
3	0.90	4225	4354.50	4400.00	4398.64	4445.00	4541
4	0.95	2717	2827.00	2865.50	2864.52	2912.25	3009
5	0.99	903	980.00	1043.50	1063.64	1124.25	1483

Table 8: Simulation 3 - Summary statistics for LCC's subsample size (results from 100 runs).

Simulation 4

				$P(Y = 0) = 0.9$			
	N	Min.	X1st.Qu.	Median	Mean	X3rd.Qu.	Max.
1	10^5	8957	9087.25	9140.00	9146.78	9189.50	9407
2	10^4	859	898.00	915.00	913.70	927.00	970
3	5000	420	445.75	460.00	460.26	471.25	521
4	2000	155	180.00	188.50	188.97	198.25	225
5	1500	116	133.75	141.00	142.15	149.00	174
				$P(Y = 0) = 0.7$			
1	10^5	16489	16686.75	16789.50	16775.96	16842.00	17102
2	10^4	1565	1647.00	1675.50	1673.52	1703.25	1756
3	5000	778	819.75	837.00	838.79	852.50	918
4	2000	302	326.00	337.00	336.24	344.00	372
5	1500	221	241.75	252.00	250.36	257.25	285
				$P(Y = 0) = 0.8$			
1	10^5	13611	13813.75	13868.50	13875.55	13942.25	14066
2	10^4	1314	1366.50	1391.50	1390.81	1413.25	1481
3	5000	621	677.75	693.00	692.44	708.25	757
4	2000	244	270.75	279.00	279.00	289.00	322
5	1500	181	199.75	211.00	209.88	219.25	244
				$P(Y = 0) = 0.95$			
1	10^5	5454	5576.50	5621	5626.16	5680.00	5798

2	10^4	525	553.00	570	572.25	584.25	643
3	5000	255	281.00	295	293.38	302.00	358
4	2000	98	121.75	129	132.25	139.25	201
5	1500	78	92.75	104	106.37	115.25	151

Table 9: Simulation 4 - Summary statistics for LCC's subsample size (results from 100 runs).

9.2.2 Additional tables from simulation section

Simulation 3: $\mathbf{N = 10^5}$ and $\mathbf{k = 20}$

N	k	$P(Y=0)$	α	Method	N_s	$\widehat{bias}^2$	$\widehat{var}$	$\bar{a}(\tilde{\theta})$
10^5	20	0.7	-5.85	CC	15,000	0.0009	0.0409	-
				WCC	15,000	0.0014	0.0447	-
				LCC	7,500	8×10^{-6}	0.0122	0.0757
				Logit	10^5	2×10^{-5}	0.0035	-
10^5	20	0.8	-6.39	CC	12,600	0.0012	0.0496	-
				WCC	12,600	0.0016	0.0586	-
				LCC	6,300	1×10^{-5}	0.0148	0.0638
				Logit	10^5	2×10^{-5}	0.0041	-
10^5	20	0.9	-7.19	CC	8,800	0.0024	0.0680	-
				WCC	8,800	0.0095	0.1117	-
				LCC	4,400	2×10^{-5}	0.0218	0.0441
				Logit	10^5	6×10^{-5}	0.0054	-
10^5	20	0.95	-7.94	CC	5,600	0.0042	0.1004	-
				WCC	5,600	0.0561	0.2671	-
				LCC	2,800	6×10^{-5}	0.0405	0.0289
				Logit	10^5	3×10^{-5}	0.0078	-
10^5	20	0.99	-9.60	CC	2,000	0.0423	0.2723	-
				WCC	2,000	3.8440	2.1392	-
				LCC	1,000	0.9599	423.48	0.0133
				Logit	10^5	0.0007	0.0191	-

Table 10: Simulation 3 - Results from variation in class marginal imbalance, holding population size $N = 10^5$ and $k = 20$ constant (results from $1,000$ runs).

Simulation 4: $\mathbf{k = 10}$ and $\mathbf{P(Y = 0) = 0.7}$

N	k	$P(Y = 0)$	α	Algorithm	N_s	$\widehat{bias^2}$	$\widehat{var}$	$\widehat{\bar{a}(\tilde{\theta})}$
10^5	10	0.7	-3.35	CC	33,400	1×10^{-5}	0.0029	-
				WCC	33,400	1×10^{-5}	0.0032	-
				LCC	16,700	3×10^{-6}	0.0021	0.1678
				Logit	10^5	4×10^{-7}	0.0003	-
10^4	10	0.7	-3.35	CC	3,280	0.0002	0.0303	-
				WCC	3,280	0.0004	0.0324	-
				LCC	1,640	1×10^{-5}	0.0228	0.1678
				Logit	10^4	1×10^{-5}	0.0031	-
5,000	10	0.7	-3.35	CC	1,600	0.0017	0.0635	-
				WCC	1,600	0.0017	0.0716	-
				LCC	800	0.0002	0.0531	0.1678
				Logit	5,000	7×10^{-5}	0.0063	-
2,000	10	0.7	-3.35	CC	640	0.0110	0.1783	-
				WCC	640	0.0134	0.1979	-
				LCC	320	0.0010	0.1616	0.1682
				Logit	2,000	0.0004	0.0165	-
1,500	10	0.7	-3.35	CC	480	0.0159	0.2335	-
				WCC	480	0.0174	0.2703	-
				LCC	240	0.0025	0.2335	0.1684
				Logit	1,500	0.0006	0.0208	-

Table 11: Simulation 4 - Results from variation in full population size N, holding $k = 10$ and $P(Y = 0) = 0.7$ constant (results from $1,000$ runs).

Simulation 4: $\mathbf{k = 10}$ and $\mathbf{P(Y = 0) = 0.8}$

N	k	$P(Y=0)$	α	Algorithm	N_s	$\widehat{bias^2}$	$\widehat{var}$	$\widehat{\bar{a}(\tilde{\theta})}$
10^5	10	0.8	-3.35	CC	27,600	4×10^{-6}	0.0033	-
				WCC	27,600	4×10^{-6}	0.0043	-
				LCC	13,800	3×10^{-6}	0.0027	0.1387
				Logit	10^5	2×10^{-7}	0.0004	-
10^4	10	0.8	-3.35	CC	2,720	0.0004	0.0350	-
				WCC	2,720	0.0010	0.0455	-
				LCC	1,360	4×10^{-5}	0.0297	0.1388
				Logit	10^4	4×10^{-5}	0.0039	-
5,000	10	0.8	-3.35	CC	1,340	0.0015	0.0711	-
				WCC	1,340	0.0039	0.0982	-
				LCC	670	0.0003	0.0648	0.1389
				Logit	5,000	9×10^{-5}	0.0079	-
2,000	10	0.8	-3.35	CC	540	0.0088	0.1989	-
				WCC	540	0.0230	0.2585	-
				LCC	270	0.0037	0.2158	0.1400
				Logit	2,000	0.0003	0.0184	-
1,500	10	0.8	-3.35	CC	400	0.0193	0.2749	-
				WCC	400	0.0564	0.3839	-
				LCC	200	0.0013	0.3389	0.1407
				Logit	1,500	0.0003	0.0265	-

Table 12: Simulation 4 - Results from variation in full population size N, holding $k = 10$ and $P(Y = 0) = 0.8$ constant (results from $1,000$ runs).

Simulation 4: $\mathbf{k = 10}$ and $\mathbf{P(Y = 0) = 0.9}$

N	k	$P(Y=0)$	α	Algorithm	N_s	$\widehat{bias}^2$	$\widehat{var}$	$\bar{a}(\tilde{\theta})$
10^5	10	0.9	-4.69	CC	18,000	8×10^{-6}	0.0044	-
				WCC	18,000	2×10^{-5}	0.0090	-
				LCC	9,000	3×10^{-6}	0.0042	0.0915

				Logit	10^5	2×10^{-7}	0.0006	-
10^4	10	0.9	-4.69	CC	1,800	0.0008	0.0458	-
				WCC	1,800	0.0063	0.0987	-
				LCC	900	0.0001	0.0499	0.0921
				Logit	10^4	7×10^{-5}	0.0062	-
5,000	10	0.9	-4.69	CC	900	0.0035	0.0939	-
				WCC	900	0.0239	0.2012	-
				LCC	450	0.0003	0.1184	0.0928
				Logit	5,000	0.0004	0.0114	-
2,000	10	0.9	-4.69	CC	360	0.0177	0.2664	-
				WCC	360	0.1673	0.6115	-
				LCC	180	0.0120	0.6192	0.0975
				Logit	2,000	0.0013	0.0305	-
1,500	10	0.9	-4.69	CC	270	0.0416	0.4036	-
				WCC	270	0.2702	0.8646	-
				LCC	130	4.0434	1079.1	0.1025
				Logit	1,500	0.0035	0.0407	-

Table 13: Simulation 4 - Results from variation in full population size N, holding $k = 10$ and $P(Y = 0) = 0.9$ constant (results from $1,000$ runs).

Simulation 4: $\mathbf{k = 10}$ and $\mathbf{P(Y = 0) = 0.95}$

N	k	$P(Y = 0)$	α	Algorithm	N_s	$\widehat{bias^2}$	$\widehat{var}$	$\widehat{\bar{a}(\tilde{\theta})}$
10^5	10	0.95	-5.44	CC	10,000	4×10^{-5}	0.0075	-
				WCC	10,000	0.0005	0.0244	-
				LCC	5,600	4×10^{-5}	0.0070	0.0563
				Logit	10^5	5×10^{-6}	0.0009	-
10^4	10	0.95	-3.35	CC	1,000	0.0018	0.0787	-
				WCC	1,000	0.0395	0.2546	-
				LCC	570	0.0002	0.0954	0.0580
				Logit	10^4	0.0002	0.0092	-

5,000	10	0.95	-3.35	CC	500	0.0109	0.1713	-
				WCC	500	0.1629	0.5959	-
				LCC	290	0.0020	0.3065	0.0610
				Logit	5,000	0.0005	0.0182	-
2,000	10	0.95	-3.35	CC	200	0.0703	0.5660	-
				WCC	200	1.2634	2.0068	-
				LCC	130	2×10^24	2×10^27	0.0728
				Logit	2,000	0.0033	0.0469	-
1,500	10	0.95	-3.35	CC	150	0.1440	0.8435	-
				WCC	150	2.4870	3.1536	-
				LCC	100	1×10^25	7×10^27	0.0829
				Logit	1,500	0.0060	0.0646	-

Table 14: Simulation 4 - Results from variation in full population size N, holding $k = 10$ and $P(Y = 0) = 0.95$ constant (results from $1,000$ runs).

9.2.3 Additional tables from data application

Statistic	N	Mean	St. Dev.	Min	Max
age	11,208	44.006	10.341	19	90
fnlwgt	11,208	188,398.000	102,492.100	13,769	1,226,583
education_num	11,208	11.599	2.367	1	16
capital_gain	11,208	3,991.792	14,616.540	0	99,999
capital_loss	11,208	193.487	592.642	0	3,683
hours_per_week	11,208	45.690	10.798	1	99
class_1	11,208	1.000	0.000	1	1

Table 15: Summary Statistics from Income dataset for individuals with an annual income of more than 50K.

Statistic	N	Mean	St. Dev.	Min	Max
age	34,014	36.749	13.565	17	90
fnlwgt	34,014	190,175.200	106,653.700	13,492	1,490,400
education_num	34,014	9.631	2.420	1	16
capital_gain	34,014	149.023	927.447	0	41,310
capital_loss	34,014	54.032	312.220	0	4,356
hours_per_week	34,014	39.372	11.974	1	99
class_0	34,014	1.000	0.000	1	1

Table 16: Summary Statistics from Income dataset for individuals with an annual income of 50K or less.

	m	$N_{s,CC}$	$N_{s,WCC}$	$N_{s,LCC}$
1	0.95	21,295	21,295	12,551
2	0.90	20,175	20,175	11,882
3	0.85	19,054	19,054	11,242
4	0.80	17,937	17,937	10,575
5	0.75	16,819	16,819	9,899
6	0.70	15,699	15,699	9,260
7	0.65	14,575	14,575	8,583
8	0.60	13,451	13,451	7,918
9	0.55	12,336	12,336	7,262
10	0.50	11,212	11,212	6,602
11	0.45	10,090	10,090	5,936
12	0.40	8,969	8,969	5,283

Table 17: Data application - Fixed subsample sizes used in the data application analysis, where the result for $N_{s,lcc}$ is taken from the average subsample size after 100 runs, and $N_{s,CC}$ and $N_{s,WCC}$ are the results of sampling all cases and the same amount of controls for the mth sample.

9.3 Additional R packages

List of additional packages used for the analysis and visualizations of results:

MASS (Venables and Ripley 2002), tidyr (Wickham, Vaughan, and Girlich 2023), dplyr (Wickham, François, et al. 2023), ggplot2 (Wickham 2016), ggpubr (Kassambara et al. 2020), Wes Anderson color palette (Ram and Wickham 2018), stargazer (Hlavac 2022) and table (Dahl et al. 2019).

Statement of authorship:

I hereby confirm that the work presented has been performed and interpreted solely by myself except for where I explicitly identified the contrary. I assure that this work has not been presented in any other form for the fulfillment of any other degree or qualification. Ideas taken from other works in letter and in spirit are identified in every single case.

Bonn, 14.06.2023 ______________________________

Carolina Alvarez