I, Zhiyuan Dong, hereby submit this original work as part of the requirements for the degree of:

Doctor of Philosophy

in Business Administration

It is entitled:

Three Essays in Quantitative Analysis

Student Signature: Zhiyuan Dong

This work and its defense approved by:

Committee Chair: Martin Levy, PhD

Yan Yu, PhD

Jeffrey Camm, PhD

James J. Cochran, PhD

Tim K. Keyes, PhD
Three Essays in Quantitative Analysis

A dissertation submitted to the

Graduate School

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Department of Quantitative Analysis and Operations Management

of the College of Business

by

Zhiyuan Dong

MBA, MSQA. University of Cincinnati

July 2010

Committee Chair: Martin S. Levy, Ph.D.
Abstract

Essay 1: We develop a matrix approach via an eigen analysis for comparing the variances of estimates of a population total based on the arc-weight, Horvitz-Thompson, and conventional methods under a many-to-many frame structure on a data free basis. A benefit of this representation is that it results in an improved method for computing the 1st order inclusion probabilities for the Horvitz-Thompson method, which is known to be computationally expensive to obtain in a many-to-many context. We also propose an improved method for characterizing the eigen-structure of the Arc-Weight method. A Mathematica based package that performs the corresponding analysis is developed to facilitate the computations.

Essay 2: We propose a constrained M-estimation approach in logistic regression models to improve prediction accuracy in situations where useful prior information is available in the form of linear inequality and equality constraints on the parameters. While our approach is intended for use in small to moderate sample size situations, it also has practical virtues in helping to avoid the issue of non-convergence of parameter estimates. We also supply large sample size properties of the estimators, i.e., consistency and asymptotic distributions under certain assumptions. We apply the constrained approach on a real corporate bankruptcy prediction data set by embedding prior information such as Moody and S&P credit ratings and demonstrate that the prediction results are promising. We also conduct a simulation study and obtain favorable prediction performance via several metrics over the traditional approach.

Essay 3: In this essay, we develop a sample average approach to estimate the optimal coverage probability in an optimal coverage problem. We demonstrate the statistical properties in estimation bias reduction of this approach and provide some asymptotic results. We report the results of an empirical study on our proposed approach.
Acknowledgements

The Ph.D. study in the QAOM Department at the University of Cincinnati greatly improved my capacity and strength. I firmly believe that this experience will benefit me for the rest of my life. At the stage of concluding the Ph.D. dissertation work, I feel grateful to a lot of individuals who helped me along the way.

First and foremost, my heartfelt gratitude goes to Prof. Martin Levy. His constant guidance, knowledge, encouragement and patience helped me achieving one milestone after another in the Ph.D. program. My sincere thanks must also go to Prof. Jeff Camm, not only because he is in my dissertation committee group and has offered excellent expertise, but also because he has introduced me to a number of assignments and projects which have shaped me fundamentally over the years.

I am grateful to Prof. Yan Yu. Her constantly valuable insights and meaningful critiques helped me bring myself to a higher standard. I also thank Prof. James Cochran for the inspiring discussions and interesting the topic he offered for my third dissertation essay. Dr. Tim Keyes supplied the project which evolves into the second dissertation essay in its current shape and offered industrial strength advices.

It is also important to mention the professors in the University of Cincinnati math department who have taught me most of the graduate level math courses: Prof. James Deddens, Prof. Siva Sivaganesan, Prof. Paul Horn, Prof. Magda Peligrad, Prof. Costel Peligrad, and Prof. Charles Groetsch. They all helped deeply shaping my mathematical sense in one way or another.

Last but not the least, I am deeply indebted to my mother, Qi Zhang, and my wife, Meng Zhou. Without their constant support and love, it is impossible for me to get to where I am now, even remotely.

Zhiyuan Dong

July 1st 2010, Cincinnati, USA
## Contents

1 A Matrix Approach for Comparing Estimates of a Population Total Under a Many-to-Many Frame Structure .......................... 1

1.1 Introduction ........................................................................ 1

1.2 Background and Notational Set-up ........................................ 6

   1.2.1 Many to many sampling frame ..................................... 6

   1.2.2 The Arc-weight estimator of the population total ............... 7

   1.2.3 The Horwitz-Thompson estimator for the population total .... 13

   1.2.4 More on the Horwitz-Thompson Method .......................... 17

   1.2.5 The Inclusion Weight Point of View ............................... 20

   1.2.6 A General Linear Weight Estimator ............................... 22

1.3 Relationship Between GWSM and GLWM ............................. 24

   1.3.1 Standardized Link Matrix $\tilde{\Theta}_{AB}$ .......................... 24

   1.3.2 A Numerical Example ................................................. 26

   1.3.3 An Example in Variance Comparison ............................... 27

   1.3.4 Inclusion Relationship ............................................... 27

1.4 Matrix Approach for Variance Calculation and Characterization 29

   1.4.1 Arc-Weight Method ..................................................... 30

   1.4.2 Horwitz-Thompson Method ......................................... 31

   1.4.3 General Linear Weight Method ...................................... 31

   1.4.4 Variance Dominance Study and Empirical Findings .......... 31
1.5 Variance Calculation Revisited ........................................ 33
1.6 A Mathematica Add-On Package ..................................... 37
1.7 Discussion, Conclusion, and Future Work .......................... 37

2 On Constrained Logistic Regression with an Application to Corporate
Bankruptcy Prediction ......................................................... 40

2.1 Introduction ............................................................... 40
2.2 The Constrained Logistic Regression Model ................. 45

2.2.1 The Logistic Regression Model ................................. 45
2.2.2 The Parameter Constraints ................................... 47
2.2.3 The Constrained Logistic Regression Estimation Problem ........................................ 48
2.2.4 Constraint Qualifications ...................................... 48

2.3 Computations ............................................................. 50

2.3.1 Numerical Algorithms ............................................ 50
2.3.2 Computing Platforms ............................................ 51
2.3.3 Discussion .......................................................... 52

2.4 Application to Corporate Bankruptcy Dataset .................. 54

2.4.1 Data Set from WRDS ............................................. 56
2.4.2 Parameter Constraints ............................................ 57
2.4.3 Time Window and Sample Selection Methods ............. 58
2.4.4 Results .............................................................. 60

2.5 Simulation ................................................................. 62

2.6 Large Sample Properties .............................................. 64

2.7 Conclusion ................................................................. 68

3 On the statistical characteristics of coverage optimization based on sample
mean approach ................................................................. 70

3.1 Introduction ............................................................... 70
List of Tables

1.1 different ways of choosing 3 frame units out of 5 ........................................ 11
1.2 the population total estimate based on each of the 10 sample of size 3 drawn as shown in Table 1.1 .......................................................... 13
1.3 the HT population total estimate based on each of the 10 sample of size 3 drawn as shown in Table 1.1 .......................................................... 16
1.4 population element values: original v.s. apportioned ................................. 34

2.1 Summary statistics of data we use. Total N = 11059 firms. ....................... 57
2.2 Empirical results of the WRDS Dataset. Testing sample size is 30 for all fitting sample sizes. Number of replications made is 1000 for all fitting sample sizes. Percentage $AUC_{con} > AUC_{mle}$ column shows the percentages that the constrained approach has larger area under the ROC curve than the MLE approach. ............................... 60
2.3 Empirical Results 1 of the Simulated Dataset. Each entry shows the percentage of times that the constrained approach has a higher area under the ROC curve result than the unconstrained approach. $n_1^*$ is the training sample size. $max(p_i)$ is the highest event probability in all the samples, both training sample and testing sample. 63
2.4 Empirical Results 2 of the Simulated Dataset. Each entry shows the percentage of times that the mean of $||\tilde{\beta}_{con}^n - \beta^*||_2$ for the constrained approach is smaller than the mean of $||\tilde{\beta}_{mle}^n - \beta^*||_2$ for the unconstrained approach. $n_1^*$ is the training sample size. $max(p_i)$ is the highest event probability in all the samples, both training sample and testing sample. 63
2.5 Empirical Results 3 of the Simulated Dataset. Each entry shows the percentage of times that the mean of \( \sum_{i=1}^{n_2} |\hat{p}_{i}^{\text{con}} - p_1^*| \) for the constrained approach is smaller than the mean of \( \sum_{i=1}^{n_2} |\hat{p}_{i}^{\text{mle}} - p_1^*| \) for the unconstrained approach. \( n_1^* \) is the training sample size. \( \text{max}(p_i) \) is the highest event probability in all the samples, both training sample and testing sample.
List of Figures

1.1 A taxonomy of frame structures .............................................. 8
1.2 a Complex Many-to-Many structure, adopted from Clough (2000), pp. 81 ... 9
1.3 the arc-weight estimate of population total corresponding to row one in Table 1.1 ................................................................. 12
1.4 the HT estimate of population total corresponding to row one in Table 1.1 . 16
1.5 details of computing $\pi_1$ ................................................................. 18
1.6 another way of expressing the AW estimate of population total corresponding to row one in Table 1.1 ...................................................... 20
1.7 original frame-population and the apportioned frame-population, referring to the Figure 1.2 example ......................................................... 34

2.1 Area Under ROC for the results in Table 2.2. Mean is the sample average of the area under the ROC curve computed for the testing sample; Std Dev is the standard deviation of the area under the ROC curve computed for the testing sample. Each boxplot column represents the sampling picture of the area under the ROC curve under combinations of training sample size, i.e. 50, 100, 150, 200, 250, 300, 350, and estimation methods, i.e. constrained (CON) or MLE. ...................... 61

C.1 $m = 1$ ......................................................................................... 100
C.2 $m = 5$ ......................................................................................... 101
C.3 $m = 10$ ......................................................................................... 101
C.4  m = 15 ................................................. 101
C.5  m = 20 ................................................. 102
C.6  m = 25 ................................................. 102
C.7  m = 30 ................................................. 102
C.8  m = 35 ................................................. 103
C.9  m = 40 ................................................. 103
C.10 m = 45 ................................................. 103
C.11 m = 50 ................................................. 104
C.12 m = 55 ................................................. 104
C.13 m = 60 ................................................. 104
C.14 m = 65 ................................................. 105
Chapter 1

A Matrix Approach for Comparing Estimates of a Population Total Under a Many-to-Many Frame Structure

1.1 Introduction

In classical sample survey settings, it is commonly assumed that there is a complete and perfect frame available. Completeness means that all of the population elements are included in the frame and perfectness refers to the fact that for each frame unit, there is one and only one population element associated with it.

In reality, there are situations where this ideal situation is violated. For example, suppose one needs to estimate the total household income of a particular city and we do not have a complete and perfect frame of households. Instead, we may have dual frames with one frame being the telephone number list of a particular city and the other frame being the areal list of the same city, assuming that the union of these two frames will cover the entire
set of households in the city. In essence, this is an example where there is no complete frame available. Hartley (1962) and Hartley (1974) discussed the method of sampling from both frames to achieve better efficiency compared to sampling directly from a single complete and perfect frame. Fuller and Burmeister (1972) discussed various estimators for samples selected from two overlapping frames thereby minimizing the variance among the class of linear unbiased estimators. Biemer (1984) gave a systematic treatment over methodology for optimal dual frame sample designs. Lepkowski and Groves (1986) developed an error model for dual frame surveys and a cost model that aims to capture the complexity of a full scale dual frame survey. Bankier (1986) extended the treatment where observations are weighted according to their inclusion probabilities to stratified random sampling and multiple frame surveys. Skinner and Rao (1996) introduced the pseudo-maximum likelihood estimator in handling dual frame surveys. The asymptotic properties of this pseudo-maximum likelihood estimator are also discussed. Lohr and Rao (2000) discussed various estimators in dual frame surveys and compared the performance in a unified setup. They also proposed a jackknife-estimator, established the consistency and explored its finite sample properties. Lohr and Rao (2006) extended their treatment to multiple frame surveys. They derived and studied an optimal linear estimator and a pseudo-maximum likelihood estimator.

Another example, a variant of the above, immediately leads to a situation where there is no perfect frame. Imagine that for some instance of the areal list, we want to take samples to estimate the total household income. In some cases, one address corresponds to one household. In other cases, one address corresponds to a group of households, e.g. a building complex. Then, multiplicity comes into play. That is, for that one frame unit, i.e. the address, there are multiple population elements, i.e. households in the building complex, associated with it. Multiplicity in this context is referred to as duplication by some authors. For treatment of sampling with multiplicity, see Rao (1968), Sirken (1970), Sirken (1972a), Sirken (1972b), and others. In Rao (1968), the Hansen-Hurwitz estimator, an estimator similar to the Horwitz-Thompson (HT) estimator, was used to solve the problem
of multiplicity. Sirken (1970), Sirken (1972a) and Sirken (1972b) proposed an estimator similar to the arc-weight (AW) estimator later independently proposed by Byczkowski et al. (1998). Sirken (1972b) also extended the treatment to the stratified random sampling case.

A third example with increased complexity is as follows. Imagine that our primary point of interest is the total size of RAM installed in all of the computers in a building in terms of giga-bytes. If we have a complete list of computers installed in the building, then we could directly add the individual RAMs to get the exact answer. However, suppose instead, what we have is only a user-computer list that associates all the computers within the building to the corresponding persons within the building, assuming that all the computers are used by someone working within the building. If the list is not physically available, then we can draw a sample from this virtual complete list by interviewing a selected group of people. Now, multiplicity comes into play. That is, some person uses more than one computer and some computer is used by more than one person. So, when we collect data from the interviews, the question is how to handle the multiplicity in order to get an estimate with nice statistical properties such as unbiasedness, efficiency, etc.

In a complex building characteristics survey that motivated the research in Byczkowski et al. (1998), Byczkowski and Levy (2005) and Byczkowski and Levy (2009), where there was one complete but imperfect frame, multiplicities in both the frame units (building addresses corresponding to a list of electrical meters) and the population elements (actual buildings) were present. That survey was the genesis of the so-called AW estimator for the population total designed to handle both kinds of multiplicities in the context of simple and stratified random sampling. In those papers, a taxonomy for various many-to-many structures was introduced. Unbiasedness of the AW estimator was also established under some sampling designs. They also supplied the variance formulas for the estimated population total for the designs they discussed. They focused primarily on a so-called simple many-to-many case. To compare the performance of the AW estimator with the perfect frame estimator, an empirical study was conducted by means of a designed simulation experiment. One of the
factors was the size of the correlation between the population element values and the arc-
weights which was varied at different levels to investigate the performance change in terms of
relative efficiency between the AW estimator and the perfect frame estimator. The amount
and types of frame imperfections were other factors studied. They concluded that under
some correlation profiles and levels of the other factors, the AW estimator is preferred, while
under other correlation profiles and levels of the other factors, the perfect frame estimator is
preferred. The response variable in these experiments was the variance ratio of the AW and
perfect frame estimators. Recommendations were then outlined for practitioners who might
know or at least speculate on the levels of these factors present in their particular study.

In the messy frame literature, Deville and Lavellee (2006) proposed Indirect Sampling
and the Generalized Weight Share Method (GWSM) to deal with complexities in the many-
to-many settings. There are similarities between the Generalized Weight Share Method and
our proposed approach, the General Linear Weight Method (GLWM), as will become clear
later in this paper. We will compare these two approaches in later sections.

Clough (2000) extended the empirical study reported in Byczkowski and Levy (2005)
and Byczkowski and Levy (2009) to the complex many-to-many case and considered the HT
estimator as well as the AW and perfect frame methods for simple and stratified random
sampling designs. A pairwise three-way comparison was made among these three estimators
based on variance ratios. Clough (2000) found that even slight changes in the so-called com-
plex many-to-many case could reverse his previously observed empirical comparison results.
That is, the addition or deletion of even a single arc linking a frame unit to a population
element entailed a different recommendation of estimation method.

Not all many-to-many estimator efficiency research findings employ a variance ratio.
Another surprisingly revealing indicator of variance domination is the difference score i.e.,
the difference between the variance of estimator variances, such as the AW and the perfect
frame estimators. Byczkowski and Levy (2009) extended their earlier research based on a
variance ratio by a careful examination of the difference score. They discovered that when
properly represented, the difference score provided a coherent and interpretable basis for
deciding between the uses of the AW and conventional estimation methods. They also found
that these difference score results led to very similar recommendations based on their earlier
variance ratio results for planning a sample survey regarding the estimator to use in practice.

Primarily motivated by the startling findings in Clough (2000), together with the utility
of the difference score in Byczkowski and Levy (2009), we embarked on this current line of
inquiry. For the moment, we will limit our presentation to the case of simple random sampling
without replacement (SRSWOR) in order to facilitate an introduction to our matrix based
approach. In subsequent research, we will report on extensions of this work to the case of
stratified random sampling. Also, we limit discussion to the HT and AW methods since
both can be used in an imperfect frame setting, whereas use of the conventional estimator
of the population total entails a perfect frame, so its comparisons to HT and AW must be
dependent upon the nature of the data. Here, we will:

1. Introduce a matrix based formulation of the AW and HT estimators for the population
total. This formulation will facilitate the variance comparison or characterization process
via a difference score (as in Byczkowski and Levy (2009)) and more importantly makes it
independent of data. In the presence of multiplicities, the HT estimator is known to be
expensive to calculate since it requires calculation of the first order inclusion probabilities.
We will present an improved way of computing the first order inclusion probabilities. The
advantage of this improved calculation method is that it will facilitates increases in the frame
and population sizes of the problems in any researcher’s efforts to compare the variance of
the estimated population total when using the AW and the HT estimators (or others) on a
data free basis.

2. Demonstrate the connection between the AW and the HT estimators via a so-called
inclusion weight methodology.

3. Introduce a general linear weight family of estimators for the population total, special
cases of which yield the AW, HT, and conventional estimators.
4. Introduce an improved method of making the variance comparisons and/or characterizations under a data free setting using moderately large sample size.

5. Introduce our algorithms implemented by a Mathematica based package that perform all of the aforementioned calculations and characterizations, and an associated program that graphically generates all frame-population scenarios of moderate sizes having the whole panoply of many-to-many structures.

Our approach here is to present the principles underlying our algorithms using simple, easily digestible examples rather than generally stating results in theorematic generalities.

1.2 Background and Notational Set-up

1.2.1 Many to many sampling frame

Let $N_F = \text{the number of frame units}, N_F \in \mathbb{N}$; $n = \text{the number of frame units sampled}, n \in \mathbb{N}; N_P = \text{the number of population elements}, N_P \in \mathbb{N}$. Let $I_U$ be an index set associated with frame units, $I_U \subset \mathbb{N}, I_U \neq \emptyset, N_F = |I_U| < \infty$, and wolog assume that $I_U$ consists of consecutive integers and always starts with 1. In a similar fashion, define $I_E$ be an index set associated with the population elements, $I_E \subset \mathbb{N}, I_E \neq \emptyset, N_P = |I_E| < \infty$, and wolog assume that $I_E$ consists of consecutive integers and always starts with 1. Let $\Omega_U = \{U_i | i \in I_U\}$ be the set of all frame units; $\Omega_E = \{E_j | j \in I_E\}$ be the set of all population elements.

Under usual sampling settings, $I_U = I_E$. Sampling one frame unit $U_i, i \in I_U$, can be defined using a bijective choice function: $f: \Omega_U \mapsto \Omega_E, f(U_i) = E_i, f^{-1}(E_i) = U_i$. Under many-to-many settings, $I_U \neq I_E$ in general. In that case, sampling one frame unit $U_i, i \in I_U$, can be defined using a surjective choice function $g: \Omega_U \mapsto \Omega_E$.

$$g(U_i) = \bigcup_j \{E_j | E_j \text{ is associated with } U_i\} \quad (1.1)$$

$$g^{-1}(E_j) = \bigcup_i \{U_i | U_i \text{ is associated with } E_j\} \quad (1.2)$$
\[|g(U_i)| \geq 1 \text{ and } |g^{-1}(E_j)| \geq 1 \text{ for all } i \in I_U \text{ and } j \in I_E.\]

This characterization subsumes the usual sampling settings and excludes scenarios where \(|g(U_i)| = 0\) or \(|g^{-1}(E_j)| = 0\) for some \(i \in I_U\) or \(j \in I_E\). A taxonomy of various frame-population structures that we will refer to from time to time in this paper can be seen from Figure 2.1 adapted from Byczkowski and Levy (2005).

Let \(y^j \in \mathbb{R}\) be a value attached to \(E_j, j \in I_E\). We are interested in estimating the population total, \(Y = \sum_{j \in I_E} y^j\).

### 1.2.2 The Arc-weight estimator of the population total

According to Byczkowski and Levy (2005), the AW estimator of the population total is,

\[
\hat{Y}_{AW} = \frac{N_E}{n} \sum_{j=1}^{n} y^j s^j_{AW}
\]

where \(s^j_{AW} = \sum_{i: U_i \in g^{-1}(E_j)} s^{ji}\), \(s^{ji}\) is the conditional probability that a randomization (described in Byczkowski and Levy (2005)) leads to the choice of population element \(j\) given that frame unit \(i\) has been selected. These conditional probabilities were called arc probabilities in that paper.

Throughout this paper, we assume that one has a complete frame from which to draw samples in order to estimate the population total. What makes the estimation difficult are the multiplicities concurrently occurring in both the frame and the population, e.g. simple many-to-many and complex many-to-many cases as shown in Figure 2.1.

Consider the frame population scenario shown in Figure 1.2, an example of a complex many-to-many case. The frame units are \(\Omega_U = \{U_1, U_2, U_3, U_4, U_5\}\), the population elements are \(\Omega_E = \{E_1, E_2, E_3, E_4, E_5\}\), and the adjacency matrix is \(A\) in (1.35), which we now define. Let \(a_{ij} \in \{0, 1\}\) be the \((i, j)\)th entry of \(A\). Define \(a_{ij} = 1\) if \(U_i\) is associated with \(E_j\); \(a_{ij} = 0\) if \(U_i\) is NOT associated with \(E_j\). By construction, rows of \(A\) correspond to elements in \(\Omega_U\) and columns of \(A\) correspond to elements in \(\Omega_E\). It is easy to see that the adjacency matrix
A uniquely determines the frame population scenario and vice versa. To avoid isolated frame units or population elements, we restrict $A$ to be such that $\sum_j a_{ij} \geq 1$ for all $i$ and $\sum_i a_{ij} \geq 1$ for all $j$.

In Figure 1.2, the number of frame units is $N_F = 5$; the number of population elements is $N_P = 5$. Let the sample size drawn via SRSWOR from $\Omega_U$ be $n = 3$. Then, the total number of different ways of choosing 3 frame units out of 5 is $\binom{N_F}{n} = \binom{5}{3} = 10$. Let $M$ be the row stochastic incidence matrix, which results from the adjacency matrix $A$ derived by dividing each row by its corresponding row sum. That is, the $(i,j)$th element of $M$ is given by $m_{ij} = a_{ij}/\sum_j a_{ij}$. By this definition, it follows that:

$$\sum_i \sum_j m_{ij} = N_F$$

(1.4)

In (1.35), we display $A$ and $M$ corresponding to Figure 1.2. In that figure, note that $N_F = N_P$, which makes both $A$ and $M$ square matrices. But, in general, $N_F$ does not need to be equal to $N_P$, so that $A$ and $M$ may not be square matrices. Our treatment in this paper still holds if we have non-square matrices.
Using the above matrix notation, let \( w_{AW} \) be the column vector of the arc-weights of population elements. The \( j \)th coordinate of \( w_{AW} \) is denoted by \( w_{jAW} \), \( j \in I_E \). We have,

\[
w_{AW}^t = \mathbf{1}_{N_F}^t M
\]

(1.6)

where \( \mathbf{1}_{N_F} \) denotes a row vector of ones having size \( N_F \). The symbol \((\cdot)^t\) denotes the matrix transpose operation. It follows immediately that \( \mathbf{1}_{N_F}^t w_{AW} = \mathbf{1}_{N_F}^t M \mathbf{1}_{N_F} = N_F \) by (1.4).

With respect to Figure 1.2, we have \( w_{AW}^t = [1, 1/2, 1/2, 5/2, 1/2] \) by (1.6).

Denote by

\[
\Omega_{U,n} = \{ \{U_{i_1}, U_{i_2}, ..., U_{i_n}\} \mid U_{i_k} \in \Omega_U, i_k \in I_U, i_k \text{ distinct} \}
\]

(1.7)

all possible samples of size \( n \) from the frame units. Then, \( |\Omega_{U,n}| = \binom{N_F}{n} \). The \( \xi \)th element
of $\Omega_{U,n}$ is denoted by

$$\Omega_{U,n}^\xi = \{U_{i_1}^\xi, U_{i_2}^\xi, \ldots, U_{i_n}^\xi\}$$

for $1 \leq \xi \leq \binom{N_F}{n}$. We say $\Omega_{U,n}^{\xi_1} = \Omega_{U,n}^{\xi_2}$, if $\{U_{i_1}^{\xi_1}, U_{i_2}^{\xi_1}, \ldots, U_{i_n}^{\xi_1}\} = \{U_{i_1}^{\xi_2}, U_{i_2}^{\xi_2}, \ldots, U_{i_n}^{\xi_2}\}$. In other words, order is irrelevant as long as the two sets contain the same frame units. For any particular ordering of the set $\Omega_{U,n}^\xi$, the $\zeta$th element of $\Omega_{U,n}^\xi$ is denoted by $\Omega_{U,n}^{\xi,\zeta}$, $1 \leq \zeta \leq n$.

Define the combinatorial instance matrix, $C_{N_F,n}$ to be a matrix of $(N_F^n)$ rows and $N_F$ columns with its $\xi$th row, denoted by $C_{N_F,n}^\xi$, $1 \leq \xi \leq \binom{N_F}{n}$,

$$C_{N_F,n}^\xi = \begin{bmatrix} I(U_1 \in \Omega_{U,n}^\xi) & I(U_2 \in \Omega_{U,n}^\xi) & \ldots & I(U_{N_F} \in \Omega_{U,n}^\xi) \end{bmatrix} \quad (1.8)$$

where $I(\cdot)$ is the indicator function. On the population elements side, for each $\Omega_{U,n}^\xi \in \Omega_{U,n}$, define

$$\Omega_{E,n}^\xi = \{\{E_{j_1}, E_{j_2}, \ldots, E_{j_n}\} \mid \Omega_{U,n}^{\xi,k} \in g^{-1}(E_{j_k}), \ 1 \leq k \leq n\} \quad (1.9)$$

Denote the $\nu$th subset of $\Omega_{E,n}^\xi$ by $\Omega_{E,n}^{\xi,\nu}$. Again, order is irrelevant in the set $\Omega_{E,n}^\xi$ and further within its subsets $\Omega_{E,n}^{\xi,\nu}$. Note that, $\Omega_{E,n}^\xi$ consists of distinct elements $U_{i_k}$. However, in $\Omega_{E,n}^{\xi,\nu}$, $E_{j_k}$ may not be distinct. This can also be seen in the Table 1.1 example.

Before we estimate the population total using the AW estimator, we need to go through an enumeration process as in Byczkowski and Levy (2005). That is, for each sample $\Omega_{U,n}^\xi$, we need to enumerate $\Omega_{E,n}^\xi$ defined by (1.9). This is illustrated in Table 1.1. The column labeled Frame Units explicitly lists the set $\Omega_{U,n}$. Each row represents a sample, $\Omega_{U,n}^\xi$. The set $\Omega_{E,n}^\xi$ is enumerated under the column labeled Population Elements. The column labeled Individual Inclusion Probability is defined to be $1/|\Omega_{E,n}^\xi|$.

Define a function $\lambda_{AW}(\Omega_{E,n}^\xi, E_k)$ to be the number of times that $E_k$ appears in the set $\Omega_{E,n}^\xi$ under the AW method. Let $\Omega_{E,n}^{\xi,\nu,\zeta}$ denote the $\zeta$th element of $\Omega_{E,n}^{\xi,\nu}$.

$$\lambda_{AW}(\Omega_{E,n}^\xi, E_k) = \sum_\nu \sum_\zeta I(\Omega_{E,n}^{\xi,\nu,\zeta} = E_k) \quad (1.10)$$
Table 1.1: different ways of choosing 3 frame units out of 5

<table>
<thead>
<tr>
<th>Frame Units</th>
<th>Population Elements and Individual Inclusion Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$U_1$, $E_1$, $E_4$, $E_3$, $E_2$, $E_4$</td>
</tr>
<tr>
<td>2</td>
<td>$U_1$, $E_1$, $E_4$, $E_3$, $E_1$, $E_4$, $E_3$, $E_2$, $E_4$</td>
</tr>
<tr>
<td>3</td>
<td>$U_1$, $E_1$, $E_4$, $E_1$, $E_1$, $E_4$, $E_1$, $E_2$, $E_5$</td>
</tr>
<tr>
<td>4</td>
<td>$U_1$, $E_1$, $E_4$, $E_3$, $E_4$, $E_3$, $E_4$, $E_3$, $E_5$</td>
</tr>
<tr>
<td>5</td>
<td>$U_1$, $E_1$, $E_4$, $E_1$, $E_1$, $E_4$, $E_1$, $E_2$, $E_5$</td>
</tr>
<tr>
<td>6</td>
<td>$U_1$, $E_1$, $E_4$, $E_1$, $E_1$, $E_4$, $E_1$, $E_2$, $E_5$</td>
</tr>
<tr>
<td>7</td>
<td>$U_1$, $E_1$, $E_4$, $E_1$, $E_1$, $E_4$, $E_1$, $E_2$, $E_5$</td>
</tr>
<tr>
<td>8</td>
<td>$U_1$, $E_1$, $E_4$, $E_1$, $E_1$, $E_4$, $E_1$, $E_2$, $E_5$</td>
</tr>
<tr>
<td>9</td>
<td>$U_1$, $E_1$, $E_4$, $E_1$, $E_1$, $E_4$, $E_1$, $E_2$, $E_5$</td>
</tr>
<tr>
<td>10</td>
<td>$U_1$, $E_1$, $E_4$, $E_1$, $E_1$, $E_4$, $E_1$, $E_2$, $E_5$</td>
</tr>
</tbody>
</table>

where $I(\cdot)$ is the indicator function. Later we will define a similar function $\lambda_{HT}(\Omega_{k,n}^\xi, E_k)$ under the HT method. Given a sample of $n$ frame units $\Omega_{U,n}^\xi$, the population total estimate under AW method is:

$$\hat{Y}_{AW}(\Omega_{U,n}^\xi) = \frac{1}{|\Omega_{E,n}^\xi|} \sum_{k \in I_E} \lambda_{AW}(\Omega_{E,n}^\xi, E_k)y_k N_F \frac{n}{n}$$ (1.11)

It can be shown that (1.11) is equivalent to (1.3). As an example of (1.11), the estimate of the population total using the AW method based on the sample $\{U_1, U_2, U_3\}$ is obtained as shown in Figure 1.3.

Denote $y_{AW}' = [y^1_{AW} / w^1_{AW}, y^2_{AW} / w^2_{AW}, ..., y^{N_F}_{AW} / w^{N_F}_{AW}]$. Using matrix notation, we can rewrite (1.11) as follows:

$$\hat{Y}_{AW}(\Omega_{U,n}^\xi) = C^\xi_{N_F,n} M y_{AW} \frac{N_F}{n}$$ (1.12)

In Figure 1.2, the estimate of the population total using the AW method based on the sample
Figure 1.3: the arc-weight estimate of population total corresponding to row one in Table 1.1

\{U_1, U_2, U_3\} can be obtained as follows by (1.12):

\[
\hat{Y}_{AW}(\{U_1, U_2, U_3\}) = \begin{bmatrix}
1/2 & 0 & 1/2 & 0 & 0 \\
1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1/2 & 1/2 \\
\end{bmatrix} \begin{bmatrix}
y^1/1 \\
y^2/(1/2) \\
y^3/(1/2) \\
y^4/(5/2) \\
y^5/(1/2) \\
\end{bmatrix}
\]

(1.13)

By following (1.12), in a similar fashion, we obtain the estimated population total for these ten samples as shown in Table 1.2 after expanding matrix products as in (1.13).

By adding these 10 individual estimates and dividing by 10, it is easy to see that the arithmetic average of these ten samples yields an unbiased estimate of the population total. That is, the average value of the estimated population total is \(\sum_{i=1}^{5} y^i\). Note that, Byczkowski and Levy (2005) showed that the AW method (1.3) yields an unbiased estimator of the population total. Due to the equivalence of (1.3), (1.11) and (1.12), it is easily seen that the matrix approach (1.12) also produces an unbiased estimator of the population total. Here we offer a proof that demonstrates the power of a matrix approach.

**Theorem 1.** \(\hat{Y}_{AW}(\Omega^\xi_{U,n}) = C^\xi_{N_F,n} M y_{AW} \frac{N_F}{n}\) is a unbiased estimator of the population total \(Y = \sum_{j \in t_E} y^j\)
Table 1.2: the population total estimate based on each of the 10 sample of size 3 drawn as shown in Table 1.1

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5/3(y₁ + y² + y³ + 2y⁴)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5/3(y₁ + y² + y³ + 2y⁴)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5/3(y₁ + y² + y³ + 4/3y⁴ + y⁵)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1/6(5y₁ + 10y³ + 8y⁴)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5/3(1/2y₁ + y³ + 3/2y⁴ + y⁵)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5/3(1/2y₁ + y³ + 3/2y⁴ + y⁵)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1/6(5y₁ + 10y³ + 8y⁴)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5/3(1/2y₁ + y² + 3/2y⁴ + y⁵)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>5/3(1/2y₁ + y² + 3/2y⁴ + y⁵)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5/3(y⁴ + y⁵)</td>
<td></td>
</tr>
</tbody>
</table>

Proof.

\[
\frac{1}{|Ω_{U,n}|} \sum_ξ \tilde{Y}_{AW}(Ω_ξ^{Ω_{U,n}}) = \frac{1}{|Ω_{U,n}|} \sum_ξ C_ξ^{Ω_{U,n}} M y_{AW} \frac{N_F}{n} \\
= \frac{1}{\binom{N_F}{n}} \left( \binom{N_F - 1}{n - 1} \right) 1_{N_F}^t M y_{AW} \frac{N_F}{n} \\
= \frac{n}{N_F} w_{AW}^t y_{AW} \frac{N_F}{n} \\
= \sum_{j \in I_E} y^j
\]

This completes the proof.

1.2.3 The Horwitz-Thompson estimator for the population total

According to [Cochran (1977)] pp. 259, the HT estimator of the population total, denoted by \( \hat{Y}_{HT} \), is,

\[
\hat{Y}_{HT} = \sum_{j=1}^{n} \frac{y^j}{\pi_j}
\]

where \( \pi_j \) is the 1st order probability, i.e., the probability that the \( j \)th unit is in the sample of size \( n \). We will explain precisely what \( \pi_j \) means in the many-to-many settings via a constructive argument in this section.
In order to apply the HT method in estimation, we need to calculate the 1st order inclusion probability $\pi_j$ for each population element $E_j$, $j \in I_E$. In the literature, the $\pi_j$ are known to be computationally expensive to obtain. For a detailed calculation of the 1st order inclusion probability in the many-to-many setting, see Clough (2000), pp. 66, Example 5-6, pp. 72, Example 5-7, pp. 74, Example 5-8 and pp. 79, Example 5-9. The calculated 1st order inclusion probabilities of the five population elements, $\{E_1, E_2, E_3, E_4, E_5\}$, in Figure 1.2 are: $\{21/40, 3/10, 3/10, 19/20, 3/10\}$, respectively. We will show how to get these probabilities later in this section. We will also demonstrate a more computationally efficient approach.

Denote by the vector $\pi^t = [\pi_1, \pi_2, ..., \pi_{N_P}]$ the 1st order inclusion probabilities of population elements $E_1, E_2, ..., E_{N_P}$. Note that, $\{\pi_j\}_{j=1}^{N_P}$ does not sum to $n$. This generally cannot happen in the standard complete perfect frame case. In fact, as shown in Cochran (1977), pp. 259, (9A.36), using SRSWOR, $\sum_{j=1}^{N_P} \pi_j = n$.

Define a function $\lambda_{HT}(\Omega_{E,n}^\xi, E_k)$ to be the number of times that $E_k$ appears in the set $\Omega_{E,n}^\xi$ under the HT method, i.e.,

$$\lambda_{HT}(\Omega_{E,n}^\xi, E_k) = \sum_\nu \mathbb{I}(E_k \in \Omega_{E,n}^\xi, \nu) \tag{1.15}$$

where $\mathbb{I}(\cdot)$ is the indicator function. When compared to (1.10), (1.15) performs duplicate correction. That is, if $E_k$ appears in $\Omega_{E,n}^\xi, \nu$ multiple times, (1.10) will count it as many times as it appears while (1.15) will just count it once. From an algorithm implementation point of view, duplicate correction is a major disadvantage. The 1st order inclusion probability $\pi_k$ for population element $E_k$ can be computed as:

$$\pi_k = \frac{1}{|\Omega_{U,n}|} \sum_{\xi} \frac{\lambda_{HT}(\Omega_{E,n}^\xi, E_k)}{|\Omega_{E,n}^\xi|} \tag{1.16}$$

In (1.16), $\lambda_{HT}(\Omega_{E,n}^\xi, E_k)/|\Omega_{E,n}^\xi|$ is the probability that $E_k$ is included in the sample given that the sample $\Omega_{U,n}^\xi$ is taken. So, $\pi_k$ is just the arithmetic average of these inclusion probabilities over the sample space $\Omega_{U,n}$. 


The steps behind (1.16) are illustrated in Figure 1.2. Here, population element $E_1$ appears a total of 4 times in the sample $\{U_1, U_2, U_3\}$ if we apply (1.10). This fact can be easily seen from the first row in Table 1.1 by counting how many times $E_1$ appears. However, by applying (1.15), we have the number of times $= 3$. By applying (1.16), we have:

$$
\pi_1 = \frac{1}{10}(\frac{3}{4} + \frac{3}{4} + \frac{6}{8} + \frac{1}{2} + \frac{2}{4} + \frac{2}{4} + \frac{2}{4} + \frac{2}{4} + 0) = \frac{21}{40} \quad (1.17)
$$

In a similar fashion, we can obtain $\pi^t = [21/40, 3/10, 3/10, 19/20, 3/10]$. One can easily note that since $1^t_{N_P} \pi = 19/8 \neq 3 = n$, in many-to-many settings, $\{\pi_j\}_{j=1}^{N_P}$ does not generally sum to $n$. This is in contrast to the fact in the AW method we have $1^t_{N_P} w_{AW} = N_F$. Let $w_{HT} = \pi$ whose coordinates are defined in (1.16). Denote $y_{HT}^t = [y^1/w_{HT}, y^2/w_{HT}, ..., y^{N_P}/w_{HT}]^T$. Then, in a similar fashion to (1.11), for the HT method, we have:

$$
\hat{Y}_{HT}(\Omega_{U,n}^\xi) = \frac{1}{|\Omega_{E,n}^\xi|} \sum_{k \in l_E} \frac{\lambda_{HT}(\Omega_{E,n}^\xi, E_k)y^k}{w^k_{HT}} \quad (1.18)
$$

Let $\Gamma^t_{\xi} = \frac{1}{|\Omega_{E,n}^\xi|}[\lambda_{HT}(\Omega_{E,n}^\xi, E_1), \lambda_{HT}(\Omega_{E,n}^\xi, E_2), ..., \lambda_{HT}(\Omega_{E,n}^\xi, E_{N_P})]$. This is the inclusion probability of population elements given that sample $\Omega_{U,n}^\xi$ is taken. Then (1.18) can be written

$$
\hat{Y}_{HT}(\Omega_{U,n}^\xi) = \Gamma^t_{\xi} y_{HT} \quad (1.19)
$$

Continuing with the example as shown in Figure 1.2, when given the sample $\{U_1, U_2, U_3\}$, $\hat{Y}_{HT}(\{U_1, U_2, U_3\})$ can be obtained as shown in Figure 1.4. Similar to Table 1.2, we have Table 1.3.
Figure 1.4: the HT estimate of population total corresponding to row one in Table 1.1

Table 1.3: the HT population total estimate based on each of the 10 sample of size 3 drawn as shown in Table 1.1

<table>
<thead>
<tr>
<th></th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$10y_1/7 + 5y^2/3 + 5y^3/3 + 20y^4/19$</td>
</tr>
<tr>
<td>2</td>
<td>$10y_1/7 + 5y^2/3 + 5y^3/3 + 10y^4/19$</td>
</tr>
<tr>
<td>3</td>
<td>$10y_1/7 + 5/57(19y^2 + 19y^3 + 6y^4 + 19y^5)$</td>
</tr>
<tr>
<td>4</td>
<td>$5/399(76y_1 + 133y^3 + 84y^4)$</td>
</tr>
<tr>
<td>5</td>
<td>$5/399(76y^1 + 133y^3 + 84y^4 + 133y^5)$</td>
</tr>
<tr>
<td>6</td>
<td>$5/399(76y^1 + 133y^3 + 84y^4 + 133y^5)$</td>
</tr>
<tr>
<td>7</td>
<td>$5/399(76y^1 + 133y^3 + 84y^4)$</td>
</tr>
<tr>
<td>8</td>
<td>$5/399(76y^1 + 133y^3 + 84y^4 + 133y^5)$</td>
</tr>
<tr>
<td>9</td>
<td>$5/399(76y^1 + 133y^3 + 84y^4 + 133y^5)$</td>
</tr>
<tr>
<td>10</td>
<td>$20y^4/19 + 5y^5/3$</td>
</tr>
</tbody>
</table>
1.2.4 More on the Horwitz-Thompson Method

In this section, we will discuss how to obtain the right-hand-side of (1.19) without doing a brute force duplicates correction. Define the function $h$:

$$h(U_i, E_k) = g(U_i) \setminus E_k$$

(1.20)

where $g(\cdot)$ is defined in Section 1.2.1. Then $\lambda_{HT}(\Omega_{E,n}^\xi, E_k)/|\Omega_{E,n}^\xi|$, the inclusion probability of $E_k$ given the sample $\Omega_{U,n}^\xi$, can be rewritten:

$$\lambda_{HT}(\Omega_{E,n}^\xi, E_k) / |\Omega_{E,n}^\xi| = \prod_{U_i \in \Omega_{U,n}^\xi} |g(U_i)| - \prod_{U_i \in \Omega_{U,n}^\xi} |h(U_i, E_k)| / \prod_{U_i \in \Omega_{U,n}^\xi} |g(U_i)|$$

(1.21)

To understand (1.21), one just needs to see that the numerators and the denominators on each side are equal to each other. Specifically, the equality of numerators follows from a complementary counting argument. That is, the number of $\Omega_{E,n}^\xi$ containing $E_k$ in $\Omega_{E,n}^\xi$ is equal to the total number of $\Omega_{E,n}^\xi$ in $\Omega_{E,n}^\xi$ minus the number of $\Omega_{E,n}^\xi$ that do not contain $E_k$ in $\Omega_{E,n}^\xi$. The advantage of (1.21) is that one does not need to enumerate $\Omega_{E,n}^\xi$ in order to evaluate $\lambda_{HT}(\Omega_{E,n}^\xi, E_k)$ on the left hand side. On the other hand, $|g(U_i)|$ and $|h(U_i, E_k)|$ can be obtained without enumerating $\Omega_{E,n}^\xi$:

$$|g(U_i)| = \text{ith coordinate of } A \mathbf{1}_{N_F}$$

(1.22)

$$|h(U_i, E_k)| = (i, k)\text{th element of } A (J_{N_P} - I_{N_P})$$

(1.23)

where $J_{N_P}$ is the square unit matrix consisting of 1s having size $N_P$, and $I_{N_P}$ is the identity matrix of size $N_P$. Using (1.21), (1.22) and (1.23) one can simplify the calculation of (1.16). We continue using the example in Table 1.1 to recompute $\pi$. Compute $A \mathbf{1}_{N_F}$ and $A (J_{N_P} -$
$$A \mathbf{1}_{N_p} = \begin{bmatrix} 2 \\ 2 \\ 1 \\ 1 \\ 2 \end{bmatrix} ; \ A(J_{N_p} - I_{N_p}) = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 & 2 & 2 \\ 1 & 1 & 2 & 2 & 2 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 2 & 2 & 2 & 1 & 1 \end{bmatrix}$$

(1.24)

The details involved in computing $\pi_1$ are shown in Figure 1.5. The left panel is equivalent to the sample space $\Omega_{U,n}$. Each row represents a $\Omega_{U,n}^\xi$ for some $\xi$. Each row in the central panel is computed using the right hand side of (1.21). Entries in the central panel of Figure 1.5 come from various coordinates of $A \mathbf{1}_{N_p}$ and $A(J_{N_p} - I_{N_p})$ as indicated by (1.22) and (1.23). Applying the same steps as above to other population elements obtains the $\pi$ vector. An interesting by-product obtains when computing the $\pi$ vector, namely a matrix, denoted...
by , whose element is just (1.21). That is,

\[ \Lambda_{\xi,k} = \frac{\lambda_{HT}(\Omega_{E,n}^\xi, E_k)}{|\Omega_{E,n}^\xi|} = \frac{\prod_{U_i \in \Omega_{U,n}^\xi} |g(U_i)| - \prod_{U_i \in \Omega_{U,n}^\xi} |h(U_i, E_k)|}{\prod_{U_i \in \Omega_{U,n}^\xi} |g(U_i)|} \]  

(1.25)

Note that, if we use \( \Lambda_{\cdot,k} \) to denote the \( k \)th column of \( \Lambda \), then \( \Lambda_{\cdot,k} \) contains the inclusion probability of \( E_k \) in the sample space for each \( \xi \), \( 1_{[\Omega_{U,n}]}^{t} \Lambda_{\cdot,k} / |\Omega_{U,n}| = \pi_k \). This is exactly (1.16). In the right panel of Figure 1.5, we have one example when \( k = 1 \). It is also interesting to note that, if we use \( \Lambda_{\xi,\cdot} \) to denote the \( \xi \)th row of \( \Lambda \), then \( \Lambda_{\xi,\cdot} \) is just \( \Gamma_{\xi}^t \) as in (1.19). One needs to realize the importance of having \( \Lambda \) before applying (1.19) since both \( \Gamma_{\xi}^t \) and \( y_{HT} \) rely on \( \Lambda \).

Having observed these facts, the unbiasedness of the HT estimator follow easily in matrix notation.

**Theorem 2.** \( \hat{Y}_{HT}(\Omega_{U,n}^\xi) = \Gamma_{\xi}^t y_{HT} \) is an unbiased estimator of the population total \( Y = \sum_{j \in I_E} y_j \).

**Proof.**

\[
\frac{1}{|\Omega_{U,n}|} \sum_{\xi} \hat{Y}_{HT}(\Omega_{U,n}^\xi) = \frac{1}{|\Omega_{U,n}|} \sum_{\xi} \Gamma_{\xi}^t y_{HT} \\
= \frac{1_{[\Omega_{U,n}]}^{t} \Lambda}{|\Omega_{U,n}|} y_{HT} \\
= w_{HT}^t y_{HT} \\
= \sum_{j \in I_E} y_j
\]

This completes the proof. \( \square \)

However, one needs to note that when \( |\Omega_{U,n}| \) grows, obtaining \( \Lambda \) becomes impractical. For example, when \( n = 50 \) and \( N_F = 100 \), \( |\Omega_{U,n}| = 100891344545564193334812497256 \), so it is burdensome to compute the entries of \( \Lambda \). When we choose 15 frame units out of 30, \( |\Omega_{U,n}| = 155117520 \), which still presents a computational challenge for \( \Lambda \).
1.2.5 The Inclusion Weight Point of View

To motivate the idea of the inclusion weight, refer to Figure 1.3, and distribute the fraction \( \frac{5}{3} \) on the left hand side into each term in the denominators yielding Figure 1.6. Let us call the resulting denominator terms inclusion weights. Denote the inclusion weight vector for the population elements by \( w_{IW} \), and note:

\[
\frac{4y^4}{5} + \frac{2y^3}{25} + \frac{2y^2}{25} + \frac{4y^1}{5} = \frac{5}{3} \left( y^1 + y^2 + y^3 + \frac{2}{5} y^4 \right)
\]

It is now easy to see that the difference between Figure 1.6 and Figure 1.4 is that they have different numerators and different denominators, although both are unbiased estimates of the population total.

From Section 1.2.3 and 1.2.4, we can see that it is computationally difficult to get a population total estimate under the HT method. On the other hand, it is easy to get an AW estimate because (1.12) and (1.6) are much easier to compute.

The equalities in (1.27) give the motivation for the name inclusion weights.

\[
w_{IW} = \frac{n}{N_F} \mathbf{w}_{AW}
\]

\[
w_{IW} = \frac{n}{N_F} \mathbf{1}^{t}_{N_F} \mathbf{M} = \left( \begin{array}{c} N_F - 1 \\ n - 1 \end{array} \right) \mathbf{1}^{t}_{N_F} \mathbf{M} \frac{1}{\binom{N_F}{n}} = \mathbf{1}^{t}_{\binom{N_F}{n}} \mathbf{C}_{N_F,n} \mathbf{M} \frac{1}{\binom{N_F}{n}}
\]
where $M$ is the row stochastic incidence matrix, defined in Section 1.2.2. One example of $M$ is given in (1.35). The combinatorial instance matrix, $C_{N_F,n}$ is defined in Section 1.2.2. One example of $C_{N_F,n}$ is shown in the left panel of Figure 1.5. The second equality in (1.27) holds since $n/N_F = \binom{N_F - 1}{N_F - n}/\binom{N_F}{n}$. The third equality in (1.27) holds since each frame units repeats $\binom{N_F - 1}{n-1}$ times in the matrix $C_{N_F,n}$ and therefore $\binom{N_F - 1}{n-1} \frac{1}{N_F} = \binom{N_F}{n} C_{N_F,n}$. The idea of inclusion weight comes from the right hand side in (1.27) since the resulting vector gives the average weight over the sample space of each population element included. One example of the resulting inclusion weight vector is shown in Figure 1.6. The denominators are the coordinates of the inclusion weight vector.

Now, it is easier to interpret the difference between Figure 1.6 and Figure 1.4. In Figure 1.6, corresponding to the AW method, the numerator coefficients are the number of times the population elements are included in the sample and the denominator coefficients are the weights of the population elements included in the sample space. In Figure 1.4, corresponding to the HT method, the numerator coefficients are the duplication corrected number of times the population elements are included in the sample and the denominator coefficients are the 1st order inclusion probabilities of the population elements in the sample space.

By (1.26) or (1.27), the ratio of any inclusion weight to the corresponding arc-weight is equal $n/N_F$. Note that the sum of the inclusion weights of the population elements is equal to the sample size $n$, since $\frac{1}{N_F} w_{IW} = \frac{1}{N_F} w_{AW} n/N_F = N_F \frac{n}{N_F} = n$. Recall that for the HT method, the sum of the 1st order inclusion probabilities does NOT equal the sample size $n$. Thus, in trying to produce computational savings using the HT method, the resulting estimator with those desired computational savings, turns out to be the AW estimator! It is now evident that the AW estimator is nothing but a simplified variant of the HT estimator. The simplification results because there is no longer a need to calculate the 1st order inclusion probabilities of the population elements.
1.2.6 A General Linear Weight Estimator

This section is motivated by observing that there exists a generalization to the AW and HT formulas in (1.12) and (1.19), respectively. In Section 1.2.2, the arc-weight for population elements $\Omega_E$ is given by (1.6). Note that another way of obtaining (1.6) is:

**Lemma 3.** $w^{t}_{AW} = \frac{N_F}{n} \frac{C_{N_F,n} M}{n} \frac{\frac{N_F}{n}}{\frac{n}{N_F} \frac{n}{N_F} - 1} = 1_{\Omega_U,n}^t C_{N_F,n} M \frac{N_F}{n} = 1_{N_F}^t M$

Proof.

$$\frac{1}{|\Omega_U,n|} 1_{|\Omega_U,n|}^t C_{N_F,n} M \frac{N_F}{n} = \frac{1}{(N_F)^n} \left( \frac{N_F - 1}{n - 1} \right) 1_{N_F}^t M \frac{N_F}{n} = 1_{N_F}^t M$$

The right hand side is (1.6). This completes the proof. \qed

Now, as part of the proof of Theorem 2, we have,

$$w^{t}_{HT} = \frac{1}{|\Omega_U,n|} \Lambda$$

(1.28)

Next, compare (1.28) to the equation in Lemma 3, and note that the counter-part of $\Lambda$ in (1.28) is $C_{N_F,n} M \frac{N_F}{n}$ for the AW method. Now, if one substitutes the symbol $\Lambda_{HT}$ for $\Lambda$ in (1.28) and one lets $\Lambda_{AW}$ denote $C_{N_F,n} M \frac{N_F}{n}$ in Lemma 3, then we can rewrite (1.12) and (1.19) as:

$$\hat{Y}_{AW}(\Omega_{U,n}^\xi) = \Lambda_{AW}^\xi \cdot y_{AW}$$

(1.29)

$$\hat{Y}_{HT}(\Omega_{U,n}^\xi) = \Lambda_{HT}^\xi \cdot y_{HT}$$

(1.30)

where $\Lambda_{AW}^\xi$ and $\Lambda_{HT}^\xi$ denote the $\xi$th rows in $\Lambda_{AW}$ and $\Lambda_{HT}$, respectively.

These observations permit us to construct a more general family of estimators that contains the AW and HT estimators as special cases. Let $\Lambda_{\psi}$ be a matrix having $|\Omega_U,n|$ rows and $N_F$ columns with real entries such that $\sum_i \Lambda_{\psi}^{ij} \neq 0$ for each $j$. The symbol $\psi$ is a placeholder for some estimation method and $\Lambda_{\psi}$ is $\psi$ specific. Let the weight vector for population
elements $\Omega_\psi$ under method $\psi$ be denoted by $w_\psi$ and define,

$$w^t_\psi = \frac{1_{[\Omega_{U,n}]}^t \Lambda_\psi}{|\Omega_{U,n}|}$$

(1.31)

whose $j$th coordinate is denoted by $w^j_\psi$. Further, define the $\psi$ - weighted values of the population elements to be:

$$y^j_\psi = \left[ y^1_\psi/w^1_\psi, y^2_\psi/w^2_\psi, ..., y^N_\psi/w^N_\psi \right]$$

(1.32)

Then define a general linear estimator of the population total $Y$ given a sample $\Omega_{U,n}^\xi$ as:

$$\hat{Y}_\psi(\Omega_{U,n}^\xi) = \Lambda_\psi \cdot y_\psi$$

(1.33)

**Theorem 4.** $\hat{Y}_\psi(\Omega_{U,n}^\xi) = \Lambda_\psi \cdot y_\psi$ is an unbiased estimate of the population total $Y = \sum_{j \in I_\psi} y^j$.

**Proof.**

$$\frac{1}{|\Omega_{U,n}|} \sum_\xi \hat{Y}_\psi(\Omega_{U,n}^\xi) = \frac{1}{|\Omega_{U,n}|} \sum_\xi \Lambda_\psi \cdot y_\psi$$

$$= \frac{1_{[\Omega_{U,n}]}^t \Lambda_\psi}{|\Omega_{U,n}|} y_\psi$$

$$= w^t_\psi y_\psi$$

$$= \sum_{j \in I_\psi} y^j$$

This completes the proof. $\Box$

It is easy to see that the family of such estimators that are linear in $y$, and denoted by $\Psi$ consists of unbiased estimators of the population total. Observe that the AW and HT estimators both belong to $\Psi$. Since $\Lambda_\psi^{ij} \in \mathbb{R}$, $\Psi$ is uncountable.
1.3 Relationship Between GWSM and GLWM

In this section, we investigate the relationship between the GWSM and GLWM approach. We will give examples that show the similarity and difference between GWSM and GLWM. We will conclude this section by showing that GWSM is a sub-case of GLWM.

1.3.1 Standardized Link Matrix $\tilde{\Theta}_{AB}$

The difference between GWSM and the AW Method in GLWM originates from the way that the standardized link matrix, $\tilde{\Theta}_{AB}$, is constructed. Before we get to $\tilde{\Theta}_{AB}$, here we focus on the following link matrix, $\Theta_{AB}$, quoted from Deville and Lavellee (2006), page 166.

$$
\Theta_{AB} = \begin{bmatrix}
\theta_{11}^{AB} & \theta_{12}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\
\theta_{21}^{AB} & \theta_{22}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \theta_{33}^{AB} & \theta_{34}^{AB} & 0 & 0 & 0 & 0 \\
0 & 0 & \theta_{43}^{AB} & \theta_{44}^{AB} & \theta_{45}^{AB} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \theta_{55}^{AB} & \theta_{56}^{AB} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \theta_{67}^{AB} & \theta_{68}^{AB}
\end{bmatrix}
$$

(1.34)

In (1.34), there are 6 rows, each representing a frame unit or an element from the population $A$. There are also 8 columns, each representing a population element or an element from the population $B$. We prefer our term, frame unit, since it makes an immediate distinction between the frame and the target population.

In the AW Method in GLWM, there is a counterpart to $\Theta_{AB}$. It is the adjacency matrix $A$, defined by (1.35). One might treat the adjacency matrix $A$ as a special case of (1.34) since $A$ has the form shown in (1.35).
\[ A = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
\end{bmatrix} \]

(1.35)

However, in order to demonstrate the difference between the GSWM and the AW Method in GLWM, we choose to retain the \( \theta_{AB} \) in place of the 1s in the adjacency matrix \( A \). The reason is that maintaining symbolic terms avoids a situation in which the two results are numerically similar but not exactly the same, so we can lose track of the reason why they are different. To show the dependence, we use \( A_\theta \) instead. Thus, we have the correspondence (1.36).

\[ A_\theta = \begin{bmatrix}
\theta_{11}^{AB} & \theta_{12}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\
\theta_{21}^{AB} & \theta_{22}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \theta_{33}^{AB} & \theta_{34}^{AB} & 0 & 0 & 0 & 0 \\
0 & 0 & \theta_{43}^{AB} & \theta_{44}^{AB} & \theta_{45}^{AB} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \theta_{55}^{AB} & \theta_{56}^{AB} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \theta_{67}^{AB} & \theta_{68}^{AB} \\
\end{bmatrix} = \Theta_{AB} \]  

(1.36)

In GWSM, the standardized link matrix is then constructed as in (1.37)

\[ \tilde{\Theta}_{AB} = \Theta_{AB} \left[ \text{diag}(1_A^t \Theta_{AB}) \right]^{-1} \]  

(1.37)

Our counterpart to \( \tilde{\Theta}_{AB} \) is the row stochastic matrix, \( M \), which was introduced in equation (1.35). Using symbolic terms, our \( M_\theta \) matrix is,

\[ M_\theta = \left[ \text{diag}(A_\theta 1_B) \right]^{-1} A_\theta = \left[ \text{diag}(\Theta_{AB} 1_B) \right]^{-1} \Theta_{AB} \]  

(1.38)
Note, $M_\theta$ is not simply the transpose of $\tilde{\Theta}_{AB}$. By comparing (1.37) with (1.38), we note the fundamental difference between $\tilde{\Theta}_{AB}$ and $M_\theta$ is that:

- In GWSM, each entry of $\tilde{\Theta}_{AB}$ is standardized by using the sum of the corresponding column as a divisor.

- In the AW under GLWM, each entry of $M_\theta$ is standardized by using the sum of the corresponding row as a divisor.

In fact, in the AW under GLWM, when all the $\theta_{ji}^{AB} \equiv 1$ or 0, depending on whether or not there is a link between the frame unit $j$ and the population element $i$, $A_\theta$ and $M_\theta$ are reduced to our version of the adjacency matrix $A$ and the row stochastic matrix $M$. One numerical example of $A$ and $M$ is shown in equation (1.6). But, as stated $A_\theta$ and $M_\theta$ are general symbolic expressions that theoretically facilitate one seeing the difference between GSWM and the AW under GLWM.

### 1.3.2 A Numerical Example

Here we use the example on page 55 in Lavellee (2007). To begin with, the $\Theta_{AB}$ matrix has form,

$$
\Theta_{AB} = \begin{bmatrix}
\theta_{11}^{AB} & 0 \\
\theta_{21}^{AB} & \theta_{22}^{AB} \\
0 & \theta_{32}^{AB}
\end{bmatrix}
$$

In the following matrix, we list the results from both methods.

$$
\begin{bmatrix}
\text{Samples } s^A & \hat{Y}^B : \text{GWSM} & \hat{Y}_{AW} : \text{GLWM} \\
\{1, 2\} & \frac{3}{2} \left[ y_1 + \frac{y_2}{2} \right] & \frac{1}{2} \left[ 3y_1 + y_2 \right] \\
\{1, 3\} & \frac{3}{2} \left[ \frac{y_1}{2} + \frac{y_2}{2} \right] & y_1 + y_2 \\
\{2, 3\} & \frac{3}{2} \left[ \frac{y_2}{2} + y_2 \right] & \frac{1}{2} \left[ y_1 + 3y_2 \right]
\end{bmatrix}
$$

(1.39)
From this example, the difference between these two methods is now clear: both give unbiased population total estimates, however, they obtain the unbiasedness in different ways.

### 1.3.3 An Example in Variance Comparison

To compare the estimation variance of the two methods in (1.39), we simply need to investigate the difference between the following two quantities,

\[
S_{WS} = (1.5y_1 + 0.75y_2 - y_1 - y_2)^2 + (0.75y_1 + 0.75y_2 - y_1 - y_2)^2 + (0.75y_1 + 1.5y_2 - y_1 - y_2)^2 \\
S_{AW} = (1.5y_1 + 0.5y_2 - y_1 - y_2)^2 + (y_1 + y_2 - y_1 - y_2)^2 + (0.5y_1 + 1.5y_2 - y_1 - y_2)^2 
\]

yielding,

\[
S_{WS} - S_{AW} = -0.125y_1^2 + 0.625y_1y_2 - 0.125y_2^2
\]

So, assuming \(y_1\) and \(y_2\) are both positive, if \(y_1 \leq 0.208712y_2\) or \(y_1 \geq 4.79129y_2\) then \(S_{WS} \leq S_{AW}\), which means the Weight Share method produces smaller estimation variance. On the other hand, if \(0.208712y_2 \leq y_1 \leq 4.79129y_2\) then the conclusion is reversed.

### 1.3.4 Inclusion Relationship

In the SRSWOR, we can assume that each sample \(s^A\) is obtained with equal probability. With this assumption, we can show that \(\Psi_{GWSM} \subset \Psi_{GLWM}\), where \(\Psi_\delta\) denotes the set of population total estimators belongs to family \(\delta\).

To see the proper inclusion, we need to note that, the Horwitz-Thompson estimator, as we define it, does not in general belong to \(\Psi_{GWSM}\) since we focus on the inclusion probabilities of the population elements while \(\Psi_{GWSM}\), as Lavelle defines it, needs to be aligned with the inclusion probability of the frame units.
Generally, $\Psi_{GWSM}$ can be characterized as

$$
\Psi_{GWSM} = \{ \hat{Y}^B \mid \hat{Y}^B = 1_A' T_A \Pi_A^{-1} \Theta_{AB} Y \text{ for some } \Theta_{AB} \in R_{AB}^+ \}
$$

(1.40)

$$
= \{ \hat{Y}^B \mid \hat{Y}^B = 1_A' T_A \Pi_A^{-1} \Theta_{AB} [\text{diag}(1_A' \Theta_{AB})]^{-1} Y \text{ for some } \Theta_{AB} \in R_{AB}^+ \}
$$

(1.41)

where $\Pi_A = \text{diag}(\pi_A)$, $T_A = \text{diag}(t_A)$, $t_A = \{t_A^1, t_A^2, ..., t_A^N_A\}$ and $t_A^j = 1$ if $j \in s_A$.

In (1.41), it is easy to see that $1_A' T_A = t_A$, which is a row vector. Moreover, with equal sample probabilities in SRSWOR, it is easy to see that $\pi_A^j = (N_A - 1)/N_A = n/N_A$. This implies that $\Pi_A^{-1} = (N_A/n) I_{N_A \times N_A}$.

To be precise, we have used the additional notation, $n$, to denote the size of the sample taken over the frame units. We also introduce $\xi$ to indicate randomness. With this notation, a random sample results in the diagonal matrix $T_A^\xi$, $1 \leq \xi \leq (N_A)$.

Recall, our GLWM estimator is defined as follows.

Let $\Lambda_\psi$ be a matrix having $|\Omega_{U,n}| = \binom{N_A}{n}$ rows and $N_P = N^B$ columns with real entries such that $\sum_i \Lambda_i^j \neq 0$ for each $j$. The symbol $\psi$ is a placeholder for some estimation method and $\Lambda_\psi$ is $\psi$ specific. Let the weight vector for population elements $\Omega_E$ under method $\psi$ be denoted by $w_\psi$ and define,

$$
w_\psi^t = \frac{1_{|\Omega_{U,n}|}^t \Lambda_\psi}{|\Omega_{U,n}|}
$$

(1.42)

whose $j$th coordinate is denoted by $w_\psi^j$. Further, define the $\psi$ - weighted values of the population elements to be:

$$
Y_\psi^t = [y^1/w_\psi^1, y^2/w_\psi^2, ..., y^{N_P}/w_\psi^{N_P}]
$$

(1.43)

Then define a general linear estimator of the population total $Y$ given a sample $\Omega_{U,n}^\xi$ as:

$$
\hat{Y}_\psi(\Omega_{U,n}^\xi) = \Lambda_\psi^t Y_\psi
$$

(1.44)
where $\Lambda_{\psi}^{\xi}$ denotes the $\xi$th row of the matrix $\Lambda_{\psi}$.

Now, let $\Lambda_{\psi}^{\xi} = \frac{N_A}{n} t^{A,\xi} \Theta_{AB} = t^{A,\xi} \Pi_A^{-1} \Theta_{AB}$, then it is easy to see that

$$1'_{(N_A/n)} \Lambda_{\psi} = \frac{N_A}{n} \left( \frac{N_A}{n} - 1 \right) 1'_{N_A} \Theta_{AB} = \left( \frac{N_A}{n} \right) 1'_{N_A} \Theta_{AB} \quad (1.45)$$

Comparing (B.1) and (1.41), we can set:

$$w_{\psi}^t = \frac{1'_{(N_A/n)} \Lambda_{\psi}}{(N_A/n)} = 1'_{N_A} \Theta_{AB} \quad (1.46)$$

So, we have: $[\text{diag}(1'_{A} \Theta_{AB})]^{-1} Y$ in (1.41) can be written as the weighted vector as in (1.43). Hence the inclusion relationship.

### 1.4 Matrix Approach for Variance Calculation and Characterization

Recall that for a population of values $\{x_j\}_{j=1}^n$, one can express the variance $\sum_{j=1}^n (x_j - \bar{x})^2/n$ where $\bar{x} = \sum_{j=1}^n x_j/n$, in matrix notation as,

$$\frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 = x^t \left[ \frac{1}{n} (I_n - \frac{1}{n} 1_n 1_n^t) \right] x \quad (1.47)$$

where $x^t = [x_1 \ x_2 \ ... \ x_n]$, and $I_n$ is the identity matrix of size $n$.

In our many-to-many settings, (1.47) is useful in investigating the variance dominance of one estimator over another via a difference score. Suppose that the variance of estimator 1 can be written as $x^t B_1 x$ for some suitable square matrix $B_1$ and the variance of estimator 2 can be written as $x^t B_2 x$ for some other suitable square matrix $B_2$. Then the difference score for these two estimators is $x^t (B_1 - B_2) x$. Solving the eigen system of $B_1 - B_2$ to obtain the characterization for definiteness will determine the nature of the variance domination of
one estimator over another. In particular, if $B_1 - B_2$ is positive definite, then estimator 2 is better; if $B_1 - B_2$ is negative definite, then estimator 1 is better; if $B_1 - B_2$ is indefinite, then the comparison is inconclusive; if $B_1 - B_2$ is positive semi-definite, then estimator 2 is at least as good as estimator 1; if $B_1 - B_2$ is negative semi-definite, then estimator 1 is at least as good as estimator 2. This holds true for any data set. For a variety of reasons, one might prefer to examine a variance ratio rather than a difference score in efficiency research. But as Byczkowski and Levy (2009) discovered, this can be analytically insurmountable.

The key advantage of basing variance domination comparisons on the definiteness of $B_1 - B_2$, is that they are data independent. That is, the characterization of the variance difference does not depend on any particular realization or sample of $x$. This provides a compliment to conducting simulation studies, as in Byczkowski and Levy (2009), to characterize the variance dominance, although as stated, both approaches have their pros and cons.

1.4.1 Arc-Weight Method

We now present the matrix representation for calculating the variance of the AW estimator for the estimated population total.

Let $W_{AW} = \text{diag}(w_{AW})$, a diagonal matrix with $(j,j)$th entry equal to the $j$th coordinate of $w_{AW}$. Let $B_{AW}$ denote the matrix used to calculate the variance of the estimated population total using the AW method. Let $y^t = [y^1 \ y^2 \ ... \ y^{N_F}]$, then the variance can be written as $y^t B_{AW} y$, where

$$B_{AW} = W_{AW}^t M^t C_{N_F,n}^t \left[ I_{N_F} - \frac{1}{N_F} \binom{N_F}{n} \binom{N_F}{n}^t \right] C_{N_F,n} M W_{AW} \left( \frac{N_F}{n} \right)^2 \frac{1}{N_F}$$  (1.48)

(1.48) follows from (1.47) and (1.12).
1.4.2 Horwitz-Thompson Method

Let $W_{HT} = \text{diag}(w_{HT})$, a diagonal matrix with $(j,j)$th entry equal to the $j$th coordinate of $w_{HT}$. Let $B_{HT}$ denote the matrix used to calculate the variance of the estimated population total using the HT method. The variance can be written as $y^t B_{HT} y$, where

\[
B_{HT} = W_{HT}^t \Lambda^t \left[ I_i^{N_F} - \frac{1}{(N_F)} 1_i^{N_F} 1_i^{N_F} \right] \Lambda W_{HT} \frac{1}{(N_F)} \tag{1.49}
\]

(1.49) follows from (1.47) and (1.19).

1.4.3 General Linear Weight Method

Let $W_\psi = \text{diag}(w_\psi)$, a diagonal matrix with $(j,j)$th entry equal to the $j$th coordinate of $w_\psi$. Let $B_\psi$ denote the matrix used to calculate the variance of the estimated population total using the General Linear Weight Method. The variance can be written as $y^t B_\psi y$, where

\[
B_\psi = W_\psi^t \Lambda_\psi \left[ I_i^{N_F} - \frac{1}{(N_F)} 1_i^{N_F} 1_i^{N_F} \right] \Lambda_\psi W_\psi \frac{1}{(N_F)} \tag{1.50}
\]

Equation (1.50) follows from (1.47) and (1.44).

1.4.4 Variance Dominance Study and Empirical Findings

In the light of (1.48) and (1.49), one might ask, "Does there exist a many-to-many structure in which one method dominates the other in that it has uniformly smaller estimation variance regardless of the values assumed by population elements? Here is an answer:

**Theorem 5.** Let $B_1$ and $B_2 \in B_\psi = \bigcup_\psi B_\psi$, where $B_\psi$ is defined by (1.50) and $\psi \in \Psi$. $B_1 \neq B_2$. Then, for any many-to-many structure, $B_1 - B_2$ is either semi-definite or indefinite.
Proof. Let $\Delta$ denote the common factor in $B_1, B_2$, i.e.,

$$
\Delta = \left[ I_{\binom{N_F}{n}} - \frac{1}{\binom{N_F}{n}} 1_{\binom{N_F}{n}}1_{\binom{N_F}{n}}^T \right] \frac{1}{\binom{N_F}{n}}
$$

Note that $\Delta$ is a multiple of a symmetric idempotent matrix with rank $\binom{N_F}{n} - 1$, and hence $\text{Det}(\Delta) = 0$, i.e. $\Delta$ is singular. Furthermore, for any matrix $\Upsilon$, we have $\text{Det}(\Upsilon^T \Delta \Upsilon) = 0$ too. Also, note that $B_k$ can be factored as $\Xi_k^T \Delta \Xi_k$ for some suitably chosen $\Xi_k$, $1 \leq k \leq 2$ where $\Xi_k$ is dependent on the structure of the underlying arbitrary many-to-many frame population and method $k$, say. In general, $\Xi_k = \Lambda_k W_k / \sqrt{\binom{N_F}{n}}$. Setting $\Upsilon = \Xi_1 - \Xi_2$ it then follows that $B_1 - B_2$ is singular. Since positive or negative definite matrices CANNOT be singular, there does NOT exist a many-to-many structure such that one method has uniformly smaller estimation variance than the other method on a data independent basis. In other words, the possible kinds of variance comparison results between any two estimators in the $\Psi$ family given any many-to-many structures can only be either semi-definite or indefinite. Because the matrices flanking $\Delta$ in each of the above formula are not trivial (e.g., not an identify matrix), the non-dominance result is non-trivial as it is in the case of (1.47) where it is clear that only a vector of constants makes the formula equal to 0.

Theorem 5 also implies that there is NO UMVUE in the family $\Psi$.

Our empirical findings are consistent with the above result in which we focused on the AW and HT methods. Specifically, we performed an exhaustive case search over all possible non-isomorphic many-to-many frame population structures where the number of arcs is $\leq 13$. In terms of the adjacency matrix $A$, it is equivalent to say that we have enumerated all the non-isomorphic cases where $\sum_i \sum_j a_{ij} \leq 13$. In each of these cases, we let the sample size vary from 1 to $N_F$ exhaustively and investigated the eigen structure of $B_{AW} - B_{HT}$. The variance comparison results are exclusively either semi-definite or indefinite.
1.5 Variance Calculation Revisited

From the previous sections, we saw that when we apply the matrix approach for comparing the AW and HT sampling variances in estimating population totals, we needed to enumerate all the rows of the combinatorial instance matrix $C_{N_F,n}$ in (1.48) in the AW method and enumerate all the rows of the matrix $\Lambda$ in (1.49) in the HT method in order to characterize the variance difference on a data free basis. One downside to this approach is that when the system size increases, it becomes impractical quickly, as was illustrated at the end of Section 1.2.4, where we choose 50 frame units out of 100. Due to the high computation cost associated with obtaining the matrix $\Lambda$ for the HT method, we limit our scope here to the AW method only. We now propose a method to eliminate the necessity of enumerating the matrix $C_{N_F,n}$ when we compare the variance of the AW and perfect frame estimators. By perfect frame we mean the traditional complete and perfect frame sampling setting.

The central idea is to construct an equivalent complete and perfect frame-population relationship with suitable associated values. We call this new frame-population relationship the apportioned frame-population. Usually, the number of frame units does not equal the number of population elements. We need to fix this by somehow expanding or condensing the set of population elements. For example, we want to make both sides of the bipartite graph shown in Figure 2.1, representing the frame-population network, have the same number of nodes.

Observe that according to (1.12), for the AW method we have:

$$\hat{Y}_{AW}(\Omega_{U,n}^{c_k}) = C_{N_F,n}^{c_k} M y_{AW} \frac{N_F}{n}$$

The key is to treat $M y_{AW}$, a column vector of size $N_F$, as the new set of population elements. This will result in the change needed as shown in Figure 1.7. The difference between the original frame-population and the apportioned frame-population, among other things, is that the values for the population elements change as shown in Table 1.4. The column labeled
Figure 1.7: original frame-population and the apportioned frame-population, referring to the Figure 1.2 example

Table 1.4: population element values: original v.s. apportioned

<table>
<thead>
<tr>
<th>Frame Units</th>
<th>Original Values</th>
<th>Apportioned Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
<td>$y^1$</td>
<td>$y^1/2 + y^3$</td>
</tr>
<tr>
<td>$U_2$</td>
<td>$y^2$</td>
<td>$y^1/2 + y^2$</td>
</tr>
<tr>
<td>$U_3$</td>
<td>$y^3$</td>
<td>$2y^4/5$</td>
</tr>
<tr>
<td>$U_4$</td>
<td>$y^4$</td>
<td>$2y^4/5$</td>
</tr>
<tr>
<td>$U_5$</td>
<td>$y^5$</td>
<td>$y^4/5 + y^5$</td>
</tr>
</tbody>
</table>
Apportioned Values in Table 1.4 is obtained as:

\[
\begin{bmatrix}
y_1^A P \\
y_2^A P \\
y_3^A P \\
y_4^A P \\
y_5^A P \\
\end{bmatrix} = \begin{bmatrix}
1/2 & 0 & 1/2 & 0 & 0 \\
1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1/2 & 1/2 & 0
\end{bmatrix} \begin{bmatrix}
y^1/1 \\
y^2/(1/2) \\
y^3/(1/2) \\
y^4/(5/2) \\
y^5/(1/2)
\end{bmatrix} = \begin{bmatrix}
y^1/2 + y^3 \\
y^1/2 + y^2 \\
2y^4/5 \\
2y^4/5 \\
y^4/5 + y^5
\end{bmatrix}
\] (1.51)

where \( y_{AP}^k \) is the \( k \)th coordinate of \( y_{AP} \), the apportioned population values vector. Now, taking samples of size \( n \) from the new apportioned frame-population is comparable to the the traditional complete and perfect frame sampling setting. We now show how to use the apportioned frame-population to obtain the matrix \( B_{AW} \) in (1.48) without enumerating \( C_{NP,n} \).

Let \( \Omega_{U,n}^{\xi} \) be the sample drawn from the system. Let \( \Omega_{i_u,n}^{\xi,\nu} \) denote the \( \nu \)th element in the sample, \( 1 \leq \nu \leq n \). Fixing \( \xi \), for each \( \nu \), there is a \( i_u^\xi \in I_U \) with apportioned value \( y_{AP}^{i_u^\xi} \). The following holds trivially:

\[
\hat{Y}_{AW}(\Omega_{U,n}^{\xi}) = C_{NP,n}^{\xi} M y_{AW} \frac{N_F}{n} = \sum_{\nu=1}^{n} y_{AP}^{i_u^\xi} \frac{N_F}{n}
\] (1.52)

By Theorem 1, \( \hat{Y}_{AW}(\Omega_{U,n}^{\xi}) \) is unbiased for the population total \( \sum_{j \in I_E} y^j \). The same is true
for $\sum_{\nu=1}^n y_{i\nu}^\epsilon N_F/n$ by (1.52). Let $V$ denote the estimation variance, then we have

$$V = \frac{1}{(N_F/n)} \sum_{\xi=1}^{(N_F)} \left( \hat{Y}_{AW}(n) - \sum_{j \in I_E} y_j \right)^2$$

(1.53)

$$= \frac{1}{(N_F/n)} \left[ \sum_{\xi=1}^{(N_F)} \left( \sum_{\nu=1}^n y_{i\nu}^\epsilon \frac{N_F}{n} \right)^2 - \left( \frac{N_F}{n} \right) \left( \sum_{j \in I_E} y_j \right)^2 \right]$$

(1.54)

$$= \frac{1}{(N_F/n)} (V_1 - V_2)$$

Note that in (1.54), $V_2$ is easy to expand. We now deal with $V_1$:

$$V_1 = \sum_{\xi=1}^{(N_F)} \left( \sum_{\nu=1}^n y_{i\nu}^\epsilon \frac{N_F}{n} \right)^2 = \left( \frac{N_F}{n} \right)^2 \sum_{\xi=1}^{(N_F)} \left( y_{i1}^\epsilon + y_{i2}^\epsilon + ... + y_{iN_F}^\epsilon \right)^2$$

$$= \left( \frac{N_F}{n} \right)^2 \left[ \left( \frac{N_F - 1}{n - 1} \right) \sum_{j=1}^{(N_F)} (y_{ij}^\epsilon)^2 + \left( \frac{N_F - 2}{n - 2} \right) \sum_{i=1}^{N_F-1} \sum_{j=i+1}^{N_F} 2 y_{ij}^\epsilon y_{ij}^\epsilon \right]$$

(1.55)

Equation (1.55) is quite an economical, time saving way of computing $V_1$ since one only needs to collect square terms $N_F$ times and quadratic cross terms $N_F(N_F - 1)/2$ times. Ultimately, after algebraically expanding $V_1$ and $V_2$ and collecting terms, one ends up with $N_P$ square terms in $y_j^i$, $1 \leq j \leq N_P$, and $N_P(N_P-1)/2$ cross terms in $y_i^j y_j^i$ for $1 \leq i \leq N_P - 1$, $i+1 \leq j \leq N_P$. Let the coefficient of the $j$th square term of $y_j^i$ be denoted by $\alpha_j$, $1 \leq j \leq N_P$. Let the coefficient of the $ij$th cross term of $y_i^j y_j^i$ be denoted by $\beta_{ij}$, $1 \leq i \leq N_P - 1$, $i + 1 \leq j \leq N_P$.

On the other hand, note that the matrix $B_{AW}$ is symmetric and of size $N_P$. $y^i B_{AW} y$ is a quadratic form in $y$ and it has unique algebraic expansions in terms of the coordinates of $y$ that exactly equals $V$ as in (1.53). Therefore, we have the following relationship:
\[ B_{ij}^A_W = \alpha_i \text{ if } i = j \]
\[ B_{ij}^A_W = B_{ij}^A_W = \beta_{ij}/2 \text{ if } i < j \]

Now, the idea is clear: restore \( B_{AW} \) by obtaining \( \alpha_j \) and \( \beta_{ij}/2 \). This completes the process of obtaining \( B_{AW} \) without enumerating \( C_{N_F,n} \). Of course, one would need to rely on a state-of-the-art symbolic package, e.g., Mathematica® for doing all this algebraic work.

### 1.6 A Mathematica Add-On Package

We have implemented a Mathematica® Add-On package that does all the symbolic manipulation work outlined in the previous sections. It accepts the adjacency matrix \( A \) as input and computes \( w_{AW} \) and \( w_{HT} \). It also produces \( \hat{Y}_{AW}(\Omega_{U,n}^\xi) \), \( \hat{Y}_{HT}(\Omega_{U,n}^\xi) \) for each \( \xi \) when \( n \) is given. More importantly, it computes \( B_{AW} \) and \( B_{HT} \) and characterizes the eigen structure of \( B_{AW} - B_{HT} \). In particular for the AW method, it can produce \( B_{AW} \) when a complex many-to-many structure is given, and at sizes of \( N_P \) up into the thousands! The Add-On package also has the facility to graphically generates all frame-population scenarios of moderate sizes having the whole panoply of many-to-many structures.

The package is made available at the following URLs: http://docs.google.com/leaf?id=0By8sw065EW-GZDk0MzFiNzQtN2NkZC00YzdhLTk4NGItYjk3N2JiYTRmYmFm&hl=en and http://docs.google.com/leaf?id=0By8sw065EW-GZTUxYTA4OGQtZmVjYi00MWQ0LTkyZjctZjE3OGQiZjY3OWVl&hl=en.

For a large matrix, solving for the eigen structure is computationally expensive. Here we choose to use Gerschgorin’s Theorem to estimate the locations of the eigen values. For details, see Atkinson (1989).

### 1.7 Discussion, Conclusion, and Future Work

Our matrix based approach for representing the estimate of the population total in the many-to-many scenario serves as a characterization tool to help understand when and why there
is estimation precision dominance. In conducting an empirical study (say, by simulation) that explores various frame-population scenarios where the structural information has to be intermingled with the values of the population elements, it is a complex task to track down the exact cause of the precision dominance. Given that one finds there is significant estimation precision dominance, is it due to the population-frame structure or is it due to the values of the population elements? Furthermore, inevitable interactions may only be partially explained by recursive partitioning methods (see, e.g., Byczkowski and Levy (2009)). The approach taken in this paper separates the structural effect from the data by confining it within a variance matrix that captures only the structural information independent of any particular data set, i.e., values assumed by the population elements. Both approaches based on difference scores have their advantages and disadvantages.

In using the HT method, our approach to calculating the 1st order inclusion probability improves the methodology used in Clough (2000), where scenarios of 5 or 6 population elements are the best workable cases computationally feasible in that empirical study. Our approach provides an improvement over that study, permitting exploration of a population size of 20 with ease. However, since our approach still uses the combinatorial instance matrix $C_{N,F,n}$, which grows exponentially, it appears that no further size increase is immediately feasible.

In using the AW method, our approach to obtaining the variance matrix works efficiently. Since it does not require the combinatorial instance enumeration, it can work out complex problems of sizes with 1000 population elements. This greatly improves our ability to characterize the difference between AW and the perfect frame estimator under messy frame settings, including the complex many-to-many case, without doing any simulation study.

In this paper we have restricted attention to the case of simple random sampling without replacement so as to facilitate an introduction to our matrix based approach. The case of stratified random sampling design is much more complex since it deals with issues such as mis-stratification and a concept called crossover. Extensions of this work to the stratified
random sampling design will be forthcoming. Possible future research might also involve a model-based approach to investigate and achieve asymptotic results.
Chapter 2

On Constrained Logistic Regression with an Application to Corporate Bankruptcy Prediction

2.1 Introduction

Motivated by the problem of predicting corporate bankruptcies, we consider a constrained logistic regression approach with intelligently selected constraints on certain parameters reflecting prior beliefs or judgement. The notation for the data set is as follows. For each record representing a company in the data set, we have \([X_i, Y_i] = [X_{i,1}, X_{i,2}, ..., X_{i,k}, Y_i]\), where \(Y_i\) is binary: \(Y_i = 1\) signifies bankruptcy, \(Y_i = 0\) means otherwise; The independent variables, \(X_{i,1}, X_{i,2}, ..., X_{i,k}\), are important financial variables or financial ratios that are commonly used in these sorts of default probability prediction problems. Often, practitioners have useful prior information based on their experience, judgement, expert advice, and/or other types of proprietary resources. In addition, public information on credit ratings of individual corporations such as Moody’s Investors Service, Standard & Poor’s, and Fitch Ratings are often available. It is therefore desirable to have a systematic way of incorpo-
rating this information into the logistic regression model in the hope of improving default probability prediction. Our method can be construed as an alternative to Bayesian methods for incorporating such information.

Logistic regression models [McCullagh and Nelder (1989)] have been widely used in many disciplines to predict dichotomous outcomes. As examples of fields using logistic modeling, consider the medical literature where such models have been successfully applied in areas such as obstetrics and pediatrics, just to name a few, in which infant survival is of key interest. In the accounting and finance literature, logistic regression models have been used to create credit score-cards, to predict bankruptcy, and as evaluators of corporate financial distress, etc. In political science, logistic regression models have been used to analyze rare event data such as terrorist attacks, and of course in the prediction of election outcomes. All of these applications have been greatly aided by the widespread availability of logistic model fitting routines in major statistical packages such as SAS SPSS, Gauss, Stata, and others.

The literature on applications of the logistic regression model to the prediction of bankruptcies is quite large. See for example Altman (1968), Ohlson (1980), Wiginton (1980), etc. A popular approach in estimating the coefficients in the logistic regression model is the Maximum Likelihood Estimator (MLE) [Wedderburn (1976)]. It is well known that MLEs enjoy nice asymptotic properties, e.g. strong consistency and asymptotic normal distribution [Fahrmeir and Kaufmann (1985)]. However, in a more realistic setting, with moderate sample sizes, and possible relatively small number of bankruptcy cases (leading to possible complete separation issues), we have found that there is room to improve parameter estimation in the logistic regression model. Our constrained approach offers an attractive alternative that, 1) possesses nice asymptotic properties, 2) ameliorates the separation issues, and 3) actually improves the prediction performance.
Specifically, we propose a constrained logistic regression model in these instances. The constraints on parameters are chosen to reflect strong prior beliefs or good business judgment. In one compelling study involving a data set obtained from the Compustat through the Wharton Research Data Services (WRDS) we provide empirical evidence of the virtue of the constrained approach. In one example of a favorable finding, we show that with a sample size of $n = 350$, in 80% of the trials featuring pairs of selected training and out-of-sample prediction data sets, the constrained approach yields AUCs, that is, area under a receiver operating characteristic curve (ROC curve) in the out-of-sample set that are larger, and hence of superior predictive quality than AUCs for the classical MLE approach. We further demonstrate the predictive superiority of the constrained logistic models over the usual unconstrained MLE approach based on a number of metrics in a simulation study involving moderately sized data sets possessing characteristics similar to those of the WRDS data set, thus reinforcing our findings. For all of these metrics considered, the constrained approach outperforms the classical MLE method. Of theoretical interest, with an eye toward inference-making, we have established large sample properties of the constrained logistic parameter estimates. These show that under our constrained logistic regression model, parameters are strongly consistent. We also establish their asymptotic distribution by applying the proof techniques in the classical papers of Dupacová and Wets (1988) and Wets (1991).

(1996) considered asymptotics of least squares estimators for constrained nonlinear regression.

The regularization idea, e.g. Hoerl and Kennard (1970), Tibshirani (1996), Zou and Hastie (2005) and Zou (2006), in the logistic regression context can be considered as one form of shrinkage constrained logistic regression. They constrain the $L_1$ norm, $L_2$ norm, or convex combination of the $L_1$ and $L_2$ norm of the estimated parameter vector.

Piegosch and Casella (1988) considered adding constraints on independent variables for logistic regression models. This approach falls into the general framework of robust estimation. In some situations, it is also necessary to add constraints on dependent variables to incorporate sampling variability. The following works deal with robust estimation in logistic regression: Pregibon (1982), Stefanski et al. (1986), Copas (1988), Künsch et al. (1989), Morgenthaler (1992), Carroll and Pederson (1993), Bianco and Yohai (1996), Kordzakhia et al. (2001) and Müller and Neykov (2003), among others. The robust optimization approach is different from our approach in that the robust optimization approach adds bounds on data (e.g. dependent or independent variables) and our approach adds constraints on event probabilities and consequently on model parameters.

There are various Bayesian approaches to logistic regression modeling. These offer one alternative for incorporating useful prior distributional information about the parameter vector. For example, Bedrick et al. (1996) considered assigning certain Beta distributions as prior distributions for event probabilities in the context of Bayesian logistic regression. Other works using Bayesian approaches in logistic regression models include, Ibrahim and Laud (1991), Chen et al. (1999), Meyer and Laud (2002), and Robert and Casella (2004).

Both the Bayesian and our constrained approach to logistic model estimation enable practi-
tioners to integrate valuable prior information on event probabilities as part of the estimation process. In our constrained logistic regression approach, the parameter constraints that the modeler supplies focus the available prior information on event probabilities in the form of equality and/or inequality restrictions. Thus, the constraints in our approach are all ‘hard’, in the sense that although they arise from modelers belief, they take the form of unconditional statements about the event probabilities. This contrasts with most Bayesian approaches, in which the constraints may be described as being ‘soft’. That is, the constraints are imposed using prior probability distributions on the parameters. To ensure the validity of the added parameter constraints of our approach in applications, there are numerous flexible options for choosing reliable information on event probabilities such as soliciting expert advice, using published information, etc.

Our constrained approach also offers computational advantages in small and moderate sample cases. In particular, as discussed in Section 2.3.3, whenever there exists complete separation, or if all the responses have the same value, e.g. \( Y_i = 0 \), adding suitable parameter constraints can ameliorate the problem of parameter estimates that are boundless. Technically, in order to estimate the model parameters of our constrained logistic regression model, a nonlinear programming package is necessary. Standard computational algorithms for solving constrained non-linear optimization problems are made available as callable functions or procedures in most standard packages, e.g. SAS, Matlab, Gauss, etc.

The rest of the paper is arranged as follows. In Section 2.2, we formulate our constrained logistic regression approach. We discuss the numerical computing part of our approach in Section 2.3. In Section 2.4, we report results based on a real bankruptcy prediction data set. In Section 2.5, we report results arising from a simulated data set. In Section 2.6, we present the large sample properties of our approach.
2.2 The Constrained Logistic Regression Model

2.2.1 The Logistic Regression Model

Let \((\Omega, \mathcal{G}, P)\) be a probability space. Let \(\omega_i \in \Omega, \omega_i = [X_i, Y_i] = [X_{i,1}, X_{i,2}, ..., X_{i,k}, Y_i]\) be a random row vector of size \(k + 1, k \geq 1\) and \(k \in \mathbb{N}\), with \(X_i \in \mathbb{R}^k\) and \(Y_i \in \{0, 1\}\). \(\omega_i\) are independently and identically \(P\)-distributed, \(i = 1, 2, ...\). The relationship between \(X_i\) and \(Y_i\) is given by the logistic regression model [McCullagh and Nelder (1989)]

\[
\Pr(Y_i = 1|X_i) = \frac{1}{1 + \exp(-X_i\beta)},
\]

and

\[
\Pr(Y_i = 0|X_i) = 1 - \Pr(Y_i = 1|X_i) = \frac{1}{1 + \exp(X_i\beta)},
\]

where \(\beta\) is a column vector of size \(k\) of unknown parameters to be estimated, \(\beta \in \mathbb{R}^k\). In other words, given \(X_i\) and \(\beta\), \(Y_i\) has Bernoulli distribution with event probability equal to \(\Pr(Y_i = 1|X_i)\). We also assume that the sample design matrix, denoted by \(X\) is of full column rank.

Following the logistic regression model, the negative loglikelihood function, denoted by \(\ell(\beta, \omega_i)\), is:

\[
\ell(\beta, \omega_i) = -Y_iX_i\beta + \log(1 + \exp(X_i\beta)) .
\] (2.1)

Denoting \(\Pr(Y_i = 1|X_i)\) by \(p_i\), standard algebra yields the first order derivative of \(\ell\) as,

\[
\frac{\partial \ell(\beta, \omega_i)}{\partial \beta} = -\sum_{i=1}^{n}(Y_i - p_i)X_i ,
\] (2.2)

and the second order derivative of \(\ell\) as,

\[
\frac{\partial^2 \ell(\beta, \omega_i)}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} p_i(1 - p_i)X_iX_i' .
\] (2.3)
Since the design matrix $X$ is of full column rank, one can show that (2.3) is positive definite. This implies that $\ell(\beta, \omega_i)$ is a strictly convex function in $\beta$. The convexity property of $\ell$ has several good implications. One of them is that it ensures that when there are no parameter constraints, whenever $\ell$ has a local minimizer, it is then a global minimizer. Moreover, it is unique.

At the population level, based on the probability space $(\Omega, \mathcal{G}, P)$, if the distribution $P$ is known, the estimation problem is defined to be:

$$\min_{\beta \in \mathbb{R}^k} \int_{\Omega} \ell(\beta, \omega)P(d\omega) .$$

(2.4)

The optimal solution to (2.4) is denoted by $\beta^*$. Given $X_i$, the event probability evaluated at $\beta^*$ is denoted by $p^*_i$.

More realistically, at the sample level, since $P$ is not known and only partial information about it is available, we use the following to describe the sampling process. Let $(M, \mathcal{F}, \mu)$ be the sample space. Let $(\mathcal{F}_i)_{i=1}^{\infty}$ be an increasing sequence of $\sigma$-fields contained in $\mathcal{F}$. Let $P^n$ be a conditional probability measure: $P^n : \mathcal{G} \times M \rightarrow [0, 1]$, for each $n$. It follows from Billingsley (1995) that, for all $m$ in $M$, $P^n(\cdot, m)$ is a probability measure on $(\Omega, \mathcal{G})$. For all $S \in \mathcal{G}, m \rightarrow P^n(S, m)$ is $\mathcal{F}$-measurable.

In our setup, we take $P^n$ to be the empirical measure. It follows that for $\mu$-almost all $m$ in $M$, $P^n(\cdot, m)$ converges weakly to $P$ as $n \rightarrow \infty$. At the sample level, the estimation problem is then defined based on the empirical measure.

$$\min_{\beta \in \mathbb{R}^k} \int_{\Omega} \ell(\beta, \omega)P^n(d\omega) .$$

(2.5)

The optimal solution to (2.5) is denoted by $\widehat{\beta}^n_{\text{mle}}$. Given $X_i$, the event probability evaluated
at $\hat{\beta}_{mle}^n$ is denoted by $\hat{\beta}_{mle}^n$.

Equation (2.5) can be re-written as:

$$\text{Minimize } \hat{\beta} \in \mathbb{R}^k \frac{1}{n} \sum_{i=1}^{n} -Y_i X_i \beta + \log(1 + \exp(X_i \beta)). \quad (2.6)$$

At the sample level, under suitable regularity conditions, the existence, uniqueness as well as consistency and asymptotic normality of the parameter estimates using the maximum likelihood estimation method can be found in Nelder and Wedderburn (1972), Gourieroux and Monfort (1981), Fahrmeir and Kaufmann (1985), among others.

In practice, model parameters can be estimated using a variety of numerical methods, e.g. Newton-Ralphson’s method, Iteratively Reweighted Least Square (IRLS) [Green (1984)], Conjugate Gradient (CG). See Minka (2003) for a nice review.

### 2.2.2 The Parameter Constraints

We constrain $\beta$ on a set $\Theta$, $\Theta \subset \mathbb{R}^k$. $\Theta$ is non-empty and closed, but not necessarily bounded. Throughout, we assume that the constraint set $\Theta$ is the same at both the population level and the sample level. In applications, a general way of forming the set $\Theta$ is:

$$\Theta = \{ \beta \in \mathbb{R}^k | g_j(\beta) \leq 0, j \in I_1 \text{ and } h_j(\beta) = 0, j \in I_2 \} ,$$

where $I_1$ and $I_2$ are two finite index sets. $g_i(\cdot)$ and $h_i(\cdot)$ are assumed to be continuously differentiable functions in $\beta$. In general, $g_i(\cdot)$ and $h_i(\cdot)$ can be either affine linear or nonlinear in $\beta$ for the purpose of showing the consistency of a sequence of the estimates under constraints. However, for proving asymptotic normality or conic normality, the results evidently require $g_i(\cdot)$ and $h_i(\cdot)$ to be affine linear in $\beta$. In other words, $\Theta$ needs to be a polyhedron, together with other assumptions which are cited in later sections, in order to establish asymptotic normality or conic normality. In constrained cases, unlike unconstrained cases, asymptotic
normality or conic normality becomes an exception instead of the rule [Wets (1991), Shapiro (1989)]. We discuss the parameter constraints further in Section 2.2.4.

2.2.3 The Constrained Logistic Regression Estimation Problem

At the sample level, we can rewrite (2.5) as follows:

\[
\begin{align*}
\text{Minimize} & \quad \frac{1}{n} \sum_{i=1}^{n} -Y_i X_i \beta + \log(1 + \exp(X_i \beta)), \\
\text{s.t.} & \quad g_j(\beta) \leq 0, \quad j \in I_1, \\
& \quad h_j(\beta) = 0, \quad j \in I_2.
\end{align*}
\]

(2.7)

The optimal solution to (2.7) is denoted by \( \hat{\beta}_{con}^n \). Given \( X_i \), the event probability evaluated at \( \hat{\beta}_{con}^n \) is denoted by \( \hat{p}_{con}^i \). The model (2.7) is the constrained logistic regression model we focus upon in this paper.

2.2.4 Constraint Qualifications

In model (2.7), we can add constraints that express useful prior knowledge about the data. However, there are mathematical requirements the constraints must satisfy in order to guarantee a successful search for an optimal solution. A set of necessary conditions that is implied by optimality under constraints may be found in Bertsekas (1999), Proposition 3.3.6. These conditions guarantee the existence of Lagrange multipliers for optimality.

The following three conditions represent a slightly more powerful form of the commonly known Karush-Kuhn-Tucker (KKT) conditions. Let \( \beta^* \) be a local optimal solution. A necessary condition for the optimality is the existence of Lagrange multipliers \( \mu_j^*, j \in I_1 \) and \( \lambda_j^*, j \in I_2 \) satisfying the following three conditions [Bertsekas (1999)].
(i) First Order Condition:

$$\nabla f(\beta^\circ) + \sum_{j \in I_1} \mu_j^* \nabla g_j(\beta^\circ) + \sum_{j \in I_2} \lambda_j^* \nabla h_j(\beta^\circ) = 0.$$  

(ii) Non-negativity:

$$\mu_j^* \geq 0, \quad j \in I_1.$$  

(iii) Complimentary Slackness: In every neighborhood $N$ of $\beta^\circ$ there is a $\beta \in N$ such that $\mu_j^* \nabla g_j(\beta) > 0$ for all $j$ with $\mu_j^* \neq 0$ and $\lambda_j^* \nabla h_j(\beta) > 0$ for all $j$ with $\lambda_j^* \neq 0$.

Remark 1. **Linear and Concave Constraints**

If all the functions $h_j(\cdot)$ in $I_2$ are linear in $\beta$ and all the functions $g_j(\cdot)$ in $I_1$ are concave in $\beta$, then $\hat{\beta}_{con}^n$, the sample based optimal solution, satisfies the necessary conditions for optimality [Bertsekas (1999), Proposition 3.3.7].

Note that having all the functions $g_j(\cdot)$ in $I_1$ being linear automatically satisfies the concavity in $\beta$. Moreover, since linear constraints are in fact both concave and convex, this facilitates the search for an optimal solution to model (2.7). That is, the KKT necessary condition is also sufficient for optimality. In other words, whenever there exist multipliers satisfying conditions (i) (ii) and (iii), they are the multipliers associated with the optimal solution.

Remark 2. **Mangasarian-Fromovitz Constraint Qualification**

If the gradients $\nabla h_j(\hat{\beta}_{con}^n)$ for $j \in I_2$ are linearly independent and there exists a vector $d$ such that

$$\nabla g_j(\hat{\beta}_{con}^n)' d < 0, \quad \forall j \in A_{I_1}(\hat{\beta}_{con}^n),$$

$$\nabla h_j(\hat{\beta}_{con}^n)' d = 0, \quad \forall j \in I_2,$$
where $A_{I_1}(\hat{\beta}_{con}^n)$ is the index set for which $g_j(\cdot)$ are tight or active, i.e. $g_j(\hat{\beta}_{con}^n) = 0$, then $\hat{\beta}_{con}^n$, the sample based optimal solution, satisfies the necessary conditions for optimality [Bertsekas (1999), Proposition 3.3.8].

2.3 Computations

2.3.1 Numerical Algorithms

In order to obtain the parameter estimates of the constrained logistic regression model (2.7), i.e. the optimal solution to the constrained nonlinear programming problem (2.7), there are several available options: 1. Method of Multipliers. This method is also known as the Augmented Lagrangian Method. 2. Trust-Region Sequential Quadratic Programming. 3. Trust-Region Interior Point Method. For technical details, consult Bazaraa et al. (2006), Bertsekas (1999), Nocedal and Wright (2006). These methods are general methods that work for both linear and nonlinear constraints. In this section, we present an algorithm, Method of Feasible Directions, due to Zoutendijk in Bazaraa et al. (2006) that works for the linear constraint case since the asymptotic distribution theory in the above works for the linear constraint case.

In our model (2.7), let the matrix $A$ and the vector $a$ be such that $A\beta \leq a$ represents all the inequality constraints $g_j(\beta) \leq 0$, $\forall j \in I_1$. Let the matrix $B$ and the vector $b$ be such that $B\beta = b$ represents all the equality constraints $h_j(\beta) = 0$, $\forall j \in I_2$.

*Step 0.* Find a feasible solution $\xi_1$ with $A\xi_1 \leq a$, and $B\xi_1 = b$ and go to Step 1.

*Step 1.* Given $\xi_q$, let $A$ and $a$ vertically decompose into $A_1$, $A_2$ and $a_1$, $a_2$ such that
A_1 \xi_q = a_1 \text{ and } A_2 \xi_q < a_2. \text{ Let } d_q \text{ be an optimal solution to the following problem}

\begin{align*}
\text{Minimize} & \quad \nabla \ell(\xi_q)'d, \\
\text{s.t.} & \quad A_1 d \leq 0, \\
& \quad B d = 0, \\
& \quad -1 \leq d_i \leq 1, \ i = 1, 2, ..., k.
\end{align*}

where \( d_i \) are elements of \( d \). If \( \nabla \ell(\xi_q)'d_q = 0 \), then stop: \( \xi_q \) is a KKT point, i.e. \( \xi_q \) is the optimal solution. Otherwise, go to step 2.

**step 2.** Let \( \rho_q \) be an optimal solution to the following line search problem:

\begin{align*}
\text{Minimize} & \quad \ell(\xi_q + \rho d_q), \\
\text{s.t.} & \quad 0 \leq \rho \leq \rho_{\text{max}},
\end{align*}

where \( \rho_{\text{max}} \) is determined as follows:

\[ \rho_{\text{max}} = \min \{ \hat{a}_i / \hat{d}_i : \hat{d}_i > 0 \} \] if \( \hat{d} \not< 0 \), and \( \rho_{\text{max}} = \infty \) if \( \hat{d} \leq 0 \), where \( \hat{a} = a_2 - A_2 \xi_q \) and \( \hat{d} = A_2 d_q \). Let \( \xi_{q+1} = \xi_q + \rho_q d_q \), and update \( A_1, A_2, a_1, a_2 \) by identifying binding constraints. Replace \( q \) by \( q + 1 \) and go to Step 1 until optimal solution is found in Step 1.

Note that in step 1, we need to solve a linear program and in step 2, we need to solve a one-dimensional search problem.

### 2.3.2 Computing Platforms

For small size problems, say a couple of hundreds variables, one can rely on standard nonlinear program solvers to handle the constrained model (2.7). For example, one can use the MATLAB constrained optimization function ‘fmincon’. One can also use the Gauss constrained optimization facility CMLMT or COMT. In SAS/OR, there is a PROC NLP
available to handle our constrained logistic regression model. In SAS/IML, one can use the optimization functions such as NLPNRA.

Depending on the data, most often, one can choose to supply the solver with the gradient vector (2.2), the Hessian matrix (2.3) and 'BFGS' option for Hessian updates. One can also choose to merely provide the objective function and constraints without giving the expression for the gradient vector and the Hessian matrix. Then the solvers either use finite difference (FD) or algorithmic differentiation (AD) to compute the derivatives numerically.

For large size problems, say one thousand variables and more, one needs to code an ad-hoc type of nonlinear solver which conducts all computations including the search direction $d_q$, and the line search $\rho_q$. Typically, this is coded in C++ in conjunction with other supporting function libraries. A sequential quadratic program solver and a linear program solver are usually required if one does not want to start from scratch. Since this option requires more advanced expertise from users, which may not be an easy thing, most of the time, people can take advantage of the optimization routines provided by standard packages, e.g. MATLAB, GAUSS, SAS, etc.

2.3.3 Discussion

Variable Selection

In reality, we also have to decide what variables to include in the constrained logistic model (2.7). One can apply standard variable selection techniques: forward selection, backward selection, stepwise selection. Lasso [Tibshirani (1996)] has become a popular method in variable selection lately. For logistic regression models, one can use the ada-lasso logistic regression [Zou (2006)] method since it possesses the attractive oracle properties. Selected
variables are given by,

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} [-Y_i X_i' \beta + \log(1 + \exp(X_i' \beta))] + \lambda_n \sum_{j=1}^{k} \hat{w}_j |\beta_j|, \quad (2.8)$$

where $\hat{w}_j$ is the weight placed on the $j$th variable. Usually, $\hat{w}_j$ is set to be $1/|\hat{\beta}_j^{mle}|^\gamma$ for some $\gamma > 0$. $k$ is the total number of candidate variables. One needs to note that an implicit assumption for (2.8) to work as desired is that the set of candidate variables subsumes all the correct variables needed to be included in the model. Under this assumption then, the key is to correctly set the tuning parameter ($\lambda_n, \gamma$). This is typically done by a cross-validation based grid search.

In order to use those standard nonlinear constrained solvers for the ada-lasso regression, e.g. solver in MATLAB, GAUSS and SAS, one can apply the following standard transformation to avoid the non-differentiability issue that is caused by $||\theta||_1$, in (2.8). Let $\beta^+ = \max(\beta, 0)$ and $\beta^- = \max(-\beta, 0)$, then $\beta = \beta^+ - \beta^-$ and $|\beta| = \beta^+ + \beta^-$. Using this transformation, we have the following equivalent program (2.9) which is immediately solvable using standard nonlinear constrained solvers. To implement the gradient vector, one just needs to pay attention to the size change, i.e. from $k$ to $2k$ and implement accordingly.

$$\text{Min}_{\beta^+, \beta^-} \left\{ -\frac{1}{n} \ell(\beta^+ - \beta^-) + \frac{1}{n} \lambda_n \sum_{j=1}^{k} \hat{w}_j (\beta^+_j + \beta^-_j) \right\},$$

s.t. $\beta^+ \geq 0$ and $\beta^- \geq 0$. \quad (2.9)

Once the set of variables is selected, we can start to estimate the parameters for our constrained model (2.7).

**Initial Feasible Solution**

Another practical issue in computation is to find the initial feasible solution to the constrained logistic regression model (2.7) For unconstrained programs, due to strict convexity one can
basically select any convenient initial solution. This is not the case for constrained programs. Unless there exists a trivial feasible solution, one needs to provide an initial feasible solution to those standard nonlinear solvers. This problem is simplified when all the constraints are linear constraints. In this case, one can use a linear programming package to find an initial feasible solution.

**Improving numerical experience in MLE**

For the unconstrained case, and in the small sample case in particular, it often happens that there is a complete separation or quasi-complete separation of the data that causes the sample optimal solution to escape to infinity. Here, complete separation refers to the fact that there exists a $\beta^t$ such that for all $Y_i = 1$, $X_i \beta^t < 0$ and for all $Y_i = 0$, $X_i \beta^t > 0$ at the same time. If any of these two strict equalities contains cases for which one is close to having an equal sign, then the situation is referred to as quasi-complete separation. The consequence of complete separation or quasi-complete separation is that the parameter estimates escape to infinity under MLE. Another less common situation occurs when all the responses are the same, for example, for all $i$, $Y_i = 0$. Then it can be shown that parameter estimates also escape to infinity using MLE.

Using our constrained model (2.7), in both of the above situations, the numerical properties of constrained MLEs are improved by adding suitable 'lock-type' constraints. That is, we can lock the fitted event probabilities between upper and lower bounds such that $p_{low} \leq \hat{p}_i \leq p_{high}$ for some constants $p_{low}$ and $p_{high}$, $i \in I$, where $I$ is some index set. In the sequel, we also discuss 'order-type' constraints of the form $p_i \leq p_j$, $i, j \in I$ for some index set $I$.

### 2.4 Application to Corporate Bankruptcy Dataset

Needless to say, bankruptcy prediction is of great importance from both financial and social considerations. Prediction failure has resulted in an astronomical amount of financial loss to
society. Logistic regression is a popular method used for bankruptcy prediction, e.g. Altman (1968), Ohlson (1980) etc.

Typically in bankruptcy prediction data sets, it is usually the case that there are a very small number of companies suffering bankruptcy while the vast majority of companies do not. This poses practical problems when fitting the unconstrained logistic regression model since the parameter divergence problem will likely to be in evidence.

Another practical issue is that it is likely that the sample used to fit the unconstrained logistic regression model is biased in one way or the other: either having too many default companies or having too many non-default companies. This is so since applying the MLE method to the unconstrained logistic regression model with an intercept term entails $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{p}^{mle}_i$.

Researchers have identified various ways of partially dealing with these issues. See the Introduction section for references. But very little of the existing literature deals with how to incorporate useful or quality prior information on event probabilities in a systematic way into the parameter estimation process in the logistic regression context, the notable exception being the Bayesian methods. Without using useful or quality prior information, it is often the case that for a given data set, the model fitting results under MLE may be incongruent with existing useful or quality prior information.

When one is presented with useful prior information, e.g. Moody’s rating, S&P ratings, or expert advice, our constrained logistic regression model (2.7) can help improve the model fitting in a finite sample context by removing the restriction $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{p}^{mle}_i$ and also by constraining the model parameters in a way that is consistent with the useful or quality prior information.
To measure the performance, we compute the AUC for the ROC curve [Fawcett (2003)] on both the unconstrained and the constrained ML approaches. AUC can be computed as follows. Let $n_0$ be the number of observed events and $n_1$ be the number of observed non-events. Sort the event probability predictions ascendingly and assign ranks $r_j$ to those sorted predicted event probabilities. Then, by Hanley and McNeil (1982), AUC is equivalent to the Wilcoxon-Mann-Whitney statistic and can be computed by

$$AUC = \sum_{j \in \{j | y_j = 1\}} r_j - \frac{n_0(n_0 + 1)}{2n_0 n_1}.$$

### 2.4.1 Data Set from WRDS

In this study, we use the annual corporate financial information to conduct bankruptcy prediction for companies that have entries in the WRDS. In the selected data set, we also have 9 independent variables (including the intercept) and 1 binary response variable indicating default or bankruptcy. These 9 financial accounting related variables are based on Campbell et al. (2008) and other standard references. The selected data come from the Compustat/Annual Fundamentals/North America sub-database. They refer to companies in North America whose financial report information was collected by the WRDS over the years 1967 - 2007. Some companies had defaulted and some did not. For those companies for which no defaults have been observed, we used their most recent financial information in WRDS. For those companies that had defaulted, we used their financial information prior to the default in the WRDS. We use the information in its original scale without standardizing or rescaling. The independent variables are

- $X_1 =$ intercept,
- $X_2 =$ working capital/total assets,
- $X_3 =$ retained earnings/total assets,
- $X_4 =$ earnings before interest and taxes/total assets,
- $X_5 = \frac{\text{market value of equity}}{\text{book value/total liabilities}},$

- $X_6 = \frac{\text{sales}}{\text{total assets}},$

- $X_7 = \frac{\text{net income}}{\text{total assets}},$

- $X_8 = \frac{\text{total liabilities}}{\text{total assets}},$

- $X_9 = \frac{\text{current assets}}{\text{current liabilities}}$

The response variable $Y$ is 0 if no default occurred in the observed time window, and is 1 if there is a default by the end of the time window. We have only one record for each company in our data set. This makes our data cross-sectional as opposed to time series.

<table>
<thead>
<tr>
<th></th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
<th>$X_8$</th>
<th>$X_9$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>-4.72</td>
<td>-50.66</td>
<td>-4.44</td>
<td>15.51</td>
<td>1.64</td>
<td>-16.50</td>
<td>5.73</td>
<td>3.85</td>
<td>0.02</td>
</tr>
<tr>
<td>max</td>
<td>0.99</td>
<td>140.58</td>
<td>4.49</td>
<td>33894.80</td>
<td>5500</td>
<td>766</td>
<td>9788</td>
<td>2234</td>
<td>1</td>
</tr>
<tr>
<td>min</td>
<td>-9787</td>
<td>-134863</td>
<td>-23956.50</td>
<td>0.00</td>
<td>-0.38</td>
<td>-130077</td>
<td>0.00</td>
<td>-4</td>
<td>0</td>
</tr>
<tr>
<td>median</td>
<td>0.16</td>
<td>-0.16</td>
<td>0.06</td>
<td>2.12</td>
<td>0.81</td>
<td>-0.01</td>
<td>0.51</td>
<td>1.67</td>
<td>0</td>
</tr>
<tr>
<td>std</td>
<td>157.48</td>
<td>1486.70</td>
<td>239.77</td>
<td>337.72</td>
<td>52.47</td>
<td>1260.23</td>
<td>158.49</td>
<td>26.50</td>
<td>0.13</td>
</tr>
</tbody>
</table>

### 2.4.2 Parameter Constraints

We also use the credit rating data (Moody’s or S&P) for a subset of the selected companies to help in the model fitting. Intuitively, more highly rated companies have lower fitted default probability. This can help construct the order-type constraints. For lock-type constraints, one can rely on expert advice or other useful sources to constrain the default probability in a particular interval. In the following sampling and fitting process, we chose firms using a random sampling scheme and then applied the associated and appropriate constraints to those firms contained in the sample. Six of them are order-type constraints and six of them are lock-type constraints. We also update the constraints whenever we randomly choose a new training sample.
There is nothing magical about adding six order-type or lock-type constraints. We do it only for demonstration purpose. In reality, one can choose to add any arbitrary number of constraints. From a computational point of view, as long as the constraints satisfy the constraint qualifications listed previously, finding optimal solutions inevitably becomes easier. One can also choose to fix the set of constraints throughout if there is not sufficient prior information from which to sample. One can also choose to add all the constraints to express all of the existing useful information.

### 2.4.3 Time Window and Sample Selection Methods

1. **Time Window.**

The time window we have chosen in this study is the period from 1967 to 2007. We choose not to focus on a specific 10 year, 5 year or even 1 year time period since we want to build a model that does not have strong time dependence on the periods where the data arise. In other words, we let the data represent more of a cross-sectional data set. Part of our goal is to show the efficacy of the constrained estimation approach on an overall time span basis as opposed to any particular time window of specific length.

2. **Existing Sample Selection Methods.**

In the corporate bankruptcy prediction literature, for one reason or another, the number of firms that went bankrupt is relatively small when compared to the financially healthy firms. There are two ways of doing the sampling: choice-based data sampling and complete data sampling.

In choice-based data sampling, first fix the number of default or bankrupt companies. Then for each existing default or bankrupt company, a non-default or non-bankrupt company is either chosen completely at random or chosen from a certain preset pool of companies where there is a one-to-one match between the incumbent default or bankrupt company and the
newly chosen non-default and non-bankrupt company. Complete data sampling refers to the usual method of simple random sampling without replacement.

Both of these methods have their pros and cons. Choice-based sampling may introduce extra estimation bias in the estimated parameters while avoiding the problem embedded in the complete data sampling: parameter estimation may crash or escape to infinity due to a low number of default or bankrupt firms in the logistic regression model using the MLE method. See the relevant literature for a discussion, e.g. Zmijewski (1984). In our study, we take a revised version of the complete data sampling method to side step the problem of non-convergence of the estimated parameters. Our data and the sampling approach proceed as follows:

We did not use all the data available in 1967 - 2007 in the collected data set for model training. Instead, each time, we randomly chose some small size subset. Of course this approach highlights the strength of the constrained approach in the small sample size case. However, efforts have been made to ensure that the data in the training sample pool come from different business sectors. We let our training sample sizes range from 50 to 350. In the various sizes of training samples, we have ensured that the ratio of the number of default or bankrupt companies to the number of non-default or non-bankrupt companies reflects what realistically occurs in the overall population, i.e. fewer than 3% defaults. We also have credit rating information available for the observations in the fitting data set pool $\Omega_f$, $|\Omega_f| = 397$. This permits us to assign prior default probabilities to each observation in $\Omega_f$. The testing data set pool is $\Omega_t$ and $|\Omega_t| = 11059$. For each row in the testing data pool, we have the independent variables ($X_1$ to $X_9$) and the response variable indicating a default or a non-default associated with the independent variable observed at the same financial period. In the testing data set pool $\Omega_t$, we do not have rating information available. For each iteration we randomly select 50 to 350 observations from $\Omega_f$ to fit the model both with constraints
and without constraints. With the fitted parameter $\hat{\beta}^n_{\text{con}}$ and $\hat{\beta}^n_{\text{mle}}$, we randomly select 30 observations from $\Omega_t$ and evaluate the AUC. With the same $\hat{\beta}^n_{\text{con}}$ and $\hat{\beta}^n_{\text{mle}}$, we repeat the above procedure to obtain 5 AUC measures for both methods and record their respective averages. Overall, we repeat the process 1000 times in order to eliminate random errors.

### 2.4.4 Results

Table 2.2: Empirical results of the WRDS Dataset. Testing sample size is 30 for all fitting sample sizes. Number of replications made is 1000 for all fitting sample sizes. Percentage $\text{AUC}_{\text{con}} > \text{AUC}_{\text{mle}}$ column shows the percentages that the constrained approach has larger area under the ROC curve than the MLE approach.

<table>
<thead>
<tr>
<th>Fitting Sample Size</th>
<th>Percentage AUC$<em>{\text{con}} &gt;$ AUC$</em>{\text{mle}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>66.2</td>
</tr>
<tr>
<td>100</td>
<td>67.0</td>
</tr>
<tr>
<td>150</td>
<td>70.6</td>
</tr>
<tr>
<td>200</td>
<td>77.1</td>
</tr>
<tr>
<td>250</td>
<td>78.7</td>
</tr>
<tr>
<td>300</td>
<td>81.1</td>
</tr>
<tr>
<td>350</td>
<td>87.7</td>
</tr>
</tbody>
</table>

Figure 2.1 contains the detailed AUC statistics of the runs listed in Table 2.2. From Table 2.2, we can see that the constrained approach is definitely preferable to the unconstrained approach. Note that the AUC measure is not as ideal, i.e., as large as one typically expects. But it does show the improvement of the constrained approach over the unconstrained MLE approach. We also note that, in the comparison, we did not include any case where the MLE estimates fail to converge while the constrained estimates do converge. Our empirical experience indicates that it is common to see MLE estimates failing to converge while the constrained estimates do. As one increases the quality of the constraints, i.e. by adding more order-type constraints and adding tighter lock-type constraints, the constrained estimates tend to perform even better than the unconstrained MLE. This is not only conceivable but also confirmed in our empirical studies.
Figure 2.1: Area Under ROC for the results in Table 2.2. Mean is the sample average of the area under the ROC curve computed for the testing sample; Std Dev is the standard deviation of the area under the ROC curve computed for the testing sample. Each boxplot column represents the sampling picture of the area under the ROC curve under combinations of training sample size, i.e. 50, 100, 150, 200, 250, 300, 350, and estimation methods, i.e. constrained (CON) or MLE.
### 2.5 Simulation

In the simulated dataset, we have 9 independent variables (including the intercept) and 1 binary response variable. Each row vector is \( \mathbf{X}_i = [1 \ X_{i,2} \ X_{i,3} \ldots X_{i,9}] \) and the corresponding response variable is \( Y_i \). The independent variables \( \mathbf{X}_i \) are purposely scaled to have a similar range as those we pick from the WRDS dataset [Section 2.4]. The response variable \( Y_i \) is generated following independent Bernoulli distribution, \( Y_i \sim \text{Bernoulli}(p^*_i) \) with parameter \( p^*_i \) equal to

\[
p^*_i = \Pr(Y_i = 1|\mathbf{X}_i) = \frac{1}{1 + \exp(-\mathbf{X}_i^T \beta^*)},
\]

where the population level model parameter \( \beta^* \) is randomly selected over a trial and error process to ensure that the resulting average proportion of observations having \( Y_i = 1 \) is fewer than 6%.

We partition the data into a testing data set pool, denoted by \( \Omega_t \), \( |\Omega_t| = 1000 \) and a fitting data set pool, denoted by \( \Omega_f \), \( |\Omega_f| = 500 \). The empirical study procedure is described as follows: In each trial, we add six order-type constraints and six lock-type constraints:

**step 1.** take \( n_1 = n_1^* \in \{50, 100, 200, 300, 400\} \) rows randomly from \( \Omega_f \) and fit the models, and obtain \( \hat{\beta}^*_c \) for the constrained model and \( \hat{\beta}^*_m \) for the unconstrained model. If there are no events, i.e. \( Y = 1 \), in the \( n_1^* \) chosen rows, then we resample until there are enough events that can lead to \( \hat{\beta}^*_m \) convergence, since the comparison is of little meaning when the unconstrained approach fails to converge.

**step 2.** take \( n_2 = 30 \) rows within which there is at least one event instance from \( \Omega_t \) and do prediction using \( \hat{\beta}^*_c \) and \( \hat{\beta}^*_m \) in step 1 and collect the evaluation criteria: AUC for both models, \(|\hat{\beta}^*_c - \beta^*|_2, |\hat{\beta}^*_m - \beta^*|_2, \sum_{i=1}^{n_2} |\hat{p}^*_c - p^*_i|, \sum_{i=1}^{n_2} |\hat{p}^*_m - p^*_i|\).

**step 3.** fix \( \hat{\beta}^*_c \) and \( \hat{\beta}^*_m \) in step 1, repeat step 2 \( n_3 = 5 \) times, and record the mean of the AUC over \( n_3 = 5 \) times for both models, the mean of \(|\hat{\beta}^*_c - \beta^*|_2, \text{the mean of } |\hat{\beta}^*_m - \beta^*|_2, \text{the mean of } \sum_{i=1}^{n_2} |\hat{p}^*_c - p^*_i|, \text{and the mean of } \sum_{i=1}^{n_2} |\hat{p}^*_m - p^*_i|\).
step 4. repeat step 1 to step 3 \( n_4 = 1000 \) times.

In the simulated data set, we generated values of \( p_i \) from the uniform distributions over the intervals \([0, \psi]\), where \( \psi \in \{0.02, 0.05, 0.10, 0.50\}\).

Table 2.3: Empirical Results 1 of the Simulated Dataset. Each entry shows the percentage of times that the constrained approach has a higher area under the ROC curve result than the unconstrained approach. \( n_1^* \) is the training sample size. \( \max(p_i) \) is the highest event probability in all the samples, both training sample and testing sample.

<table>
<thead>
<tr>
<th>( \max(p_i) )</th>
<th>( n_1^* = 50 )</th>
<th>( n_1^* = 100 )</th>
<th>( n_1^* = 200 )</th>
<th>( n_1^* = 300 )</th>
<th>( n_1^* = 400 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \leq 0.02 )</td>
<td>67.7</td>
<td>52.6</td>
<td>69.4</td>
<td>53.3</td>
<td>77</td>
</tr>
<tr>
<td>( \leq 0.05 )</td>
<td>61.1</td>
<td>59.9</td>
<td>68</td>
<td>83.3</td>
<td>51.5</td>
</tr>
<tr>
<td>( \leq 0.1 )</td>
<td>64.8</td>
<td>63.3</td>
<td>54.5</td>
<td>66.8</td>
<td>52.5</td>
</tr>
<tr>
<td>( \leq 0.5 )</td>
<td>86.5</td>
<td>81.4</td>
<td>77.1</td>
<td>76.2</td>
<td>90.3</td>
</tr>
</tbody>
</table>

Table 2.4: Empirical Results 2 of the Simulated Dataset. Each entry shows the percentage of times that the mean of \( ||\tilde{\beta}_{con}^n - \beta^*||_2 \) for the constrained approach is smaller than the mean of \( ||\tilde{\beta}_{mle}^n - \beta^*||_2 \) for the unconstrained approach. \( n_1^* \) is the training sample size. \( \max(p_i) \) is the highest event probability in all the samples, both training sample and testing sample.

<table>
<thead>
<tr>
<th>( \max(p_i) )</th>
<th>( n_1^* = 50 )</th>
<th>( n_1^* = 100 )</th>
<th>( n_1^* = 200 )</th>
<th>( n_1^* = 300 )</th>
<th>( n_1^* = 400 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \leq 0.02 )</td>
<td>87.2</td>
<td>94.2</td>
<td>95.7</td>
<td>86.3</td>
<td>99.6</td>
</tr>
<tr>
<td>( \leq 0.05 )</td>
<td>95.8</td>
<td>98.8</td>
<td>97.4</td>
<td>95.2</td>
<td>99.9</td>
</tr>
<tr>
<td>( \leq 0.1 )</td>
<td>98.2</td>
<td>98.5</td>
<td>91.8</td>
<td>92.2</td>
<td>91.6</td>
</tr>
<tr>
<td>( \leq 0.5 )</td>
<td>99.7</td>
<td>98.6</td>
<td>98.3</td>
<td>99.3</td>
<td>99.2</td>
</tr>
</tbody>
</table>

For each \( \psi \), i.e. \( \max(p_i) \), we run simulations for different training sample size \( n_1^* \) and measure 1) the percentage of times that the constrained approach has a higher area under the ROC curve than the unconstrained approach, 2) the percentage of times that the mean of \( ||\tilde{\beta}_{con}^n - \beta^*||_2 \) for the constrained approach is smaller than the mean of \( ||\tilde{\beta}_{mle}^n - \beta^*||_2 \) for the unconstrained approach, 3) the percentage of times that the mean of \( \sum_{i=1}^{n_2} |\tilde{p}_{con}^i - p_i^*| \) for the constrained approach is smaller than the mean of \( \sum_{i=1}^{n_2} |\tilde{p}_{mle}^i - p_i^*| \) for the unconstrained approach.
Table 2.5: Empirical Results 3 of the Simulated Dataset. Each entry shows the percentage of times that the mean of $\sum_{i=1}^{n_i} |\hat{p}_i^{con} - p_i^*|$ for the constrained approach is smaller than the mean of $\sum_{i=1}^{n_i} |\hat{p}_i^{ml} - p_i^*|$ for the unconstrained approach. $n_i^*$ is the training sample size. $max(p_i)$ is the highest event probability in all the samples, both training sample and testing sample.

<table>
<thead>
<tr>
<th>$max(p_i)$</th>
<th>$n_i^* = 50$</th>
<th>$n_i^* = 100$</th>
<th>$n_i^* = 200$</th>
<th>$n_i^* = 300$</th>
<th>$n_i^* = 400$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq 0.02$</td>
<td>99.3</td>
<td>99.6</td>
<td>100</td>
<td>99.7</td>
<td>99.9</td>
</tr>
<tr>
<td>$\leq 0.05$</td>
<td>100</td>
<td>99.3</td>
<td>100</td>
<td>100</td>
<td>99.9</td>
</tr>
<tr>
<td>$\leq 0.1$</td>
<td>99.9</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>$\leq 0.5$</td>
<td>99.9</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.9</td>
</tr>
</tbody>
</table>

We choose to use the $AUC$, the area under the ROC curve, instead of other measures since it does not depend on any specific cut-off probability and it is also a good measure of the classification capability.

From Tables 2.3, 2.4 and 2.5, we can see that the constrained approach consistently performs better than the unconstrained approach in three different goodness-of-fit measures: $AUC$, $||\hat{\beta} - \beta^*||_2$ and $\sum |\hat{p}_i - p_i^*|$.

2.6 Large Sample Properties

In this section, we discuss the large sample properties of parameter estimates of the constrained logistic regression model (2.7), i.e. consistency and (conic) normality of the optimal solution of the constrained nonlinear optimization problem (2.7).

Recall that the loglikelihood function (2.1), $\ell(\beta, \omega_i) = -Y_iX_i\beta + \log(1 + \exp(X_i\beta))$, $\ell : \mathbb{R}^k \times \Omega \mapsto (0, \infty]$ is a random lower semicontinuous function in the sense that for all $\omega \in \Omega$, $\ell(\cdot, \omega)$ is lower semicontinuous in the first argument, and $\ell(\cdot, \cdot)$ is $\mathcal{F}^k \otimes \mathcal{G}$-measurable. Lower
semicontinuity follows from the continuity of $\ell$ in $\beta$.

It holds that for all $\beta \in \Theta$, $\omega \mapsto \ell(\beta, \omega)$ is continuous on $\Omega$. It also holds that $\ell$ is locally Lipschitz on $\Theta$ in the following sense: for each $\beta$ in $\Theta$, there exists a neighborhood $\mathcal{U}$ of $\beta$ and a finite constant $L$ such that for all $\beta' \in \mathcal{U} \cap \Theta$ and $\omega \in \Omega$, $|\ell(\beta, \omega) - \ell(\beta', \omega)| \leq L \|\beta - \beta'\|$.

One version of the proof showing that $\ell$ is locally Lipschitz under the logistic regression model context can be found in Gourieroux and Monfort (1981). A more general proof can be found in Ruszczynski (2006). Being locally Lipschitz means both locally lower Lipschitz and locally upper Lipschitz. In fact, it is only the locally lower Lipschitz part that is required for proving consistency of the parameter estimates of the constrained logistic regression model (2.7). However, the locally Lipschitz part, i.e. both locally lower Lipschitz and locally upper Lipschitz, is required for establishing asymptotic distributions of the parameter estimates of the constrained logistic regression model (2.7).

**Assumption for Consistency - A1.**

Let $\zeta = (\omega_i)_{i=1}^{\infty}$ be an i.i.d. sample from the sample space. For each $\beta \in \Theta$ and every $\epsilon > 0$, there exists a compact set $\Delta_\epsilon \subset \Omega$ such that for $n = 0, 1, 2, \ldots$, with $P = P^0(\cdot, \zeta)$,

$$
\int_{\Omega \setminus \Delta_\epsilon} |\ell(\beta, \omega)| P^n(d\omega, \zeta) < \epsilon.
$$

The above assumption means that the sequence of measures $\{P^n(\cdot, \zeta)\}_{n=0}^{\infty}$ is $\ell(\beta, \cdot)$-tight (asymptotic negligibility), see Dupacová and Wets (1988).

Define $E^n\ell(\beta) = \int_{\Omega} \ell(\beta, \omega) P^n(d\omega)$ and $E\ell(\beta) = \int_{\Omega} \ell(\beta, \omega) P(d\omega)$. Then, as a consequence of Assumption A1, $\mu$-almost surely, we have [Dupacová and Wets (1988)]

(i) the expectation functionals $E^n\ell$ are random lower semicontinuous functions; and
\( E\ell = \text{epi-lim} E^n\ell = \text{ptwse-lim} E^n\ell \).

**Strong Consistency**

**Theorem 6** (Wets (1991)). Under Assumption A1, there exists \( M_0 \in \mathcal{F} \) of measure 1 such that for all \( \zeta \in M_0 \), if \( \beta^* \) is a cluster point of any sequence \( \{\hat{\beta}_{\text{con}}^n\} \) with \( \hat{\beta}_{\text{con}}^n \in \text{argmin} E^n\ell(\cdot, \zeta) \) then \( \beta^* \in \text{argmin} E\ell(\cdot, \zeta) \).

Note that, unlike the conditions we need to derive the asymptotic distribution of the parameter estimate of the constrained logistic regression model (2.7), it is not necessary that \( \Theta \) be polyhedral to establish consistency.

In order to establish the asymptotic distribution of parameter estimates of the constrained logistic regression model (2.7), we rely on the following setup. Define \( \ell_\Theta(\beta) = \ell(\beta) + \Psi_\Theta(\beta) \), where \( \Psi_\Theta(\beta) = 0 \), if \( \beta \in \Theta \); \( \Psi_\Theta(\beta) = \infty \) if otherwise. Let \( \partial h(\alpha) = \{\nabla h(\alpha) | \alpha \in \mathbb{R}^k\} \), where \( h(\alpha) \) is continuously differentiable and \( \nabla h(\alpha) \) denotes the gradient vector, i.e. the vector of partial derivatives with respect to \( \alpha \). If \( \beta \in \Theta \), then \( \partial \Psi_\Theta(\beta) \) is the polar of the tangent cone to \( \Theta \) at \( \beta \), denoted by \( T_\Theta(\beta) \) [Ruszczynski (2006), Example 2.81].

At the population level, \( 0 \in \partial E\ell_\Theta(\beta^*) \) implies that, \( P \)-almost surely \( 0 = E\{\partial \ell(\beta^*, \omega)\} + v_\Theta(\beta^*) \), where \( v_\Theta(\beta^*) \in \partial \Psi_\Theta(\beta^*) \). At the sample level, \( 0 \in \partial E^n\ell_\Theta(\hat{\beta}_{\text{con}}^n) \) implies that \( P^n \)-almost surely \( 0 = E^n\{\partial \ell(\hat{\beta}_{\text{con}}^n, \omega)\} + v_\Theta(\hat{\beta}_{\text{con}}^n) \), where \( v_\Theta(\hat{\beta}_{\text{con}}^n) \in \partial \Psi_\Theta(\hat{\beta}_{\text{con}}^n) \).

**Assumption for Asymptotic Normality - A2.**

[Dupacová and Wets (1988)] We assume that for some \( v^n \in \partial E^n\ell_\Theta(\beta^*, \zeta) \) and \( v \in \partial E\ell(\hat{\beta}_{\text{con}}^n, \zeta) \), \( \sqrt{n}[v(\hat{\beta}_{\text{con}}^n(\zeta)) + v^n(\beta^*, \zeta)] \) converges to 0 in probability, and \( \sqrt{n}[v_\Theta(\hat{\beta}_{\text{con}}^n(\zeta)) - v_\Theta(\beta^*)] \) converges to 0 in probability.

Note that \( v(\hat{\beta}_{\text{con}}^n(\zeta)) + v^n(\beta^*, \zeta) \) deals with the sample level asymptotic behavior of the sum.
of the gradient vectors evaluated at sample optima and the gradient vectors evaluated at the population optima. \( v_\Theta(\hat{\beta}_{con}^n(\zeta)) - v_\Theta(\beta^*) \) deals with difference between the gradient vector of indicator functions evaluated at the sample optima and the population optima. Combined with the consistency of \( \hat{\beta}_{con}^n \), these two conditions basically ensure the smoothness of the convergence.

Since we are dealing with the logistic regression model, we have that \( \sqrt{n}v^n(\beta^*, \zeta) \) is asymptotically Gaussian with mean vector \( 0 \) and covariance matrix \( K^{-1} \), where \( K \) is the Fisher information matrix. We also have \( E\ell \) is \( C^2 \) at \( \beta^* \) with Hessian matrix, positive definite, equal to the Fisher information matrix \( K \).

**Asymptotic Normality**

**Theorem 7** (Wets (1991)). Under Assumption A1 and A2, \( \Theta \) polyhedral,

\[
\sqrt{n}(\hat{\beta}_{con}^n - \beta^*) \Rightarrow N(0, K^{-1}) .
\]

where \( K \) is the Fisher information matrix.

**Assumption for Asymptotic Conic Normality - A3.**

Assume that \( \Theta \) is polyhedral, also assume that

\[
E\|\nabla_\Theta(\beta^*, \omega)\|^2 < \infty, \quad \text{and} \quad \omega \rightarrow \sup_{\beta^1, \beta^2 \in \Theta} \frac{|\nabla_\Theta(\beta_1^1, \omega) - \nabla_\Theta(\beta_2^1, \omega)|}{|\beta_1^1 - \beta_2^1|} \in L^2 .
\]

**Asymptotic Conic Normality**

Let \( \theta \) be a normally distributed random vector with distribution \( N(0, K) \), where \( K \) is positive definite and is indeed the Fisher information matrix as in Assumption A2.

**Theorem 8** (King (1988)). Under Assumption A3, \( \sqrt{n}(\hat{\beta}_{con}^n - \beta^*) \) is asymptotically conically
The asymptotic distribution is the distribution of the vector $\tilde{\theta}$,

$$
\tilde{\theta} = \arg\min_{u \in \tilde{\Theta}} (u - K^{-1} \theta)' K (u - K^{-1} \theta),
$$

where $\tilde{\Theta}$ is the cone, $\tilde{\Theta} = \{ u \in T_\Theta(\beta^*) \mid E \nabla \ell(\beta^*, \omega) u = 0 \}$, and $T_\Theta(\beta^*)$ is the tangent cone to $\Theta$ at $\beta^*$ and $\theta \sim N(0, K)$, with the Fisher information matrix $K$ as variance.

One can also find results of a similar nature in King and Rockafellar (1993) and Geyer (1994).

2.7 Conclusion

In this paper, we proposed a constrained M-estimation approach to improve prediction accuracy in binary logistic regression models. When there is useful information available on event probabilities, we can improve the prediction results over unconstrained logistic regression models by adding parameter constraints consistent with existing quality information. We have provided various large sample properties of this constrained logistic regression approach. We applied this constrained logistic regression approach on a simulated data set and a bankruptcy prediction data set. Empirical findings support that our constrained logistic regression model is useful in practical settings and evidently superior at prediction when compared to MLE, based on several metrics.

In the future, we will consider including the variable selection as a model component in the constrained M-estimation approach. We will also consider finding the asymptotic bias in AIC when using the constrained M-estimation approach. It would also be helpful to include statistical inference under the constrained settings as well. Another study may focus on robustness issues, i.e. in cases where constraints are chosen but are not appropriate.

Moreover, it seems desirable to include the idea of conditional value at risk, $CVaR$, as part of
the estimation process. By finding the set of parameters that minimizes $CVaR$, we should be able to incorporate the idea of loss given default into the constrained parameter estimation process.
Chapter 3

On the statistical characteristics of coverage optimization based on sample mean approach

3.1 Introduction

In this paper, we are concerned with a binary integer linear program called the coverage optimization problem shown in (3.2). To make matters more concrete, consider the following magazine covering problem. Suppose that a marketing company must decide which $m$, magazines out of a total of $M$ magazines to market in a certain area. To avoid triviality, it is assumed that $1 \leq m < M$. Let $N$ be the number of households in the area, which is often assumed to be unknown but fixed. In practical settings, $N$ is usually on the order of a hundred thousand or more. The objective is to have as many households covered as possible. In order to define being covered, we first define the $j$th subscription pattern, denoted by $G_j$, as the $j$th element in a set $G$ defined in (3.1) that contains all such possible subscription patterns.

$$G_j \in G = \{(g_1, g_2, ..., g_M)|g_k \in \{0, 1\}, 1 \leq k \leq M\} \quad (3.1)$$
Here, \( g_k = 1, 1 \leq k \leq M \), means that magazine \( k \) belongs to some subscription pattern indexed by \( j \). Note that this does not mean that magazine \( k \) is chosen in the final \( m \) magazines. Also, \( g_k = 0, 1 \leq k \leq M \), indicates otherwise. In other words, a subscription pattern is a subset with distinct elements formed by these \( M \) magazines. Therefore, the total number of different subscription patterns, denoted by \( T \), is \( 2^M \), i.e. \(|G| = T = 2^M\).

We assume that the set \( G \) is lexicographically ordered. Let \( \mathbf{x} \) be a row vector of size \( M \), representing the set of magazines chosen to market. Let \( \mathbf{x}_i \) be the \( i \)th coordinate of \( \mathbf{x} \), \( \mathbf{x}_i \in \{0,1\}, 1 \leq i \leq M \). The interpretation is that magazine \( j, 1 \leq j \leq M \), is chosen if \( \mathbf{x}_j = 1 \). It is not chosen if \( \mathbf{x}_j = 0 \). Define \( \mathbf{x} \cap G_j = \emptyset \) if

\[
\bigcup_{k=1}^{M}\{k \mid \mathbf{x}_k = 1 \text{ and } G_{j,k} = 1, G_j \in G \} = \emptyset
\]

where \( G_{j,k} \) denotes the \( k \)th coordinate of \( G_j \). The \( j \)th subscription pattern in \( G \) is said to be covered by \( \mathbf{x} \) if \( \mathbf{x} \cap G_j \neq \emptyset \). Let \( \mathbf{y} \) be a column vector of size \( T \) with \( k \)th coordinate \( \mathbf{y}_k \in \{0,1\}, 1 \leq k \leq T \). The interpretation is that the \( k \)th subscription pattern in \( G \) is covered if \( \mathbf{y}_k = 1 \). It is not covered if \( \mathbf{y}_k = 0 \). Let \( \mathbf{c} \) be a row vector of size \( T \) with \( j \)th coordinate representing the number of households having subscription pattern \( G_j \) in \( G \). Each household has one and only one subscription pattern in \( G \). Assume that \( \mathbf{c}_j \in \{0\} \cup \mathbb{N} \). Let \( A \) be a matrix of size \( T \times M \) whose rows are \( G_j \in G, 1 \leq j \leq T \). Denote by \( \mathcal{G} \) the following set of vectors,

\[
\mathcal{G} = \{(\mathbf{x}, \mathbf{y})| A\mathbf{x}^T \geq \mathbf{y} \text{ and } \mathbf{1}_M\mathbf{x}^T = m \text{ and } \mathbf{x} \in \{0,1\}^M, \mathbf{y} \in \{0,1\}^T\}
\]

where \((\cdot)^T\) denotes matrix transpose. \( \mathbf{1}_M \) is a column vector of \( 1 \) of size \( M \) and of whose entries is \( 1 \). \( \mathcal{G} \) is a finite but often very large set that makes full enumeration impractical.
We have the following general formulation, denoted by Problem $\mathcal{P}_{N,1}$.

$$\text{Problem } \mathcal{P}_{N,1} : \quad \text{Maximize } \sum_{(x,y) \in \mathcal{S}} cy \quad (3.2)$$

When solving $\mathcal{P}_{N,1}$, the following relaxation $\mathcal{S}_r$ is used instead of $\mathcal{S}$ and can be shown to have the same optimal solution.

$$\mathcal{S}_r = \{ (x,y) | Ax^r \geq y \text{ and } 1_Mx^r = m \text{ and } x \in \{0,1\}^M, \ y \in [0,1]^T \}$$

Note that the difference between $\mathcal{S}$ and $\mathcal{S}_r$ is that the binary set $\{0,1\}^T$ in $\mathcal{S}$ is relaxed to the interval $[0,1]^T$ in $\mathcal{S}_r$.

Problem $\mathcal{P}_{N,1}$ is a variant of the maximal covering problem (MCP) that was first formulated as a mixed integer problem by Church and ReVelle (1974). In the literature, the MCP problem has wide industrial applications. A partial list of various kinds of applications can be found in Sweeney et al. (1979), Dwyer and Evans (1981), Bennett et al. (1982), Eaton et al. (1985), Chung (1986), Pirkul and Schilling (1988), Current and O’Kelly (1992), Schilling et al. (1993), Adenso-diaz and Rodriguez (1997) and the references therein.

In reality, it is typically impractical or impossible to take a census of the entire area, consisting of $N$ households, to determine the $c$ vector in Problem $\mathcal{P}_{N,1}$. Usually, a random sample of $c$, containing subscription pattern information for $n$ households, denoted by $\hat{c}_n$, is taken instead. Let $\hat{c}_{n,j}$ denote the $j$th coordinate of $\hat{c}_n$, $1 \leq j \leq T$. $\hat{c}_{n,j}$ is comprised of the number of households whose subscription pattern is $G_j \in G$, $1 \leq j \leq T$, as given in (3.1). The resulting sample based optimization problem, denoted by Problem $\mathcal{P}_{n,1}$, is then solved. The set of optimal solutions is denoted $\text{Argmax } \mathcal{P}_{n,1}$. Similarly, the set of optimal solutions to Problem $\mathcal{P}_{N,1}$ is denoted by $\text{Argmax } \mathcal{P}_{N,1}$. It is often of interest to use the estimated probability of coverage, denoted by $p_{n,1}$, to estimate the true coverage probability, denoted by
\( p_{N,1} \), as in (3.3), respectively.

\[
p_{n,1} = \frac{\hat{c}_ny_1^*}{n} \quad \text{and} \quad p_{N,1} = \frac{cy_2^*}{N}
\] (3.3)

for \( y_1^* \in \text{Argmax} \mathcal{P}_{n,1} \) and \( y_2^* \in \text{Argmax} \mathcal{P}_{N,1} \). In (3.3), it is also interesting to note that, for \( n \) fixed, \( \hat{c}_n/n \) is the UMVUE of \( c/N \).

In this paper, we will follow Cochran (1997) and Cochran et al. (2008) to focus on how to estimate \( p_{N,1} \) as defined in (3.3) using sample based solutions.

### 3.2 Our Setup

In order to better estimate \( p_{N,1} \), we propose to solve the following optimization problem, denoted by Problem \( \mathcal{P}_{n,t} \),

\[
\text{Problem } \mathcal{P}_{n,t} : \quad \text{Maximize } \frac{1}{t} \sum_{j=1}^{t} \frac{\hat{c}_j}{n} y
\]

where in Problem \( \mathcal{P}_{n,t} \), \( \hat{c}_j \) denotes the subscription pattern compiled from the \( j \)th collected set of sample data, \( 1 \leq j \leq t \). \( t \) can be interpreted as the total number of time periods over which data are collected, \( t \in \mathbb{N} \) and \( t \geq 1 \). But other interpretations are possible. In time period \( j \), sample data on \( n \) households in the area are collected yielding information on subscription patterns defined in (3.1). The results are compiled into \( \hat{c}_n \), a row vector of size \( T \). Throughout, we assume that over time periods, the number of magazines under consideration, \( M \), is constant. We also assume that peoples’ subscription patterns in the area are constant. This implies that the vector \( c \) is also constant over times, by our assumption.

To start, we first set up the following framework. Let \( (\Xi_n, \mathcal{A}_n, P_{\xi_n}) \) be the probability
space, where $\Xi_n$ is the support of the probability measure $P_{\xi_n}$ and

$$\Xi_n = \{ (\xi_{n,1}, \xi_{n,2}, \ldots, \xi_{n,T}) \mid \sum_{j=1}^{T} \xi_{n,j} = n, \xi_{n,j} \in \{0\} \cup \mathbb{N}, 1 \leq j \leq T \}$$

$\mathfrak{A}_n$ is the $\sigma$-field relative to $\Xi_n$. In fact $\mathfrak{A}_n = 2^{\Xi_n}$, the power set on the elements of $\Xi_n$. $P_{\xi_n}$ corresponds to the multinomial probability distribution function, a suitable approximation for large $M$. That is, for a sample vector $\hat{c}_n \in \Xi$,

$$P(\hat{c}_{n,1} = \xi_{n,1}, \hat{c}_{n,2} = \xi_{n,2}, \ldots, \hat{c}_{n,T} = \xi_{n,T}) = \frac{n!}{\xi_{n,1}!\xi_{n,2}! \cdots \xi_{n,T}!} p_{\xi_{n,1}}^{\xi_{n,1}} p_{\xi_{n,2}}^{\xi_{n,2}} \cdots p_{\xi_{n,T}}^{\xi_{n,T}}$$

where $[p_1, p_2, \ldots, p_T] = c/N$. $n \ll N \ll T$. Define a function $f(\cdot, \cdot): \{0, 1\}^T \times [0, n]^T \mapsto [0, 1]$ as follows:

$$f(y, \xi_n) = \frac{1}{n} \xi_n y$$

where $\xi_n$ is a row vector of size $T$ and $\xi_n \in \Xi_n \subset [0, n]^T$. $y$, as defined before, is the binary column vector of size $T$. For each $y \in \{0, 1\}^T$, the random function $f(y, \xi_n)$ is measurable in $\xi_n$ with respect to $\mathfrak{A}_n$. Further, define the random function $\vartheta(\cdot): [0, n]^T \mapsto [0, 1]$ as,

$$\vartheta(\xi_n) = \max_{(x,y) \in \Theta} f(y, \xi_n)$$

Define $\mathbb{P}_{n,t}\vartheta$ and $P_n\vartheta$ as,

$$\mathbb{P}_{n,t}\vartheta = \frac{1}{t} \sum_{j=1}^{t} \vartheta(\xi_{n,j}^i), \quad P_{n,1}\vartheta = \int_{\Xi_n} \vartheta(\xi_n) P_{\xi_n}(d\xi_n)$$

Let the sample average of $\xi_{n,1}^i, \xi_{n,2}^i, \ldots, \xi_{n,T}^i$ be denoted by $\bar{\xi}_n$. Then Problem $\mathcal{P}_{n,t}$ is just $\vartheta(\bar{\xi}_n)$. Note that $\mathbb{P}_{n,t}\vartheta$ is NOT $\vartheta(\bar{\xi}_n)$. Denote by $(\Xi_{n,t}, \mathfrak{A}_{n,t}, P_{\xi_n})$ the probability space for $\bar{\xi}_n$ induced
by the sample average operation, i.e. \( \bar{\xi}_n = (1/t) \sum_{i=1}^t \xi_{ni} \). Define \( P_{n,t} \vartheta \) as

\[
P_{n,t} \vartheta = \int_{\Xi_{n,t}} \vartheta(\bar{\xi}_n) P_{\xi_n}(d\bar{\xi}_n)
\]

Under this setup, Problem \( \mathcal{P}_{n,t} \) can be treated as the sample average approximation of the following stochastic program, denoted by Problem \( \mathcal{P}_0 \),

\[
\text{Problem } \mathcal{P}_0 : \quad \text{Maximize } \quad E_{P_{\xi_n}} f(y, \xi_n)
\]

It is easy to see that Problem \( \mathcal{P}_0 \) is equivalent to Problem \( \mathcal{P}_{N,1} \).

### 3.3 Basic Properties

**Theorem 9.** \( \vartheta(\xi_n) = \max_{(x,y) \in \mathcal{Y}} f(y, \xi_n) \) is convex in \( \xi_n \).

Proof of Theorem 9 is deferred to the appendix. By Jensen inequality, it follows immediately that,

**Corollary 10.** \( E[\vartheta(\xi_n)] \geq \vartheta(E[\xi_n]) \)

Since \( P_{n,1} = E[\vartheta(\xi_n)] \) and \( \vartheta(E[\xi_n]) = p_{N,1} \), Corollary 10 says that there exists nonnegative bias in using the expected value of \( \vartheta(\xi_n) \) to estimate \( p_{N,1} \). [cf. Cochran (1997), Norkin et al. (1998), Mak et al. (1999) and Kleywegt et al. (2001)]

**Theorem 11.** Given fixed \( n \), for \( t \geq 1 \), the following hold:

a. \( p_{N,1} \leq P_{n,t} \vartheta \leq P_{n,1} \vartheta \);

b. \( P_{n,t+1} \vartheta \leq P_{n,t} \vartheta \).

Proof of Theorem 11 is deferred to the appendix.

By Corollary 10 and Theorem 11, we can see the justification for adopting the sample approximation approach to solve Problem \( \mathcal{P}_{n,t} \) since it offers less bias when using the optimal
objective function value to estimate the true coverage probability, \(p_{N,1}\) in (3.3). Later in Section 3.4.1, we will show that as \(t \to \infty\), the optimal objective function value of the Problem \(\mathcal{P}_{n,t}\) converges almost surely to \(p_{N,1}\).

On the other hand, Theorem 11 is also useful in addressing the question, "Is \(P_{n+1,1} \leq P_{n,1}\)?" In other words, by Theorem 11, we know that \(P_{n,1}\) has non-negative bias when used to estimate the true coverage probability. So, the question is, does the non-negative bias shrink when we add the additional information of one household? This is obvious when using part b of Theorem 11 which says that \(P_{1,t+1} \leq P_{1,t}\) since \(P_{1,t+1} = P_{n+1,1}\) and \(P_{1,t} = P_{n,1}\) when \(t = n\).

### 3.4 Some Asymptotics

In this section, we give some asymptotic properties of the sample average approach.

#### 3.4.1 Main Approximation

**Theorem 12.** The optimal objective function value of Problem \(\mathcal{P}_{n,t}\) is strongly consistent in estimating \(p_{N,1}\). That is, for fixed \(n\), with an i.i.d. sample \(\hat{c}_{n,j}^i\), \(1 \leq j \leq t\),

\[
\text{Maximize} \quad \left[ \lim_{t \to \infty} \frac{1}{nt} \sum_{j=1}^{t} \hat{c}_{n,j}^i \right] y = p_{N,1} \quad \text{a.s.}
\]

Proof of Theorem 12 is given in the appendix.

In order to investigate the asymptotic distribution of the optimal objective function value of Problem \(\mathcal{P}_{n,t}\), we note that since \(\hat{c}_{n,j}^i\), \(1 \leq j \leq t\), are i.i.d. multinomial with

\[
E(\hat{c}_{n,j}^i) = \frac{nc_i}{N} \quad \text{and} \quad V(\hat{c}_{n,j}^i) = \Sigma,
\]

\[
\Sigma_{i,i} = n(c_i/N)(1 - c_i/N) \quad \text{and} \quad \Sigma_{i,j} = -n(c_i/N)(c_j/N), \quad \text{if} \quad i \neq j.
\]

Since \(\sum_{i=1}^{T} c_i/N = 1\), \(\Sigma\) can
be shown to be less than full rank and therefore positive semi-definite. Denote the sample average of $\hat{\mathbf{c}}^j_n$, $1 \leq j \leq t$, by $\bar{\mathbf{c}}_n$. Therefore, for $n$ fixed, as $t \to \infty$, by C.L.T., we have,

$$\sqrt{t} \left( \bar{\mathbf{c}}_n - \frac{n \mathbf{c}}{N} \right) \Rightarrow_d N(0, \Sigma)$$

Note that the objective function of Problem $\mathcal{P}_{n,t}$ is a function of $\bar{\mathbf{c}}_n$, namely, $\vartheta(\bar{\mathbf{c}}_n)$. Therefore, for $n$ fixed, as $t \to \infty$, by the Delta method, we have,

**Theorem 13.** The asymptotic distribution of the optimal objective function value of Problem $\mathcal{P}_{n,t}$ is,

$$\sqrt{t} \left( \vartheta(\bar{\mathbf{c}}_n) - p_{N,1} \right) \Rightarrow_d N \left( 0, \frac{1}{n^2} \mathbf{y}_*^\top \Sigma \mathbf{y}_* \right)$$

for unique $\mathbf{y}_* \in \text{Argmax} \mathcal{P}_{N,1}$. Consequently, the optimal objective function value of Problem $\mathcal{P}_{n,t}$ is asymptotically unbiased and root-$t$ weakly consistent, That is,

$$\vartheta(\bar{\mathbf{c}}_n) - p_{N,1} = O_p(t^{-1/2}).$$

Proof of Theorem 13 is given in the appendix.

Note that there is a critical assumption in Theorem 13, i.e. a unique optimal solution exists for Problem $\mathcal{P}_0$. In order to allow for multiple optimal solutions to the Problem $\mathcal{P}_0$, among other things, we need the notion of Hadamard directional differentiability and the Functional Delta Method [van der Vaart (1998)].

Now, we handle the case where there are multiple optimal solutions.

**Theorem 14.** When $|\text{Argmax} \mathcal{P}_{N,1}| > 1$, i.e. there exists multiple optimal solutions to Problem $\mathcal{P}_{n,t}$, the asymptotic distribution of the optimal objective function value of Problem $\mathcal{P}_{n,t}$ is:

$$\sqrt{t} \left( \vartheta(\bar{\mathbf{c}}_n) - p_{N,1} \right) \Rightarrow_d \vartheta_{\mathbf{y}_*}^{\mathbb{P}} \left( N(0, \Sigma) \right)$$
where $\theta'_{nc}(N(0, \Sigma))$ is the directional derivative of the function $\theta(\cdot)$ at $nc/N$ for a random direction $N(0, \Sigma)$.

Proof of Theorem 14 is given in the appendix.

### 3.4.2 Other Approximations

We next present some asymptotic results related to other types of approximations. For i.i.d. $\xi_n^i$ and $\xi_n \in \Xi_n$. It is easy to see that $E(\theta(\xi_n)) = P_{n,1} \theta \leq 1$ and $E(\theta^2(\xi_n)) \leq 1$. Therefore, since $\theta(\xi_n^i)$ are i.i.d. and having finite first and second moment, by S.L.L.N., it follows that for $n$ fixed,

$$P_{n,t} \theta = \frac{1}{t} \sum_{j=1}^{t} \theta(\xi_n^i) \to \int_{\Xi_n} \theta(\xi) P_{\xi_n} (d\xi) = P_{n,1} \theta, \quad \text{as } t \to \infty, \quad a.s.$$ 

Also by C.L.T., for $n$ fixed,

$$\sqrt{t}(P_{n,t} \theta - P_{n,1} \theta) \Rightarrow_d N(0, V(\theta(\xi_n)))$$

where $V(\theta(\xi_n)) = E(\theta^2(\xi_n)) - E^2(\theta(\xi_n))$. So, root-t weak consistency follows immediately since the limiting distribution is $O_p(1)$. That is, for $n$ fixed,

$$P_{n,t} \theta - P_{n,1} \theta = O_p(t^{-1/2})$$

Define $P_{n,t,q} \theta$ as follows,

$$P_{n,t,q} \theta = \frac{1}{q} \sum_{j=1}^{q} \theta(\xi_n^i)$$

where $\xi_n^i$ belongs to the probability space $(\Xi_n, \mathcal{A}_n, P_{\xi_n})$. By the same argument as above but with more tedious notation, we have the following theorem (we omit the proof since the results follow from S.L.L.N and C.L.T.)
Theorem 15. For \( n, t \) fixed,

\[
P_{n,t,q} \vartheta = \frac{1}{q} \sum_{j=1}^{q} \vartheta(\xi_j^n) \to \int_{\Xi_{n,t}} \vartheta(\xi_n) P_{\xi_n}(d\xi_n) = P_{n,t} \vartheta, \quad \text{as } q \to \infty, \quad a.s.
\]

\[
\sqrt{q}(P_{n,t,q} \vartheta - P_{n,t} \vartheta) \Rightarrow_d N(0, V(\vartheta(\xi_n))) \quad \text{and} \quad P_{n,t,q} \vartheta - P_{n,t} \vartheta = O_p(q^{-1/2})
\]

### 3.5 Discussion

In the previous section, we established the weak convergence and consistency results for Problem \( \mathcal{P}_{n,t} \) and other related asymptotic results when using other approaches. Our sample mean approximation approach produces asymptotically unbiased estimation results which contrasts with other approaches that can not.

### 3.6 Empirical Study

In this section, we use the data set from Dwyer and Evans (1981) to illustrate the effectiveness of our proposed approach in section 3.2. Specifically, our proposed approach produces smaller bias than the approach in which one takes the average of the optimal objective function values in each period.

There are 2241 households in the data set, 2131 of them have certain non-null magazine subscription patterns. To begin with, for each number of magazines chosen, \( m, m \in \{1, 2, \ldots, 74\} \), we first solve the overall problem. That is, we use the subscription pattern of all the 2131 households to form the constraints and objective function coefficients and then solve the problem \( \mathcal{P}_{2131,1} \), (see (3.2)).

Next, we fix the number of time periods \( t \) to be 12. It is not practical to have a large number of time periods since magazine subscription patterns may change over time. That is,
we use \( t = 12 \) to represent a conservative time span where the population-wise subscription patterns can be assumed to be constant. For each time period, we take random samples of size \( n = 300 \) households from the \( N = 2131 \) households, and form the problem \( \mathcal{P}_{300,1} \). Note that, for any time period, according to our sampling plan, there is no duplicated household in the 300 households. However, over different time periods, there exist households being sampled more than once. After solving the problem \( \mathcal{P}_{300,1} \) individually 12 times with 300 households each time, we aggregate the data from the 12 time periods to form the problem \( \mathcal{P}_{300,12} \) and solve it once.

After the one trial of 12 time periods followed by solving the problem 12 times individually and then solving it once collectively as described the above, we repeat the entire trial 10 times. Overall, for each \( m \) fixed, after repeating the process 10 times, we end up with 120 individual estimates: \( p_{300,1} \) and 10 aggregate estimates: \( p_{300,12} \). In the boxplots in the attachments, for each \( m \) fixed, we plot the distribution of \( p_{2131,1}, p_{300,1} \) and \( p_{300,12} \).

From the boxplots in the appendix, we can see that our proposed approach does show smaller estimation bias. And we conclude that our approach is better.

### 3.7 Future Work

In this section, we propose the problems,

\[
\chi_1(\theta) = \min_{(x,y) \in \mathbb{S}} \sum_{j=1}^{t} |f(y, \xi_{n}^{j}) - \theta|
\]

and

\[
\chi_2(\theta) = \min_{(x,y) \in \mathbb{S}} \sum_{j=1}^{t} \left(f(y, \xi_{n}^{j}) - \theta\right)^2
\]

where \( \chi_1(\theta) \) can be interpreted as minimizing the mean absolute difference between a fixed coverage probability \( \theta \) and \( \chi_2(\theta) \) can be interpreted as minimizing the mean square error.
Initial empirical investigation indicates that for \( i = 1, 2 \), with a large probability, \( p_{2131,1} \) is contained in the interval,

\[
[0, \sup \{\theta | \chi_i(\theta) = 0\}],
\]

It is not clear yet which choice of \( \chi_1 \) or \( \chi_2 \) yields better estimates than our current aggregated one, \( p_{n,t} \). This deserves more research efforts.

We propose the application of the sample average approach to solving the MCP over very large population. Specifically, one might randomly (or otherwise) split the population into \( t \) subpopulations and solve the MCP as we proposed here. Note that instead of interpreting \( t \) as the number of \( t \) periods as previously suggested, here one sees that \( t \) is the number of subpopulations. Further investigation of this idea is planned.

### 3.8 Conclusion

In this paper, we propose a sample average approximation based approach to solve the MCP problem. We have established asymptotic results and various other properties of this sample average approximation based approach.
Bibliography


Appendices
Appendix A

Appendix - Proofs

A.1 Proof of Theorem 9

Take $\xi^i_n$ and $\xi^j_n \in [0, n]^T$. Also take $\alpha \in [0, 1]$. We have $\alpha \xi^i_n + (1 - \alpha) \xi^j_n \in [0, n]^T$ and

\[
\vartheta(\alpha \xi^i_n + (1 - \alpha) \xi^j_n) = \max_{(x, y) \in \mathcal{S}} f(y, \alpha \xi^i_n + (1 - \alpha) \xi^j_n) = \max_{(x, y) \in \mathcal{S}} \frac{1}{n} \left( \alpha \xi^i_n + (1 - \alpha) \xi^j_n \right) y
\]

\[
\leq \max_{(x, y) \in \mathcal{S}} \frac{1}{n} \alpha \xi^i_n y + \max_{(x, y) \in \mathcal{S}} \frac{1}{n} (1 - \alpha) \xi^j_n y = \alpha \vartheta(\xi^i_n) + (1 - \alpha) \vartheta(\xi^j_n) \quad \square
\]

A.2 Proof of Theorem 11

In part b, observe the following identity: for any $a_i \in \mathbb{R}$, $1 \leq i \leq t + 1$, $t \in \mathbb{N}$, $t \geq 1$,

\[
\frac{1}{t + 1} \sum_{i=1}^{t+1} a_i = \frac{1}{t + 1} \sum_{i=1}^{t+1} \sum_{j=1, j \neq i} a_j
\]
Part a:

\[ p_{N,1} = \max \frac{cy}{N} = \max \left( E\left[ \frac{1}{n} \xi_n \right] \right) y = \max \left( \frac{1}{t} \sum_{j=1}^{t} E\left[ \frac{1}{n} \xi_n^j \right] \right) y \]

\[ = \max \left( E\left[ \frac{1}{t} \sum_{j=1}^{t} \frac{1}{n} \xi_n^j \right] \right) y \]

\[ \leq E \left[ \max \left( \frac{1}{t} \sum_{j=1}^{t} \frac{1}{n} \xi_n^j \right) \right] y = P_{n,t} \theta \quad \text{by Corollary 10} \]

\[ \leq E \left[ \frac{1}{t} \sum_{j=1}^{t} E \left[ \max \left( \frac{1}{n} \xi_n^j \right) \right] y \right] \quad \text{by Convexity, Theorem 9} \]

\[ = \frac{1}{t} \sum_{j=1}^{t} E \left[ \max \left( \frac{1}{n} \xi_n^j \right) y \right] = P_{n,1} \theta \quad \square \]

Part b:

\[ P_{n,t+1} \theta = E \left[ \max \left( \frac{1}{t+1} \sum_{j=1}^{t+1} \xi_n^j \right) y \right] = E \left[ \max \left( \frac{1}{t+1} \sum_{i=1}^{t+1} \sum_{j=1, j \neq i}^{t+1} \frac{\xi_n^j}{n} \right) y \right] \]

\[ = E \left[ \frac{1}{t+1} \max \left( \frac{1}{t+1} \sum_{i=1}^{t+1} \sum_{j=1, j \neq i}^{t+1} \frac{\xi_n^j}{n} \right) y \right] \]

\[ \leq E \left[ \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{t} \max \left( \sum_{j=1, j \neq i}^{t+1} \frac{\xi_n^j}{n} \right) y \right] \quad \text{by Convexity, Theorem 9} \]

\[ = \frac{1}{t+1} \sum_{i=1}^{t+1} E \left[ \max \left( \frac{1}{t} \sum_{j=1, j \neq i}^{t+1} \frac{\xi_n^j}{n} \right) y \right] \]

\[ = \frac{1}{t+1} \sum_{i=1}^{t+1} P_{n,t} \theta = P_{n,t} \theta \quad \square \]

A.3 Proof of Theorem 12

For i.i.d. \( \xi_n^j \) with finite first moment, by S.L.L.N., we have, for fixed \( n \),

\[ \lim_{t \to \infty} \frac{1}{nt} \sum_{j=1}^{t} \xi_n^j = \frac{c}{N} \quad \text{a.s.} \]
Then for each \( y \) fixed, we have,
\[
\left[ \lim_{t \to \infty} \frac{1}{nt} \sum_{j=1}^{t} \hat{c}_j \right] y = \frac{c}{N} y \quad \text{a.s.} 
\]

Therefore, by Problem \( \mathcal{P}_{n,\infty} \) and Problem \( \mathcal{P}_{N,1} \) having the same feasible region \( \mathcal{S} \) or \( \mathcal{S}_r \), if \( y^* \) (may not be unique) solves Problem \( \mathcal{P}_{n,\infty} \), then it must also solve Problem \( \mathcal{P}_{N,1} \), and vice versa. Hence the conclusion follows:

\[
\text{Maximize}_{(x,y) \in \mathcal{S}} \left[ \lim_{t \to \infty} \frac{1}{nt} \sum_{j=1}^{t} \hat{c}_j \right] y = p_{N,1} \quad \text{a.s.} \quad \square
\]

### A.4 Proof of Theorem 13

First, note that by Danskin’s Theorem [Bertsekas (2003)],
\[
\nabla \vartheta(\hat{c}_n) = \frac{d}{d \hat{c}_n} \vartheta(\hat{c}_n) = \frac{1}{n} y^* \quad \text{and} \quad \vartheta\left(\frac{nc}{N}\right) = p_{N,1},
\]
where \( y^* \) is assumed to be unique and maximizes \( \vartheta(\hat{c}_n) \). When \( \nabla \vartheta(\cdot) \) is evaluated at \( nc/N \), we have \( \nabla \vartheta(nc/N) = y_*/n \) where \( y_* \in \text{Argmax} \ \mathcal{P}_{N,1} \). Therefore, by the Delta Method [Shao (2003)], we have:
\[
\sqrt{t} \left( \vartheta(\frac{nc}{N}) - p_{N,1} \right) \Rightarrow_d N \left( 0, [\nabla \vartheta(\frac{nc}{N})]^T \Sigma [\nabla \vartheta(\frac{nc}{N})] \right)
\]

After simplifying the right-hand-side, we have
\[
\sqrt{t} \left( \vartheta(\frac{nc}{N}) - p_{N,1} \right) \Rightarrow_d N \left( 0, \frac{1}{n^2} y_*^T \Sigma y_* \right)
\]
for some \( y_* \in \text{Argmax} \ \mathcal{P}_{N,1} \). So, the asymptotic normality is proven.

Since \( N(0, y_*^T \Sigma y_* / n^2) \) is \( O_p(1) \), \( \vartheta(\hat{c}_n) - p_{N,1} = O_p(t^{-1/2}) \) \quad \square
A.5 Proof of Theorem 14

By Danskin’s Theorem [Bertsekas (2003)], for each direction $h \in [0, n]^{T}$, take the directional derivative $\vartheta'_{nc/N}(h)$ to be

$$
\vartheta'_{nc/N}(h) = \max_{y^* \in \text{Argmax } \mathcal{P}_{N,1}} \frac{h y^*}{n} .
$$

Since $\vartheta'_{nc/N}(h)$ is a map continuous in $h$, it is easy to see that:

$$
\| \frac{\vartheta(nc/N + th_t) - \vartheta(nc/N)}{t} - \vartheta'_{nc/N}(h) \| \to 0, \quad \text{as } t \downarrow 0 \text{ and every } h_t \to h.
$$

Hadamard differentiability of $\vartheta(\cdot)$ at $nc/N$ is established. Then by the Functional Delta Method [van der Vaart (1998)], we have:

$$
\sqrt{t} \left( \vartheta(\overline{c}_n) - p_{N,1} \right) \Rightarrow_{d} \vartheta'_{wc}(N(0, \Sigma))
$$

where $\vartheta'_{wc}(N(0, \Sigma))$ is the directional derivative of function $\vartheta(\cdot)$ at $nc/N$ for a random direction $N(0, \Sigma)$. 

95
Appendix B

Appendix - Some Facts

B.1 Directional Derivative

[Bertsekas (2003)] For any $y \in \mathbb{R}^n$, we define the one-sided directional derivative of $f$ in the direction $y$, to be

$$f'(x; y) = \lim_{\alpha \downarrow 0} \frac{f(x + \alpha y) - f(x)}{\alpha}$$

provided that this limit exists.

If the directional derivative of $f$ at a vector $x$ exists in all directions $y$ and $f'(x; y)$ is a linear function of $y$, we say that $f$ is differentiable at $x$. It is seen that $f$ is differentiable at $x$ if and only if the gradient $\nabla f(x)$ exists and satisfies

$$[\nabla f(x)]' y = f'(x; y) \quad \forall x \in \mathbb{R}^n$$

B.2 Danskin’s Theorem

[Bertsekas (2003)]. Let $Z$ be a compact subset of $\mathbb{R}^m$ and let $\phi : \mathbb{R}^n \times Z \mapsto \mathbb{R}$ be continuous and such that $\phi(\cdot, z) : \mathbb{R}^n \mapsto \mathbb{R}$ is convex for each $z \in Z$
(a) The function $f : \mathbb{R}^n \mapsto \mathbb{R}$ given by

$$f(x) = \max_{z \in Z} \phi(x, z) \quad (B.1)$$

is convex and has directional derivative given by

$$f'(x; y) = \max_{z \in Z(x)} \phi'(x, z; y)$$

where $\phi'(x, z; y)$ is the directional derivative of the function $\phi(\cdot, z)$ at $x$ in the direction $y$, and $Z(x)$ is the set of maximizing points in (B.1)

$$Z(x) = \{ z \mid \phi(x, z) = \max_{z \in Z} \phi(x, z) \}$$

In particular, if $Z(x)$ consists of a unique point $z$ and $\phi(\cdot, z)$ is differentiable at $x$, then $f$ is differentiable at $x$, and $\nabla f(x) = \nabla_x \phi(x, z)$, where $\nabla_x \phi(x, z)$ is the vector with components

$$\frac{\partial \phi(x, z)}{\partial x_i}, \quad i = 1, 2, \ldots, n.$$

**B.3 Hadamard Differentiable**

[vander Vaart (1998)]. Let $\mathbb{D}$ and $\mathbb{E}$ be normed linear spaces. A map $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$, defined on a subset $\mathbb{D}_\phi$ of a normed space $\mathbb{D}$ that contains $\theta$, is called Hadamard differentiable at $\theta$ if there exists a continuous, linear map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ such that

$$\| \frac{\phi(\theta + th_t) - \phi(\theta)}{t} - \phi'_\theta \|_{\mathbb{E}} \rightarrow 0, \quad \text{as } t \downarrow 0 \text{ and every } h_t \rightarrow h.$$ 

More precisely, for every $h_t \rightarrow h$ such that $\theta + th_t$ is contained in the domain of $\phi$ for all small $t > 0$. This definition as given requires that $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ exists as a map on the whole of $\mathbb{D}$. If this is not the case, but $\phi'_\theta$ exists on a subset $\mathbb{D}_0$ and the sequences $h_t \rightarrow h$ are restricted to...
converge to limits $h \in \mathbb{D}_0$, then $\phi$ is called Hadamard differentiable \textit{tangentially} to this subset.

B.4 Functional Delta Method

\cite{van der Vaart (1998)}. Let $\mathbb{D}$ and $\mathbb{E}$ be normed linear spaces. Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ by Hadamard differentiable at $\theta$ tangentially to $\mathbb{D}_0$. Let $T_n : \Omega_n \mapsto D_\phi$ be maps such that $r_n(T_n - \theta) \Rightarrow_d T$ for some sequence of numbers $r_n \to \infty$ and a random element $T$ that takes its values in $\mathbb{D}_0$. Then

$$r_n (\phi(T_n) - \phi(\theta)) \Rightarrow_d \phi'_\theta(T)$$

if $\phi'_\theta$ is defined and continuous on the whole space $\mathbb{D}$, the we also have

$$r_n (\phi(T_n) - \phi(\theta)) = \phi'_\theta (r_n(T_n - \theta)) + o_p(1)$$
Appendix C

Appendix - Graphs

In this section, we include the graphs. Please see the 1st graph for detailed explanations. The rest are similar.
Figure C.1: $m = 1$
Figure C.2: $m = 5$

Figure C.3: $m = 10$

Figure C.4: $m = 15$
Figure C.5: $m = 20$

Figure C.6: $m = 25$

Figure C.7: $m = 30$
Figure C.8: $m = 35$

Figure C.9: $m = 40$

Figure C.10: $m = 45$
Figure C.11: m = 50

Figure C.12: m = 55

Figure C.13: m = 60
Figure C.14: $m = 65$