Methods for Correcting Selection Bias in Non-Probability Samples via Data Integration

Lyndon John Ang

A thesis submitted for the degree of

Doctor of Philosophy of The Australian National University



Australian National University

Research School of Finance, Actuarial Studies and Statistics

College of Business and Economics

November 2026

© Copyright by Lyndon John Ang, 2026

All Rights Reserved

Originality Statement

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at the ANU or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at the ANU or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Lyndon John Ang
25 November 2026

Copyright Statement

I hereby grant The Australian National University or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or hereafter known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.

Lyndon John Ang
25 November 2026

Authenticity Statement

I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

Lyndon John Ang
25 November 2026

Acknowledgements

The saying "*It takes a village to raise a child*" holds very true for this thesis, which would not have seen the light of day without the support of many people. Firstly, I would like to express my gratitude to my supervisory chair Associate Professor Robert Clark for being willing to take me on as his doctoral student. His encouragement and sound advice over the last few years has been invaluable on this learning journey. I would also like to express gratitude to my co-supervisor Dr Bronwyn Loong. Her kindness, advice, encouragement and investment of time have been appreciated as I have navigated this PhD.

A big thanks to Dr Anders Holmberg, who as Australian Bureau of Statistics (ABS) Chief Methodologist supported my case to apply for the chance to complete this PhD research, and then agreed to become one of my co-supervisors. His insights and encouragement to connect my research to practical applications within the ABS have been valuable for keeping my research grounded in reality. It was at his suggestion that I was able to attend and present at three international conferences during my studies.

I express sincere thanks to my thesis examiners for their time, careful reading of this work, and constructive feedback, which helped to refine and strengthen the final version of this thesis.

While undertaking this PhD research, I have been fortunate to meet and have discussions with a number of fellow survey methodologists, and these have helped to shape this thesis into its final form. My thanks to Dr Arnout van Delden at Statistics Netherlands for agreeing to host a visit from me early on in my research. It was on this visit that I was also able to meet Dr An-Chiao (Anne) Liu and Dr Sander Scholtus, two others working on similar research topics to mine. Anne was kind enough to ask if I would like to contribute to one of her papers, and so began my first international collaboration. The Statistical Sciences Research Institute (S3RI) at the University of Southampton hosted me for a day of discussions as part of my first overseas trip. Towards the end of my PhD study I had the opportunity to spend three weeks at the University of Waterloo working with Professor Changbao Wu. I appreciated his hospitality and the insights I gained from discussions with him. While at Waterloo, it was also my pleasure to meet Professor Pengfei Li, to whom I am grateful for sharing some R code for a method he and Professor Wu co-developed with their PhD student.

I am grateful to the ABS for their endorsement of my PhD research topic, and for their financial support during my PhD study. The Sir Roland Wilson Foundation provided me with funding to undertake my international travel while studying. I am particularly grateful to the SRW Foundation team - Bec, Meg, Liz, Dunkan, Georgina and Cassii - for their advice and support for all the Foundation scholars. To my fellow SRW scholars,

a big note of appreciation for your companionship, advice, and sharing of ideas. I have learnt so much from you all. I am grateful especially to my fellow scholars from the 2022 cohort who helped me navigate the change from public servant to student.

I would like to thank my APS mentor Dr Phillip Gould, as well as my previous mentor Teresa Dickinson, for their guidance and support. A big thanks to Dr Siu-Ming Tam, former ABS Chief Methodologist, for bringing me on board with some of his research a number of years ago. This was a catalyst for me to scratch this "PhD itch". Siu-Ming has also remained interested in my research progress, and for this I am grateful. Thank you also to my ABS colleagues: in particular, to my previous ABS supervisor Justin who supported me to apply for this PhD program, to Jack who kept me connected with the ABS while away, to Joseph who provided me with good advice on the PhD research process, and to James and Ryan for good discussions on related topics.

The ANU Research School of Finance, Actuarial Studies and Statistics (RSFAS) provided me with an environment to progress my research. I am grateful for the academics and fellow students who made my time at the ANU much more colourful than it would otherwise have been.

Finally, I wish to thank my family, who have been such a bedrock of support on this incredible journey. To my lovely wife Lily and my children Daniel and Michelle for their support and for sharing this journey with me, to my parents and parents-in-law for their help and assistance at various points in the last few years, and to my brother Russell and his family for their support - a big thank you.

Statement of Contributions

A version of Chapter 2 has been published in the Journal of Official Statistics. A shortened version of Chapter 3 was presented at the 7th International Conference on Establishment Statistics, while a more detailed version has been submitted to an international journal. Part of Chapter 4 is included in the Proceedings of the 2024 Statistics Canada International Methodology Symposium. A version of Chapter 5 has been submitted to an international journal.

The papers have been co-authored with my supervisors. My contributions include generating the original ideas, deriving theoretical results, implementing new methods in R, performing simulations and real data analysis to assess the performance of the new methods, writing initial manuscript drafts, undertaking revisions with advice from my supervisors, and finalising submissions for publication.

R code for the simulation studies in Chapters 3 and 4 can be found at https://github.com/skrawk/np-OE-inference, while code for the simulation studies in Chapter 5 is at https://github.com/skrawk/np-errorprone-aux.

List of Publications

Ang, L., R. Clark, B. Loong, and A. Holmberg. 2026. "Inference for Non-Probability Samples Using Error-Prone Auxiliary Information". In Review.

Ang, L., R. Clark, B. Loong, and A. Holmberg. 2026. "Overlap-Excluded Estimation of Participation Probabilities for Non-Probability Samples". In Review.

Ang, L., R. Clark, B. Loong, and A. Holmberg. 2025. "An Empirical Comparison of Methods to Produce Business Statistics Using Non-Probability Data". *Journal of Official Statistics*, 1–32. https://doi.org/10.1177/0282423X241298243.

Ang, L., R. Clark, B. Loong, and A. Holmberg. 2024a. "Estimating Propensities of Selection for Big Datasets via Data Integration". Presentation at the 7th International Conference on Establishment Statistics, Glasgow, United Kingdom.

Holmberg, A., L. Ang, R. Clark, and B. Loong. 2025. "Propensity Score Estimation and Optimal Sampling Design when Integrating Probability Samples with Non-probability Data". *Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics*. https://www150.statcan.gc.ca/n1/pub/11-522-x/2025001/article/00032-eng.pdf.

Abstract

There is a growing trend among statistical agencies to explore non-probability data sources for producing more timely and detailed statistics, while reducing costs and respondent burden. Selection bias and measurement error are two issues that may be present in such data. These imperfections may be corrected using available information relating to the population of interest, such as a census or a reference probability sample. This dissertation explores methods to correct selection bias and measurement error in non-probability datasets, in particular when it is possible to integrate these datasets with an existing reference probability sample.

The thesis begins with an examination of currently available methods for correcting selection bias through a realistic simulation study. The comparison is undertaken in a business survey context, where the non-probability sample is a large portion of the population. The study explores the effects on the performance of estimators in the case of nonignorable non-response and measurement error in the target variable within the non-probability sample. When missingness in the non-probability sample is ignorable and there is no measurement error, the methods examined are shown to be effective at removing selection bias. When nonignorable missingness and measurement error are present, however, it is found that a model-assisted estimator based on the reference sample provides a safe approach for making statistical inference using non-probability samples.

Next, a new approach is presented to estimate participation probabilities for the non-probability sample, when it is possible to identify the overlap between the non-probability and reference samples. In this approach, a pseudo-likelihood is maximised where the non-probability sample is combined with the reference sample units that do not also exist in the non-probability sample. A variant of the estimator is also introduced which may be employed when it is not possible to identify the overlapping units between the two samples. The proposed approach is efficient compared with existing pseudo-likelihood approaches, especially when the non-probability sample is large, and it is effective under different reference sample designs.

Using the proposed approach for estimating participation probabilities, the optimal sample design for the reference sample is explored. Details of the optimal selection probabilities are derived, and an approach for estimating these selection probabilities in practice is described. A split-population approach is then explored for the reference sample when it is possible to identify which units in the population are not on the non-probability sample. The use of optimal selection probabilities or the split-population approach enables sample size savings for the reference sample.

Finally, the more challenging situation when some auxiliary variables in the reference

sample are prone to error is explored. An approach is presented for simultaneously addressing selection bias in the non-probability sample alongside measurement error in some auxiliary variables on the reference sample. It is shown how the framework may also be applied in the case when there is nonignorable selection in the non-probability dataset. The proposed approach is shown to effectively address these biases through a simulation study and real-world data application.

The dissertation closes with a summary and ideas for future research.

Contents

Acknowledgements iii

Statement of Contributions v

Abstract vi

Contents viii

List of Figures xi

List of Tables xii

1 Introduction 1

1.1 Setup and Notation 4

1.2 Existing methods for non-probability datasets 7

1.2.1 Weighting approaches 9

1.2.2 Imputation-based approaches 12

1.2.3 Doubly Robust estimation 14

1.3 Thesis outline 16

2 Empirical Comparison of Existing Methods 19

2.1 Alternative sampling frameworks for A 20

2.1.1 Multiple frame approach 20

2.1.2 Cut-off sampling 21

2.2 Setup for the simulation study 23

2.2.1 Creation of the synthetic population 24

2.2.2 Probability sample designs 28

2.2.3 Estimators examined 30

2.3 Simulation study results . . . 30
2.3.1 Single-frame design results . . . 34
2.3.2 Dual-frame design results . . . 38
2.3.3 Cut-off design results . . . 39
2.3.4 Big data only results . . . 41
2.4 Concluding remarks . . . 42

3 Overlap-Excluded Participation Probabilities 47
3.1 Introduction . . . 47
3.2 Overlap-Excluded Propensity Score Estimation . . . 50
3.3 Consistency and variance of the OE IPW estimator . . . 55
3.3.1 Comparison of variance against alternative methods . . . 56
3.4 OE propensity score estimation when data integration is not possible . . . 64
3.5 Simulation Study . . . 67
3.5.1 Simulation Results . . . 69
3.6 Real Data Application . . . 71
3.7 Concluding Remarks . . . 76

4 Optimal Sample Design for A 79
4.1 Introduction . . . 79
4.2 Optimal selection probabilities for OE IPW estimator . . . 81
4.2.1 Poisson sampling . . . 83
4.2.2 Probability Proportional to Size Sampling with Fixed Sample Size . . . 85
4.2.3 Stratified Simple Random Sampling . . . 87
4.2.4 Optimal sample design in practice . . . 90
4.3 Split-Population Sample Design for A . . . 91
4.4 Simulation Study . . . 93
4.5 Concluding remarks . . . 95

5 Addressing Measurement Error and Selection-Not-At-Random 97
5.1 Introduction . . . 97
5.1.1 Measurement Error . . . 98
5.1.2 Selection-Not-At-Random . . . 100
5.1.3 Work undertaken in this chapter . . . 102

5.2 Basic Setup . . . 103
5.3 Proposed Method . . . 105
5.4 Consistency and Variance Estimation . . . 111
5.4.1 Pseudo-population bootstrap for variance estimation . . . 114
5.5 Simulation Study . . . 116
5.5.1 Measurement error simulation study . . . 116
5.5.2 SNAR simulation study . . . 123
5.6 Real-data Application . . . 125
5.7 Covariate measurement error in B . . . 129
5.8 Concluding Remarks . . . 131

6 Concluding remarks **135**
6.1 Summary and Conclusions . . . 135
6.2 Future Directions . . . 139

A ABS DataLab Disclaimer **143**

B Proofs and regularity conditions **145**
B.1 Regularity Conditions for OE Participation Probabilities . . . 145
B.2 Consistency and Variance Derivations for OE IPW estimator . . . 146
B.3 Variance of the Kim and Wang (2019) estimator . . . 152
B.4 Common variance expression for Hájek-like IPW estimators under Poisson sampling for A . . . 156
B.5 Regularity conditions for split-population OE participation probabilities . . . 157
B.6 Consistency of split-population OE IPW estimator . . . 158
B.7 Derivation of kernel score function for the SNAR case . . . 160
B.8 Conditions for consistency of the kernel score function . . . 164
B.9 Proof of consistency of the survey-weighted kernel score function . . . 165

List of Figures

1.1 Domains within the population $\mathcal{U}$ 5

5.1 Binned scatterplot of Log(wages) vs Log(wages with ME) 127

List of Tables

1.1 Data structures 8

2.1 Parameter values fed into Vale-Maurelli process - Employment 25
2.2 Industry Wage Factors 26
2.3 Variance terms for wages data item 27
2.4 Summary of estimators compared - single frame sample 31
2.5 Summary of estimators compared - dual frame sample 32
2.6 Summary of estimators compared - cut-off sample 32
2.7 Summary of estimators compared - Big data and frame only 32
2.8 Summary of estimators - Measurement error corrected estimators 33
2.9 Sample sizes under different designs 33
2.10 Best-performing estimators under different scenarios 34
2.11 Monte Carlo Bias and RRMSE of estimators based on 2000 samples - SAR, No Measurement Error 35
2.12 Monte Carlo Bias and RRMSE of estimators based on 2000 samples - SAR, With Measurement Error 37
2.13 Monte Carlo Bias and RRMSE of estimators based on 2000 samples - SNAR, No Measurement Error 40
2.14 Monte Carlo Bias and RRMSE of estimators based on 2000 samples - SNAR, With Measurement Error 41

3.1 Comparative performance of estimators - $\rho(y, \boldsymbol{x}^T\beta) = 0.3$ 69
3.2 Comparative performance of estimators - $\rho(y, \boldsymbol{x}^T\beta) = 0.7$ 69
3.3 Coverage Probabilities (CP) and Average Length (AL) of 95% Confidence Intervals for OE plug-in variance estimator 71
3.4 Coverage Probabilities (CP) and Average Length (AL) of 95% Confidence Intervals - Bootstrap variance for OE-EM approach 71
3.5 Population distribution across state 72

3.6 Selected quantiles for turnover and wages 72
3.7 Population distribution across industry division 73
3.8 Real-data performance of estimators for turnover 75

4.1 Performance of different sample designs for A - $\rho(y, \boldsymbol{x}^T\beta) = 0.3$ 94
4.2 Performance of different sample designs for A - $\rho(y, \boldsymbol{x}^T\beta) = 0.7$ 94

5.1 Comparison of estimators under measurement error - Case 1 119
5.2 Comparison of estimators under measurement error - Case 2 120
5.3 Comparison of estimators under measurement error - Case 3 120
5.4 Comparison of estimators under measurement error - Case 4 121
5.5 Performance of variance estimators under measurement error 122
5.6 Comparison of estimators under SNAR 125
5.7 Performance of variance estimators under SNAR 125
5.8 Performance of proposed method using real-world data 128

List of Notation

This list provides a description of some of the commonly used notation within this dissertation.

$\boldsymbol{\theta}$, $\boldsymbol{\phi}$	Vector of parameters related to a participation probability model
$\boldsymbol{q}$	Vector of design variables for the probability sample
$\boldsymbol{x}$	Vector of correctly measured auxiliary variables
$\boldsymbol{x}^*$	Vector of auxiliary variables with measurement error
$\boldsymbol{z}$	Vector of auxiliary variables not subject to measurement error
δ_i	Indicator variable denoting inclusion in B
$\mathcal{U}$	The population of interest
μ_y	Population mean of the variable of interest y
π_i^A	Reference sample inclusion probability for unit i
π_i^B	Participation probability for unit i to be included in the non-probability sample
A	A reference probability sample
B	A non-probability sample
c_i	Cost of collection for unit i in the population
d_i^A	Reference sample design weight for unit i
I_i	Indicator variable denoting inclusion in A
K_h	A kernel function with bandwidth h
N	The size of the population
n_A	The size of the probability sample
N_B	The size of the non-probability sample
y	Variable of interest that we want to obtain statistics for

Chapter 1

Introduction

Probability survey samples have been, for the best part of a century, the preferred data collection tool of statistical agencies for producing population estimates. Neyman (1934) laid the groundwork for design-based probability sampling theory, and the theory for estimating finite population quantities using probability survey samples is now well-established. Compared with censuses, probability samples provide less costly, more timely, and more detailed information on populations of interest. Methods have been developed to take advantage of additional information from other sources such as administrative datasets to improve the efficiency of estimates from probability samples (see, for example, Särndal et al. 1992).

In today's environment, the same reasons that drove the shift from censuses to probability samples are pushing statistical agencies to move away from probability samples and explore alternative data sources (Beaumont 2020). Running a high-quality probability sample is costly, users have a desire for more timely statistics at greater levels of detail, and there is growing expectation for statistical agencies to reduce respondent burden. For example, the Australian Bureau of Statistics (ABS) priorities for 2022-2023 included increasing the use of non-survey data sources to produce new, timely statistics and reduce burden on small and medium enterprises (Australian Bureau of Statistics 2022b). Statistics Canada has launched a modernisation initiative with five pillars, one of which involves the use of new data sources, lowering respondent burden, greater reliance on data integration and modelling, and reducing the role of surveys (Statistics Canada 2018).

These new data sources may include big data or "found data" such as sensor data, satellite imagery and transactions data, and non-probability survey samples such as online web panels. The data sources may have full coverage of the population, or may only include information from a fraction of the population. In this dissertation the term non-probability dataset is applied to any dataset where only part of the population is included, and the probability of inclusion into the dataset is unknown.

The potential benefits of non-probability data for producing statistics have been recognised, for example in Tam and Clarke (2015), Rao and Fuller (2017), and Tillé et al. (2022). In particular, these datasets promise to address the weaknesses of probability survey samples:

- Data collection can be much cheaper since the data is already captured through an existing process (generally for some purpose independent of producing statistics)
- Outputs are more timely
- Utilising existing data reduces the need for statistical agencies to ask people or businesses for the same or similar information

Unfortunately, non-probability data sources may suffer from quality issues. Two types of error pervading these data are selection bias and measurement error. Selection bias occurs when particular segments of a population have a lower propensity of inclusion in the data compared to other groups. Meng (2018) showed that selection bias in a non-probability sample can have a significant impact on the quality of an estimate. In fact, the error can get bigger the larger the dataset is, such that it may be preferable to use a small probability sample than a large dataset containing selection bias.

In this dissertation, the term measurement error will refer broadly to any misalignment between the true value of a target concept and the value that is actually captured by the sample. Measurement error may thus arise due to issues such as differences in concept definitions, recording or instrument errors, and so on. For example, a business may report their revenue in whole dollars on a survey form which is asking for reporting to be done in thousands of dollars.

In recent years, there has been a growing body of research on how best to address the

shortcomings of non-probability data, in particular the selection bias problem. Generally, the methods developed require that additional information from the population is available and can be used. This auxiliary information is used to correct for selection bias through a weighting process or to form models to impute for the missing units in the population, and could include known or estimated population totals, unit level information from a population register or frame, or data from an independent probability survey sample (which will be referred to as a *reference sample*) taken from the same population.

The use of auxiliary information taken from a reference sample, while the non-probability sample collects the variable(s) of interest, may be contrasted with the usual design-based, model-assisted and model-based survey scenarios. In these cases, the variables of interest are collected by the probability sample, while the frame contains the auxiliary information which is usually obtained from administrative data.

This dissertation takes up the baton of research in this exciting and fast-moving field. It particularly focusses on methods that are applicable in the business survey context. Business surveys generally have a number of characteristics that will influence the choice and performance of an estimation approach; see Hidiroglou and Lavallée (2009) for a good overview of these characteristics.

For instance, there tends to be a larger quantity of auxiliary information available on a business population frame, including items that are well-correlated with the data items of interest, for example through administrative data from taxation agencies. This means that available non-probability datasets sourced from administrative sources will tend to contain a sizeable proportion of the population. Business data items often also have skewed distributions, compared with the categorical data more often collected in population surveys. Their sample designs tend to be highly stratified, with large variations in the probabilities of selection. In particular, the largest contributors tend to be included with a probability of 1. Finally, there is often a reliable identifier (such as an official business number from the tax system) available, allowing different data sources to be linked. These characteristics will influence the data available from any reference sample taken from a business population.

This dissertation contributes to the existing literature in several ways. It first examines

the performance of existing solutions for correcting selection bias under the effects of measurement error and nonignorable selection bias through a simulation study. A more efficient method of inference for non-probability samples is then proposed in the situation where it is possible to identify membership of non-probability sample units on the reference sample. The problem of optimal reference sample design for estimates based on the non-probability dataset is then explored in the context of this new estimator. Finally, this thesis explores the issues of covariate measurement error in the reference sample and non-ignorable selection issues in the non-probability data, and proposes an approach to address these problems. The proposed solutions are tested empirically through simulation studies and real-world business data applications.

The rest of this chapter proceeds as follows. Section 1.1 provides an overview of the general setup and notation that will be used in this thesis. Section 1.2 provides a brief overview of existing methods of inference using non-probability samples. Finally, Section 1.3 provides an outline of the content of the rest of this thesis.

1.1 Setup and Notation

The basic setup and notation used in this thesis are as follows. For each unit i in a finite population $\mathcal{U}$ of size N, we have values $(\boldsymbol{x}_i, y_i)^T$ for a variable of interest y and some additional auxiliary data items $\boldsymbol{x}$. This thesis is primarily concerned with estimating either the population total $Y = \sum_{i=1}^{N} y_i$ or population mean $\bar{Y} = \mu_y = N^{-1} \sum_{i=1}^{N} y_i$. On occasion the set of auxiliary data items will be extended to include $\boldsymbol{z}_i$, where these may be variables associated with a probability sample design, or auxiliary data with a specific attribute (for example, not subject to measurement error).

A probability sample A is available, which will be referred to as the "reference" sample. A is a representative sample drawn from the population $\mathcal{U}$ and has a sample size n_A. A may have been collected for a different purpose than we are interested in, and so may contain information about $\boldsymbol{x}$ only. In other cases A may include both y and $\boldsymbol{x}$. The data collected in A are generally considered to be obtained without error, for example through careful planning and questionnaire wording, although this assumption is relaxed in Chapter 5. Define $I_i = I(i \in A)$ to be an indicator variable for unit i being included in the sample

A. Let $\pi_i^A = P(i \in A) = P(I_i = 1|\boldsymbol{q}_i)$ be the inclusion probability for unit i being in the probability sample, where $\boldsymbol{q}_i$ is the vector of design variables governing inclusion into A. Design weights for $i \in A$ are given as the inverse of the inclusion probabilities, $d_i^A = 1/\pi_i^A$. The π_i^A values are known from the sample design.

The non-probability sample is denoted by B. Like the probability sample, B may contain information on $\boldsymbol{x}$ only, for example if B is an administrative dataset collecting general population information. Alternatively B may include both $\boldsymbol{x}$ and y, for example if it is a web panel survey collecting data on the variables of interest. The full set of auxiliary variables $\boldsymbol{x}$ will generally be assumed to be available on both A and B. Let $\delta_i = I(i \in B)$ be an indicator variable for unit i being included in the sample B. The non-probability sample size is $N_B = \sum_{i=1}^N \delta_i$, $Y_B = \sum_{i=1}^N \delta_i y_i$ is the sum of the y values in the non-probability dataset, and $\boldsymbol{X}_B = \sum_{i=1}^N \delta_i \boldsymbol{x}_i$ is the sum of the $\boldsymbol{x}$ variables in the non-probability dataset. In contrast to π_i^A in the reference probability sample, the inclusion probabilities $\pi_i^B = P(\delta_i = 1|\boldsymbol{x}_i, y_i)$ are unknown and need to be estimated. In this dissertation the π_i^B will generally be referred to as participation probabilities or propensity scores.

Define $C = \mathcal{U} \setminus B$, the units in the population not included in B. There may or may not be overlap between the two samples A and B. Figure 1.1 depicts these domains within $\mathcal{U}$.

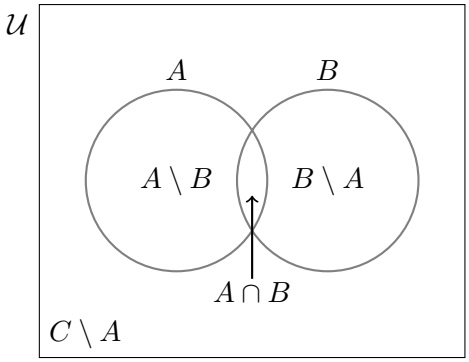

Figure 1.1: Domains within the population $\mathcal{U}$

Some assumptions are often made regarding the selection mechanism into B, in order to facilitate inferences using those datasets. Three assumptions are generally adopted for the methods described in this dissertation; see for example Chen et al. (2020) and Yang et al. (2021):

A1 *Ignorability*: Conditional on the set of covariates $\boldsymbol{x}_i$, δ_i and y_i are independent.
A2 *Positivity* or *Common Support*: Conditional on $\boldsymbol{x}_i$, $P(\delta_i = 1|\boldsymbol{x}_i) > 0, \forall i \in \mathcal{U}$.
A3 *Independence*: Conditional on $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $\delta_i \perp \delta_j, i \neq j, \forall i, j \in \mathcal{U}$.

Ignorability implies that $P(\delta_i = 1|\boldsymbol{x}, y) = P(\delta_i = 1|\boldsymbol{x})$. In other words, selection into B is ignorable conditional on the covariates $\boldsymbol{x}$. This assumption is similar to the Missing-At-Random (MAR) scenario of Rubin (1976). Andridge et al. (2019) and Boonstra et al. (2021) refer to this type of selection process as Selection At Random (SAR). When selection into B is influenced by y, then we instead have Selection Not At Random (SNAR), akin to the Missing-Not-At-Random (MNAR)/Not-Missing-At-Random (NMAR) scenario. In this dissertation, the SAR and SNAR terminology will be used to describe the type of missingness in B. In Chapter 5 the SAR assumption will be relaxed to allow the selection mechanism for B to be SNAR.

In addition to the above, the following will be assumed in some of the methods discussed:

A4 It is possible to accurately identify those records within A that are also in B
A5 There is full response in A

Assumption A4 may be satisfied if there exists a common unit identifier available on B and the population frame, or if it is possible to accurately link records in B to the population frame and/or A using a common set of linking variables. In the business setting, businesses may have a unique business number for administrative purposes such as filing taxation returns, and this identifier can be available on both the business frame and external data sources. In the population setting, personal identifiers may be available to undertake linking.

Assumption A5 will tend not to reflect reality, especially in this age of declining response rates (Beaumont 2020). The full response assumption in A is made to simplify the exposition of results, however it is noted that in general this assumption may be relaxed by assuming the response probabilities in A can be estimated using some variables available for the sample or the population; see for example Särndal and Lundström (2006). The relaxation implies that any non-response in A will be MAR and addressable through a

non-response weight adjustment or imputation process. With sufficient resources dedicated to careful non-response follow-up, it should be possible to ensure response rates are relatively high and render any remaining non-response in the probability sample as ignorable.

In addition to the above, it is assumed that the auxiliary variables $\boldsymbol{x}$ are available on the datasets A and B. In Chapter 2, y may be collected with measurement error on the non-probability dataset. Instead of observing y_i, y_i^* is observed, and the two variables are assumed to be related through the relationship (see Kim and Tam 2021)

$$y_i^* = \beta_0 + \beta_1 y_i + e_i \tag{1.1}$$

where (β_0, β_1) is unknown and $e_i \sim (0, \sigma^2)$. In the case that $(\beta_0, \beta_1) = (0, 1)$, there is no systematic measurement error in y. Further, measurement error in the reference sample covariates $\boldsymbol{x}$ will be explored in Chapter 5.

1.2 Existing methods for non-probability datasets

This section provides a brief overview of the literature on methods for making inference using non-probability datasets. A number of review papers also exist in this space that provide more detailed discussions of the various methods. The interested reader is pointed to Lohr and Raghunathan (2017), Zhang (2019), Tam and Holmberg (2020), Yang and Kim (2020), Rao (2021) and Wu (2022). More recently, Salvatore (2023) analysed a large number of documents on this topic using text mining and bibliometric techniques to identify current research trends. Rueda-Sánchez et al. (2025) conducted an empirical study to compare the performances of a range of different approaches, including calibration and model-based adjustments, and propensity score adjustments.

The method of inference used depends on the data structures of the non-probability and reference samples. There are three general approaches that will be discussed in this section:

Table 1.1: Data structures

Data type		**Response Variables y**	**Auxiliary Variables x**
Type 0	Frame	×	✓
	P Data (A)	✓	✓
Type I	NP Data (B)	✓	✓
Type II	NP Data (B)	✓	✓
	P Data (A)	×	✓
Type III	NP Data (B)	✓	✓
	P Data (A)	✓	✓
Type IV	NP Data (B)	×	✓
	P Data (A)	✓	✓

weighting-based approaches, imputation-based approaches, and doubly-robust approaches which combine a weighting-based approach for B with a model for the variable of interest y.

The data available in A and B will influence what estimation methods will be most suitable to use. Table 1.1, inspired from Tam and Holmberg (2020), describes four (Type I - Type IV) potential data structures relating to the non-probability sample and an accompanying probability sample. For completeness, a fifth data structure - Type 0 - is also included, which reflects the more standard survey situation when data is available from a probability sample and accompanying population frame, but there is no data from a non-probability sample. The Type 0 data structure is similar to the Type IV data structure, with the difference being that in the Type 0 there is full population coverage of the auxiliary variables, while in the Type IV case the coverage of the dataset B is incomplete and the selection of units into the dataset is governed by an unknown process.

Reference to these data types will occasionally be made in the discussion that follows.

1.2.1 Weighting approaches

Weighting approaches create a weight associated with each record in the reference sample or the non-probability sample, and these are used to form estimates for a target population quantity using an Inverse Probability Weighting (IPW) approach. Define w_i^A as the survey weight associated with a record i in A. w_i^A may be equal to $d_i^A = 1/\pi_i^A$, or d_i^A adjusted via a non-response and/or calibration process. Population totals may be estimated by $\hat{Y} = \sum_{i \in A} w_i^A y_i$.

Let $\hat{\pi}_k^B$ be an estimate of the propensity of selection into B for unit k in B. Then $w_k^B = 1/\hat{\pi}_k^B$ is the corresponding weight associated with record k, and we may form estimates of population totals using $\hat{Y} = \sum_{k \in B} w_k^B y_k$. Wu (2022) notes that a better estimator to use is the Hájek estimator

$$\hat{Y}_{Hájek} = \frac{N}{\hat{N}_B} \sum_{k \in B} w_k^B y_k \tag{1.2}$$

where $\hat{N}_B = \sum_{k \in B} w_k^B$.

When a Type III data structure exists, and assuming it is possible to determine δ_i for the units in A (for example, by linking the sample A to B), Kim and Tam (2021) proposed calibrating the design weights in A to the quantities $\sum_{i=1}^{N}(1, \delta_i, \delta_i y_i) = (N, N_B, Y_B)$. The calibration process involves finding new weights w_i^A for $i \in A$ that minimise a chosen distance metric subject to the calibration constraints $\sum_{i \in A} w_i^A \boldsymbol{x}_i = \sum_{i \in \mathcal{U}} \boldsymbol{x}_i = \boldsymbol{X}$ (Deville and Särndal 1992), where $\boldsymbol{X}$ is the vector of population totals for the set of auxiliary variables. In particular, the generalised regression estimator is a special case of the calibration estimator when the Chi-Square distance metric $\sum_{i \in A}(w_i^A - d_i^A)^2/2q_i d_i^A$ is used, where q_i is a tuning parameter.

Kim and Tam (2021) showed that using the calibration totals (N, N_B, Y_B) is equivalent to a post-stratified estimator combining the total Y_B with an estimated total $\hat{Y}_C$ obtained from $A \setminus B$. A benefit of their approach is that it is design-based and model-assisted,

and inference does not depend on the need for a correctly specified model. The set of calibration constraints can be expanded to include auxiliary variables (for example to address non-response) in A, and the method can also cater for measurement error in y in either A or B.

Rather than making a weight adjustment for the probability sample, a weighted non-probability sample may instead be produced. This seems attractive in particular when N_B is large in comparison with n_A. The main purpose of the reference sample A is now to aid in estimating the propensities of selection $\pi_k^B = \pi(\boldsymbol{x}_k, \boldsymbol{\phi})$ for $k \in B$, where π is a chosen parametric form such as the inverse logit function, and $\boldsymbol{\phi}$ are unknown parameters requiring estimation.

Under Assumption A1 and a Type I data structure, selection bias may be reduced by applying a weight calibration process to B which satisfies $\sum_{k\in B} w_k^B \boldsymbol{x}_k = \boldsymbol{X}$ (see, for example, Haziza et al. 2010). The initial weights feeding into the calibration may be given by $d_k^B = N/N_B$ or $d_k^B = 1$ (Rueda et al. 2023, Golini and Righi 2024). If $\boldsymbol{X}$ is not available but a Type III data structure exists, a pseudo-calibration estimator may be applied (Golini and Righi 2024; see also Righi et al. 2019) using estimated totals from the reference probability sample to feed in to the calibration equation

$$\sum_{k\in B} w_k^B \boldsymbol{x}_k = \sum_{i\in A} d_i^A \boldsymbol{x}_i \tag{1.3}$$

to produce final weights w_k^B for the units in B.

Direct estimation of propensity scores $\hat{\pi}_k^B$ may be accomplished in a variety of ways, and methods generally assume a Type II data structure. Chen et al. (2020) derive a pseudo log-likelihood function with two terms, one involving B and the other involving A, and this function can be maximised through iterative methods such as the Newton-Raphson procedure. Burakauskaitė and Čiginas (2023) detail a variation of this method which is applicable when $\boldsymbol{x}_i$ are available for $i \in \mathcal{U}$. Kim and Wang (2019) assume that it is possible to determine δ_i for units in A, and use this to estimate propensity scores based on the

probability sample.

The method of Elliot (2009), see also Elliott and Valliant (2017), estimates $\hat{\pi}_k^B$ for $k \in B$ by first pooling together the non-probability data and reference sample data. For unit i in the pooled dataset $A \cup B$, they define $Z_i = 1$ if unit i belongs to B, and $Z_i = 0$ if unit i belongs to A. Conditional probabilities $\hat{P}(Z_i = 0|\boldsymbol{x}_i)$ and $\hat{P}(Z_i = 1|\boldsymbol{x}_i)$ are estimated through a modelling process, and then combined to produce an estimated propensity of selection for units in B. This approach does not require knowledge about δ_i for units in A. On the other hand, the method assumes small sampling fractions and no overlap in the units captured within the two samples - an assumption that is not likely to hold if B is large. Liu et al. (2022) addressed the overlap issue by using only the non-overlapping units in the pooled sample to fit a model to estimate certain probabilities to feed into the creation of pseudo-weights. Their method requires knowing which units comprise $A \cap B$ (so that these can be removed from the pooled sample), and is not appropriate when the non-overlapping portion of the two samples is very small, or if one sample is a subset of the other.

Wang et al. (2021a) proposed an Adjusted Logistic Propensity (ALP) weighting method which also pools the non-probability and probability datasets together, but does not require the assumption of non-overlapping samples. However, in their method the final estimated participation probability may sometimes be greater than 1 if B is large in size. Savitsky et al. (2023) noted that the methods of Wang et al. (2021a) and Chen et al. (2020) are sub-optimal as they rely on pseudo-likelihoods for estimating propensities. Instead, Savitsky et al. (2023) constructed a likelihood defined directly on the observed pooled sample to estimate propensities. Their method does not require the samples A and B to be disjoint, allowing an unknown amount of overlap between the samples to be present. A hierarchical Bayesian approach was utilised to enable computation of all required probabilities simultaneously. In a simulation setting under moderate sample sizes, the method was found to produce more accurate estimates of participation probabilities for B compared to the pseudo-likelihood approaches used in Wang et al. (2021a) and Chen et al. (2020).

If Assumption A1 does not hold and the missingness in B is SNAR, then $\pi_k^B = \pi(\boldsymbol{x}_k, y_k, \boldsymbol{\phi})$,

and we need to include the variable of interest y in the model to estimate $\hat{\pi}_k^B$. Marella (2023) and Kim and Morikawa (2023) employed the sample empirical likelihood approach to address the SNAR selection mechanism for B. A feature of this approach is that known population means (under a Type I structure) or sample estimates of these means (under a Type II structure) of auxiliary variables can be included as calibration constraints in the maximisation process for the empirical likelihood to help address selection bias.

Machine learning techniques have also been explored to estimate the propensity scores. Ferri-García and Rueda (2020) provide a comparison of different machine learning approaches using a pooled dataset $A \cup B$ which assumes no overlapping sample, and under a Type II data structure. Castro-Martín et al. (2022) proposed including the weights derived from a propensity score estimation process, $\hat{w}_k^B = 1/\hat{\pi}_k^B, k \in B$, as part of a subsequent machine learning model training process (for example, linear regression) to predict $\hat{y}_i, i \in A$ under a Type II data structure. The imputed values obtained from the trained model may then be used to form estimates based on A:

$$\hat{Y}_{MI1} = \sum_{i \in A} w_i^A \hat{y}_i \tag{1.4}$$

where here the subscript MI denotes "Mass Imputation". More accurate estimated propensity scores may be achieved by incorporating known information about the auxiliary variables in the propensity score estimation for B. Zhu et al. (2023) assume a latent Gaussian copula model for the joint distribution of the auxiliary variables. The model is fitted using data in A, and a pseudo-population is simulated using the fitted model. The pseudo-population is used to estimate the marginal inclusion probabilities $P(\delta_k = 1), k \in B$, and these are then used to estimate propensity scores $\hat{\pi}_k^B$.

1.2.2 Imputation-based approaches

Imputation-based approaches assume that it is possible to form a reliable estimate for y_i using the available auxiliary information $\boldsymbol{x}_i$. This approach will generally be used when y

is missing from A (Type II data structure) or B (Type IV data structure). When a Type II data structure exists, the model for y is formed using the data in the non-probability sample and then applied to impute $\hat{y}$ for all the units in the reference sample, referred to as mass imputation (Chipperfield et al. 2012).

Under a Type II structure, an alternative imputation estimator to (1.4) is (see Wu 2022)

$$\hat{Y}_{MI2} = \sum_{k \in B} y_k + \Big(\sum_{i \in A} w_i^A \hat{y}_i - \sum_{k \in B} \hat{y}_k \Big) \tag{1.5}$$

where w_i^A may be design, non-response adjusted, or calibrated weights for $i \in A$, and $\hat{y}_i$ are predictions from a model, which may range from a linear model, to generalised additive models or kernel regression (Chen et al. 2022), to methods such as regression trees and random forests (Golini and Righi 2024). The estimator (1.5) may be interpreted as the sum of the true values from B and an estimated contribution for $C = \mathcal{U} \setminus B$ based on the modelled $\hat{y}$ values.

Rivers (2007) proposed a sample matching approach using nearest neighbour imputation to mass impute the missing y_i values for $i \in A$ under a Type II data structure. The non-probability dataset is treated as the donor population. A distance measure indicating the similarity of units in A and B is calculated using the available covariates $\boldsymbol{x}$, and the closest match is chosen to supply $\hat{y}_i$. Yang et al. (2021) extended the nearest neighbour imputation method to k nearest neighbours, again assuming a Type II data structure, where data from the k nearest neighbours are combined to produce a mean value $\hat{y}_i = k^{-1} \sum_{j=1}^{k} y_j$ which is used as the imputed value. When δ_i can be identified for each unit in A, the authors showed that the efficiency of the nearest neighbour imputation can be improved by also calibrating the weights in the probability sample to match known quantities in B, for example:

$$\sum_{i \in A} w_i^A (\delta_i, 1 - \delta_i, \delta_i \boldsymbol{x}_i, \delta_i \hat{y}_i) = (N_B, N - N_B, \boldsymbol{X}_B, Y_B)$$

where w_i^A are the final calibrated weights and N, N_B, $\boldsymbol{X}_B$ and Y_B are defined in Section 1.1.

When a Type IV data structure exists, the reference sample may be used to estimate the parameters of the model for $y_i|\boldsymbol{x}_i$. The fitted model is then used to provide predictions for $y_k, k \in B$. In Righi et al. (2019) and Golini and Righi (2024), a combined imputation and weighting approach is proposed whereby predicted values $\hat{y}_k$ for the variable of interest are produced from a modelling process, and are used in a pseudo-calibration estimator $\hat{Y}_{PC} = \sum_{k\in B} w_k^B \hat{y}_k$, where w_k^B is the solution to the calibration equation (1.3). The authors note that their approach can also be used in a Type III data structure when y is observed with error in B (that is, B collects y_k^* instead of y_k) and a measurement error model such as (1.1) is fitted to correct the error.

In the case where the modelling process fails to capture the true relationship between predictors and the variable of interest, $\hat{Y}_{PC}$ will be biased. Righi et al. (2019) and Golini and Righi (2024) proposed amending $\hat{Y}_{PC}$ by including a bias correction term in the estimator. This leads to the difference estimator (see also Breidt and Opsomer 2017)

$$\hat{Y}_{DPC} = \sum_{k\in B} w_k^B \hat{y}_k + \sum_{i\in A} d_i^A (y_i - \hat{y}_i) \tag{1.6}$$

Medous et al. (2023) extended the calibration approach of Kim and Tam (2021) from a Type III to a Type IV structure, proposing so-called QR predictors (see Wright 1983) to produce an estimated $\hat{Y}_B$. $\hat{Y}_B$ may then be combined with an estimated contribution from C (via A) to produce an improved population estimate.

1.2.3 Doubly Robust estimation

Many of the approaches outlined in the previous sections depend on an accurate working model. To protect against model misspecification, some authors have suggested doubly robust estimators constructed using both a propensity score model for B and a model for $y|\boldsymbol{x}$. The setup only requires one of the two models to be correctly specified in order to

be unbiased.

Assuming a Type II data structure, Chen et al. (2020) proposed two doubly robust estimators for $\bar{Y}$, with the second estimator preferred:

$$\hat{\bar{Y}}_{DR1} = \frac{1}{N}\sum_{k\in B}\frac{y_k - \hat{y}_k}{\hat{\pi}_k^B} + \frac{1}{N}\sum_{i\in A}\frac{\hat{y}_i}{\pi_i^A} \tag{1.7}$$

and

$$\hat{\bar{Y}}_{DR2} = \frac{1}{\hat{N}^B}\sum_{k\in B}\frac{y_k - \hat{y}_k}{\hat{\pi}_k^B} + \frac{1}{\hat{N}^A}\sum_{i\in A}\frac{\hat{y}_i}{\pi_i^A} \tag{1.8}$$

where $\hat{y}_i$ and $\hat{y}_k$ are the estimated values of $y_i, i \in A$ and $y_k, k \in B$ based on a fitted imputation model using $(y_k, \boldsymbol{x}_k)$ in B, $\hat{\pi}_k^B$ is the estimated participation probability for $k \in B$ using a method from Section 1.2.1, $\hat{N}_A = \sum_{i\in A} w_i^A$ and $\hat{N}_B = \sum_{k\in B} 1/\hat{\pi}_k^B$. There may or may not be overlap in the samples A and B.

More recently, Berg et al. (2025) examined doubly robust approaches which further seek to protect against model misspecification. They proposed combining a semi-parametric quantile regression approach for obtaining imputed values and a non-parametric local polynomial regression approach to obtain estimated propensities. Their approach to imputation provided several benefits over existing parametric approaches, such as protection against outliers and accommodating heteroskedasticity.

The doubly robust approach can be extended to a multiply robust scenario. Instead of employing a single propensity score model to estimate $\pi_k^B, k \in B$, and a single imputation model, Chen and Haziza (2023) suggested the use of m propensity score models and m imputation models. Each model may be based on different sets of explanatory variables. The results from the m models are "compressed" or summarised to form an overall estimated value for $\hat{y}_k$ or $\hat{\pi}_k^B$. The estimator is consistent as long as one of the imputation or propensity score models is correctly specified. Kim and Morikawa (2023) also suggest the use of multiple propensity score models and multiple constraints for bias calibration in their empirical likelihood approach.

1.3 Thesis outline

As presented earlier in this chapter, there is an increasing variety of approaches to correct for selection bias in non-probability samples. At the same time, few, if any, of these approaches are explored in the context of business data and its characteristics. The contributions laid out in this dissertation fills that gap in existing knowledge.

Chapter 2 examines and compares through a realistic and wide-ranging simulation study the effectiveness of a cross-section of existing estimation methods that have been developed for non-probability data. The comparison is undertaken in a business survey context, where the non-probability sample is a large portion of the population. The study explores the effects on the performance of estimators in the case of nonignorable non-response and measurement error in the target variable within the non-probability sample.

Chapter 3 presents a novel estimator to estimate participation probabilities for the non-probability sample. The approach involves maximising a pseudo-likelihood where the non-probability sample is combined with the reference sample units that do not also exist in the non-probability sample. The new estimator is shown theoretically to be efficient among the class of existing pseudo-likelihood approaches when the non-probability sample is large. A variant of the estimator is introduced which may be employed when it is not possible to identify the overlapping units. The efficiency of the new estimator is shown through both a simulation study and real-world data application.

Chapter 4 explores possibilities for optimising the sample design of the reference sample when the main objective is to produce statistics using the non-probability sample. Using the new estimator proposed in Chapter 3, optimal selection probabilities for the reference sample are derived under different sample design schemes. An approach is proposed to undertake the sample design in practice when the auxiliary information is only available for part of the population. A split-population design is then examined in the case when it is possible to link the non-probability dataset to a population frame. Results from a simulation study show that sample size savings can be obtained by efficient sample design of the reference sample.

Chapter 5 presents a framework for simultaneously addressing selection bias in the non-probability sample alongside covariate measurement error in the reference sample, when the data in the non-probability sample is collected without error. It is shown how the framework may also be applied in the case when there is a nonignorable selection mechanism driving the selection of units into the non-probability sample. Results from a simulation study and real-world data application show that the proposed approach effectively removes both the selection bias and covariate measurement error.

Some closing thoughts and potential future work are presented in Chapter 6.

Chapter 2

Empirical Comparison of Existing Methods

This chapter examines and compares through a realistic and wide-ranging simulation study the effectiveness of a cross-section of the estimation methods that have been developed for non-probability data, some in combination with probability sample data. The comparison is undertaken in a business survey context.

Our focus is on the situation where the non-probability dataset makes up a large (for example, 50%) portion of the population, an accompanying business probability sample from the population is available as auxiliary data, and the non-probability dataset can be linked to units on the probability sample and the population frame. The study explores the effects of nonignorable non-response and measurement error in the non-probability sample on the performance of the various estimators, with the probability sample used to help address these issues.

Other simulation studies comparing a range of estimation methods have been conducted in recent years; see for example Liu et al. (2022) and Rueda-Sánchez et al. (2025). However, an empirical comparison across such a wide range of methods and data quality scenarios, within a realistic business survey context, has to my knowledge not been done before.

The results from the study provide some practical insights for statistical agencies looking to use these estimators to make inference using their own large business non-probability datasets.

The rest of the chapter is structured as follows. Section 2.1 first outlines a number of sample design frameworks that may be used to produce reference samples to address shortcomings in the non-probability sample. Section 2.2 describes the empirical comparison study using simulated business data. The simulation is grounded in data analysis of ABS Business Longitudinal Analysis Data Environment (BLADE) data and includes empirically-derived right-skewed distributions and realistic missingness scenarios in the non-probability samples. Section 2.3 showcases the results from the simulation. Finally, some summary remarks are provided in Section 2.4.

2.1 Alternative sampling frameworks for A

In existing approaches, the reference probability samples A are generally assumed to be a given. In practice, the reference sample design may be approached in a number of ways based on the nature of the available non-probability data. In this section, I briefly discuss two sampling frameworks for the reference sample as alternatives to a traditional design-based sample from the full population. I explore how these sampling frameworks may be applied in conjunction with the various estimators in the next section.

2.1.1 Multiple frame approach

The original theory for multiple-frame surveys was developed by Hartley (1962; 1974), and has been built upon in recent years; see, for example, Lohr and Rao (2006) and Lohr (2011). In the context of non-probability data, one can consider B to be a full census from an incomplete population "frame" (Lohr 2021), which may or may not measure y, instead measuring covariates $\boldsymbol{x}$ that can be used to predict y. Lohr (2021) and Medous et al. (2023) note that the data integration estimators developed in Kim and Tam (2021)

can be considered within a multiple frame context.

Assume that a Type III data scenario where y is measured in both A and B applies, and it is possible to identify $C = \mathcal{U} \setminus B$. This may be accomplished, for example, by integrating the sample B to a population frame through a common unit identifier. Note that in general, the multiple frame approach assumes that we can determine which population segment(s) each unit belongs to. If C can be identified, a *screening* dual frame sample design may be employed whereby a probability sample A is selected from C only. A consistent estimator of Y is then simply

$$\hat{Y}_{screening} = Y_B + \hat{Y}_C \tag{2.1}$$

where $\hat{Y}_C$ may be a Horvitz-Thompson or Hájek estimator. Calibration benchmarks for the subpopulation C may be used if available to improve the efficiency of $\hat{Y}_C$. Zhang (2019) refers to (2.1) as the *split-population approach* to inference, where B and C constitute the two "populations". More generally, Zhang (2019) noted that one can produce a composite estimator for the population mean based on the mean value of y from B and the estimated mean of y from A, where A has been selected from C only.

A number of assumptions are generally made when making inferences under the multiple-frame framework; see for example Lohr (2022). Within a data integration context, some of these assumptions may be less likely to hold. Particularly relevant to this dissertation is that the variables captured on the non-probability source may not match exactly the variables of interest captured in the probability sample. If a screening dual frame design has been used, the lack of overlapping sample makes it more difficult to assess and address any measurement error in the non-probability or reference probability datasets.

2.1.2 Cut-off sampling

In cut-off sampling, a certain fraction of the population is deliberately excluded from the survey frame. A probability sample is then taken from the remainder of the population.

This practice tends to be used in business surveys when the variable of interest is highly skewed; Yorgason et al. (2011) and Elisson and Elvers (2001) provide some examples. Generally, the smallest businesses are placed into a single "take-none" stratum with zero probability of selection. The contribution of these businesses to the variable of interest is assumed to be negligible compared with the remaining part of the population, so there is a saving in terms of reduced respondent burden and cost without causing significant bias (Elisson and Elvers 2001). The rest of the population may be further divided into a "take-all" (completely enumerated) and a "take-some" (sampled) stratum. Denote the population in the take-none, take-some and take-all strata as $\mathcal{U}_E$, $\mathcal{U}_S$, and $\mathcal{U}_{CE}$ respectively, while Y_E, Y_S and Y_{CE} are the totals for y in those strata.

The non-probability dataset would often be useful for estimating the part of the population $\mathcal{U}_E$ intentionally not covered by the cut-off sample A. For example, if y is measured in both A and B, then Y_E may be estimated as

$$\hat{Y}_E = \frac{N_E}{N_{E,B}} \sum_{i \in \mathcal{U}_E \cap B} y_i$$

where N_E is the number of units in the population in the take-none stratum and $N_{E,B}$ is the number of units in B in the take-none stratum.

Note that this estimator will only be approximately unbiased if $\frac{Y_E}{Y_{CE}+Y_S} \approx \frac{Y_{B,E}}{Y_{B,CE}+Y_{B,S}}$. This approximation may be very imperfect in practice, but it may still be adequate when Y_E is small, as is usually the case by design. If auxiliary variables $\boldsymbol{x}$ are available for the whole population, then one may use this to model the propensities π_k^B, and thus form an IPW estimate for the take-none stratum by applying (1.2) to $B \cap \mathcal{U}_E$. The audit sampling ideas of Zhang (2019; 2021; 2023) may come in useful here to inform appropriate action for the cut-off population, with the idea that the audit sample only has to be taken intermittently.

2.2 Setup for the simulation study

As seen in Section 1.2, the methods for producing inferences from non-probability samples are wide-ranging. As such, any empirical comparison of methods would need to be somewhat selective. In this simulation study, a cross-section of approaches was selected that was considered to be straightforward to implement for a statistical agency. The aim of the exercise was to examine their performance under four scenarios: SAR vs SNAR missingness in the non-probability dataset, and with vs without measurement error in y on the non-probability dataset. In this simulation the non-probability dataset includes a large fraction (about 50%) of the population, so may be considered "big data".

In addition to Assumptions A4 and A5 in Section 1.1, the simulation study assumes that the population frame includes some auxiliary information, including frame employment and industry class, and these variables are available to use during the estimation process. This can be reasonably expected to hold in a business survey context. For example, some business information is often available from administrative sources (like business tax data) to attach to the population frame.

The simulation consisted of the following steps: (1) generating a number of data items for a finite population with distributional properties similar to some items on a real business survey dataset, (2) drawing a random subsample from the population to be the non-probability sample, (3) drawing a reference probability sample, and (4) applying some of the methods described in Section 1.2 to produce estimates. The process used to generate the random samples in step (2), and the associated probabilities of inclusion into the samples, are assumed to be unknown when applying the methods in step (4).

For each of the sample designs considered, $R = 2,000$ repeated samples A and B were drawn and combined in some way to produce estimates. The Monte Carlo Relative Bias (RB) and Relative Root Mean Squared Error (RRMSE) of the estimators were then calculated as

$$\text{RB} = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{Y}_r - Y}{Y}$$

$$\text{RRMSE} = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_r - Y)^2}}{Y}$$

where $\hat{Y}_r$ represents the estimate of Y from the r'th repeated sample, and Y is the true population total for y.

2.2.1 Creation of the synthetic population

This subsection describes the creation of the synthetic population. The population has $N = 900,000$ business records, consisting of two categorical domain variables (State and Industry), a frame size variable (Frame Employment), and three survey variables (Reported Employment, Total Weekly Wages/Salaries, and Overtime Pay). It resembles the real-world population of employing businesses in Australia in terms of the distribution of businesses across size categories, industry divisions and state. A combination of published survey outputs alongside employee tax data and survey microdata sourced from the ABS DataLab environment was used to help generate the survey variables of interest. The synthetic population is available at this link: https://zenodo.org/records/11095755.

Step 1: Derive input probabilities

Data from the ABS publication *8165.0 Counts of Australian Businesses, including Entries and Exits* provides counts of the number of businesses in each State by 4-digit Australian and New Zealand Standard Industry Classification (ANZSIC) code by broad employment size group. The business counts were aggregated to the ANZSIC Division level. Employee tax data (Business Longitudinal Analysis Data Environment (BLADE) 2020) for the 2019/2020 financial year was obtained from the ABS DataLab, and the business counts from this dataset were used to apportion the broad employment counts into finer size groups. Business counts were then translated into proportions: i) the proportion of businesses in each size group, ii) the proportion of businesses within a size group belonging to a particular industry, and iii) the proportion of businesses within a size group and industry that belong to a particular state. These proportions were then treated as input probabilities to assign state, industry and size categories to the units in the synthetic

population.

Step 2: Generate frame and reported employment

The frame and reported employment variables are assumed to have some relationship with each other, while also individually exhibiting certain characteristics in terms of their mean, variance, skewness and kurtosis. The method described in Vale and Maurelli (1983) (from here on referred to as Vale-Maurelli) was used to generate data for these two variables. The method generates multivariate non-normal data based on user-specified mean, covariance, skewness and kurtosis characteristics for a given set of variables. Data was generated separately for each industry within a size group. The number of records generated in each industry and size group was based on the overall population size N and the proportions generated in Step 1. Table 2.1 provides the parameter values used in the Vale-Maurelli process to generate the frame employment and reported employment data items. Separate data generation processes were run for each of 14 size groups.

Table 2.1: Parameter values fed into Vale-Maurelli process - Employment

Emp Grp	Frame Employment				Reported Employment				Covar.
	Mean	Var	Skew	Kurt	Mean	Var	Skew	Kurt	
0-4	2	3	1.1	1.2	3.5	10	1.2	1.4	4
5-19	8	4.5	1.1	2	8.5	10	1.2	1.8	4
20-49	26	28	0.8	1.8	25	35	0.9	1.6	25
50-99	68	60	0.6	3.55	51	70	0.9	3.6	55
100-149	120	70	0.6	2.2	110	70	0.8	2.4	55
150-199	160	100	0.4	1.6	145	120	0.7	1.9	70
200-249	220	2,200	0.5	1.5	200	2,400	0.5	1.3	2,000
250-299	280	6,300	0.7	1.2	250	5,400	0.7	1.5	4,800
300-349	320	14,000	0.7	1.2	270	16,000	0.9	1.3	12,000
350-399	375	25,000	0.8	0.8	300	25,000	0.9	1.1	21,000
400-449	420	45,000	1.03	1.2	370	45,000	1.02	1.1	35,000
450-499	470	58,000	1.1	1.3	420	55,000	1.1	1.5	47,000
500-999	650	80,000	0.85	0.8	550	72,000	0.8	0.8	66,000
1000+	1,350	200,000	2.3	9	1,200	200,000	2	8	190,000

Note: Parameters created based on Pay-As-You-Go (PAYG) data from the 2019-2020 financial year (Business Longitudinal Analysis Data Environment (BLADE) 2020), merged on to Economic Activity Survey data from the 2020-2021 financial year (Australian Bureau of Statistics 2021), accessed through the ABS DataLab 2 March, 2023.

State indicators were assigned to the generated data randomly based on a multinomial

distribution, with probabilities based on the state probabilities calculated in Step 1.

Step 3: Generate total wages and salaries

Data from the November 2021 cycle of the ABS publication *Average Weekly Earnings, Australia* (Australian Bureau of Statistics 2022a) was used to inform our modelling of wages and salaries data. Within each industry division and size group, a total wages figure for the i'th record was generated as

$$y_i = 1740 f_{id} x_{i,rep} + \epsilon$$

where $f_{id} = w_d + \text{Normal}(0, v_s)$ is a randomised wage inflation factor for the d'th industry division relative to the Australian-level average earnings ($1,740) based on industry wage factors w_d sourced from the AWE publication, v_s is an error variance term for size group s, $x_{i,rep}$ is the simulated reported employment for the i'th business, and $\epsilon = \text{Normal}(0, v_e x_i)$ is a random "error" for the wages and salaries figure, normally distributed with variance increasing as x_i increases. The term v_e is a multiplier which varies depending on size group.

Table 2.2 provides the wage factors used for each industry. The factors represent the relative magnitude of each industry's weekly wage compared with the national average.

Table 2.2: Industry Wage Factors

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1.5	0.9	1.1	1	0.95	0.75	0.7	1	1.25	1.2	0.9	1.2	0.9	1.1	1.1	1	0.9	0.75

Note: Factors generated using data from *Average Weekly Earnings, Australia*, November 2021 cycle (Australian Bureau of Statistics 2022a)

Table 2.3 contains the variance parameter values used in each size group for generating the wages data item. These parameter values were chosen heuristically to obtain a realistic looking distribution for the resulting wages data within each size group.

Step 4: Generate overtime payments

Table 2.3: Variance terms for wages data item

Size Group	Wage variance factor (v_e)	Error variance (v_s)
0-4	0.10	50
5-19	0.10	75
20-49	0.15	100
50-99	0.15	150
100-149	0.15	150
150-199	0.15	150
200-249	0.15	200
250-299	0.20	250
300-349	0.20	250
350-399	0.20	250
400-449	0.20	250
450-499	0.20	250
500-999	0.20	275
1000+	0.15	325

A mixture model was developed to produce the overtime payment data item. A Bernoulli distribution was first used to model the likelihood of non-zero overtime payments in each industry. Different probabilities were used for each industry, ranging from 0.1 to 0.5. For businesses with a non-zero overtime payment, an exponential model with mean 0.1 was used to derive a factor f denoting the magnitude of overtime payments as a proportion of total earnings. The total overtime payment was derived as

$$y_{\text{overtime}} = f y_{\text{earnings}}$$

Step 5: Generate data items containing measurement error

A measurement error version of total earnings, y^*_{earnings}, was produced using a two-step process. First, the original/true earnings value for each unit was perturbed by a Normal(0.85,0.025) multiplicative factor. After this, for a small random subsample of the population the earnings value was further adjusted higher or lower to simulate the presence of anomalous/outlier units in the population.

Measurement error versions of reported employment and overtime were then created by multiplying the original/true values by the factor $y^*_{\text{earnings}}/y_{\text{earnings}}$, where this factor was

calculated separately for each record. This meant that for each record the relationships between the original earnings, reported employment, and overtime variables were maintained on the measurement error versions of the variables. Although in reality measurement error may be expected to be less structured than this, the use of a constant factor within a record should not adversely impact the outcomes from the simulation.

The selection mechanism for inclusion in the big dataset is given by $\delta_k \sim \text{Bernoulli}(\pi_k^B)$. A two-stage process was used to generate the final values of π_k^B for the population. At the first stage, an initial probability π_{1k}^B was produced, where

$$\pi_{1k}^B = \frac{\exp(\phi_0 + \phi_1 x_k + \phi_2 y_k)}{1 + \exp(\phi_0 + \phi_1 x_k + \phi_2 y_k)}$$

Two types of big dataset were produced, one following a SAR process, and one following a SNAR process. $(\phi_0, \phi_1, \phi_2) = (0.09, 0.009, 0)$ for the SAR dataset, and $(\phi_0, \phi_1, \phi_2) = (0.85, 0.009, -0.1)$ for the SNAR dataset. In the calculation of π_{1k}^B for both datasets, the x variable used was Frame Employment, and the y variable used was the natural logarithm of total weekly earnings.

At the second stage, the π_{1k}^B values were adjusted downwards by a pre-specified factor in some industries, to simulate a reduced likelihood of being present on the big dataset for those industries. The resulting probabilities were the final π_k^B values. In the selection/inclusion model, units with smaller frame employment have a lower chance of being included in the big dataset B.

For each sample draw of the simulation, a big dataset sample was drawn using Poisson sampling and the final π_k^B probabilities.

2.2.2 Probability sample designs

The population was assigned to strata based on State, Industry Division and Frame Employment for each business. Size stratum categories were: 0-4, 5-19, 20-299, and 300+ employees.

Three sample design scenarios were examined:

- Single-frame - An optimal allocation of sample to strata using the full population frame, $\mathcal{U}$
- Dual-frame - An optimal allocation of sample to strata using $C = \mathcal{U} \setminus B$ as the probability sampling frame (with B comprising the other frame)
- Cut-off - An optimal allocation of sample to strata where the sampling frame is the population excluding units in the smallest (0-4 employees) size class

The sample designs were produced to meet the following accuracy constraints for the data item of interest "Total Weekly Wages/Salaries" for a single time period on the relevant sample frame:

- Relative Standard Error (RSE) of 1.5% at the National level
- RSE of 5% for each Industry Division
- RSE of 5% for each State

The Bethel-Chromy algorithm (Bethel 1989, Chromy 1987) was applied to produce optimal allocations for each of the three scenarios, treating the relevant sample frame as the population of interest. For example, in the dual-frame scenario the algorithm was applied to achieve accuracy targets based on the units in C. Note that ideally the accuracy targets for the three sample designs would be specified taking into account the non-probability data in order to achieve a required level of accuracy for the estimates over the full population. However, this was considered infeasible for this simulation study given that doing this would depend on the estimators being used, and would not be a straightforward process given the many estimators being considered. Chapter 4 further discusses efficient design of the sample A to be combined with the non-probability data B.

A minimum sample size of 6 was applied for each sampled stratum. The 300+ size strata were designated to be completely enumerated strata with a sampling fraction of 1.

2.2.3 Estimators examined

Table 2.4 provides a description of the estimators compared for the single-frame sample design.

The dual-frame probability sample design is a *screening* dual-frame design. It is thus possible to combine the estimate from the probability sample with the big data total to form an estimate for $\mathcal{U}$. Two variants were considered as described in Table 2.5.

Table 2.6 describes the estimation methods applied to the cut-off probability sample.

For comparison, estimates were also produced where population frame information on employment and industry were the auxiliary variables available in the estimation process. The estimators used are described in Table 2.7.

By adopting the measurement error model (1.1), it is possible to re-arrange and obtain an expression for a measurement error corrected version of y_i^*:

$$\hat{y}_i = \hat{\beta}_1^{-1}(y_i^* - \hat{\beta}_0) \tag{2.2}$$

In a Type III data scenario, the parameters β_0 and β_1 may be estimated using data from $A \cap B$. This simulation study examined the performance of a few estimators when measurement error was present, as detailed in Table 2.8.

2.3 Simulation study results

Table 2.9 provides the reference sample sizes resulting from each of the different sample designs. The sample reductions are shown relative to the single-frame design. For this simulation, the dual-frame designs provide good potential for reducing the sample size, with a reduction of about 40%. When using the cut-off sampling, the resulting sample reductions are about 11%. Note that no attempt was made to standardise the designs

Table 2.4: Summary of estimators compared - single frame sample

Estimator	Data Scenario	Description
GREG	Type 0	Generalised Regression estimator with frame employment as the x variable
RDI	Type III	The Regression Data Integration estimator of Kim and Tam (2021), where the probability sample is calibrated to big data totals $(N, N_B, Y_{earnings})$. In the with-measurement error scenario, the probability sample is calibrated to the measurement error versions of the data item
QR MA	Type IV	The model-assisted QR estimator described in Medous et al. (2023). Frame employment is used as the explanatory variable
KW	Type II	Estimation of $\hat{\pi}_k^B, k \in B$ by employing the method of Kim and Wang (2019). $\delta_i, i \in A$ is obtained by linking units in A and B. Frame employment and industry division are the explanatory variables in the model for $\hat{\pi}_k^B$. Estimates are produced using (1.2). In the with-measurement error scenario, the measurement error versions of the data items are used in estimation
KW-Cal	Type II	A two-step weighting process starting with creation of KW weights (see above). These weights are then calibrated to population size N and frame employment total. The with-measurement error scenario uses the measurement error versions of the data items on the big dataset
KW-Earn	Type II	The KW estimator with the addition of the natural logarithm of Earnings as an explanatory variable in the model for $\hat{\pi}_k^B$
ALP	Type II	The Adjusted Logistic Propensity method of Wang et al. (2021a), where estimation of $\hat{\pi}_k^B, k \in B$ involves pooling data in A and B together and fitting a weighted logistic model on the pooled data. Note that this pooling approach means we have an assumed model $\pi_k^B = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_k)$
Wgt_Reg_MI	Type II	A mass imputation for y in A using weighted regression modelling. It is assumed in this case that y variables are not available in A. Weighted models are fit for each y data item based on the big data, where the weights are the estimated KW weights. The models are applied to the units in the probability sample to produce $\hat{y}_i, i \in A$. (1.4) is used to produce estimates of total
DR_wgt	Type II	The doubly-robust estimator (1.8) which combines the KW and Wgt_Reg_MI approaches
HD_MI	Type II	Hot deck mass imputation method to impute $y_i, i \in A$, where industry and size groups are used to form the classes that the hot deck imputation will be performed within

Table 2.5: Summary of estimators compared - dual frame sample

Estimator	Data Scenario	Description
SP	Type III	The split-population estimator (2.1) for Y where $\hat{Y}_C$ is a Horvitz-Thompson estimator using data from A
SP_Cal	Type III	The split-population estimator (2.1) for Y, where $\hat{Y}_C$ is calibrated to population totals (N_C, X_C), with X being Frame Employment

Table 2.6: Summary of estimators compared - cut-off sample

Estimator	Data Scenario	Description
CO+BD	Type III	A Horvitz-Thompson estimate based on the cut-off sample, added to the big data total for the small size units
CO_Cal+KWFr	Type 0	An estimate from the cut-off sample calibrated to population totals which excludes the cut-off population, combined with an estimate based on the big data in the excluded part of the population which is weighted up by KW weights produced by linking to the frame to obtain $\delta_k, k \in B$ and frame information on X

Table 2.7: Summary of estimators compared - Big data and frame only

Estimator	Data Scenario	Description
AuxDiv	Type I	The big data total in each Industry Division d adjusted by an industry-specific factor $X_d/X_{B,d}$, and then summed over the D industries to produce the overall total: $\sum_{d\in D}(X_d/X_{B,d})\sum_{k\in N_{B,d}} y_k$
KWFr	Type I	Estimation of $\hat{\pi}_k^B, k \in B$ by first linking the frame data to the big dataset to find $\delta_k, k \in B$. A logistic regression is then fitted on the frame data to estimate $\hat{\pi}_k^B$, with frame employment and industry division as the explanatory variables in the model. Estimates are produced using (1.2)

in terms of the achieved precision or RMSE of estimators, as it is unclear how such a standardisation would be defined when there are many variables and estimators being

Table 2.8: Summary of estimators - Measurement error corrected estimators

Estimator	Data Scenario	Description
KW-Cor	Type III	Apply the correction (2.2) based on data from $A \cap B$. Apply the model to the data in the big dataset to create $\hat{y}_k, k \in B$. Estimates are produced by feeding these $\hat{y}_k$ into (1.2), and applying the weights from the KW estimator
KW-Cal-Cor	Type III	Similar to KW-Cor, except with an additional calibration step to population size N and frame employment total
CO_Cal+KWFr-Cor	Type III	Similar to the CO_Cal+KWFr cut-off sample approach, but with measurement error correction for the small units. The measurement error corrected data are then used to provide the contribution from the big dataset
KWFr-Cor	Type III	Similar to the KWFr estimator in Table 2.7, except now the probability sample is also utilised to fit a measurement error correction model for each data item. Measurement error corrected versions of each data item are used instead of y^* to form estimates

considered. As a result, the sample sizes achieved by the different designs only provide an indication of the reductions that could be achieved, and need to be considered in conjunction with the RMSE results achieved by those designs in Tables 2.11 to 2.14.

Table 2.9: Sample sizes under different designs

Sample Design	Sample Size	Sample Reduction
Single-frame	7,715	0%
Dual-frame - SAR	4,559*	41%
Dual-frame - SNAR	4,598*	40%
Cut-off	6,883	11%

* Average sample size over all simulations

In each iteration of the simulation, national level estimates were produced for four y variables of interest - total weekly earnings (Earn), total reported employment (Emp), total overtime (Ovt), and average weekly earnings (AWE) defined as the ratio of total weekly earnings to total reported employment. The RB and RRMSE were calculated for the different estimators described in Tables 2.4, 2.5, 2.6, 2.7 and 2.8, and are shown in

Tables 2.11 to 2.14.

Table 2.10 lists the best-performing estimator in terms of RRMSE for each data item in each of the four scenarios of SAR vs SNAR and with vs without measurement error. No one estimator consistently outperforms the others in all scenarios. However, note that the SP-Cal estimator seems to perform reasonably in the no measurement error scenario, while the GREG estimator more often performs best when there is measurement error. The results for each of the four classes of estimators are highlighted next.

Table 2.10: Best-performing estimators under different scenarios

	No Measurement Error	With Measurement Error
SAR	Earn: SP-Cal Emp: KW-Cal Ovt: KWFr AWE: AuxDiv	Earn: GREG Emp: GREG Ovt: KWFr-Cor AWE: KWFr
SNAR	Earn: SP-Cal Emp: SP-Cal Ovt: ALP AWE: ALP	Earn: GREG Emp: GREG Ovt: KW-Cal-Cor AWE: DR_wgt

2.3.1 Single-frame design results

The GREG (Type 0 data structure), RDI (Type III) and QR MA (Type IV) estimators rely on the probability sample data for inferences, and as expected they have negligible relative bias under all missingness and measurement error scenarios examined. These estimators are "safe" estimators, producing robust performance. Of these three estimators, the GREG with frame employment as the auxiliary variable performs best - it is also consistently among the best performers across all estimators. The RDI estimator, which relies on B only to provide calibration benchmarks for weighting A, achieves good performance, coming close to the GREG. It is also asymptotically unbiased. Combining the GREG and RDI benchmarks into a single calibration process can further reduce the RRMSE of the resulting estimates, however care should be observed when choosing the set of benchmarks as including too many calibration constraints may result in an increase in variance, or worse yet an infeasible calibration process (as noted in Golini and Righi 2024). The addition of

measurement error in y erodes the performance of the RDI estimator (see Tables 2.12 and 2.14).

Table 2.11: Monte Carlo Bias and RRMSE of estimators based on 2000 samples - SAR, No Measurement Error

Estimator	RB ($\times 10^2$)				RRMSE ($\times 10^2$)			
	Earn	Emp	Ovt	AWE	Earn	Emp	Ovt	AWE
Single-frame design								
RDI	0.0	0.0	0.2	0.0	1.0	1.0	6.3	0.4
GREG	0.0	0.0	0.2	0.0	0.8	0.7	6.3	0.4
QR MA	0.0	0.0	0.2	0.0	1.3	1.2	6.4	0.4
KW	0.1	-0.2	0.3	0.2	2.0	2.0	2.4	0.5
KW-Cal	0.2	0.0	0.4	0.2	0.5	0.0	1.5	0.5
KW-Earn	-0.1	-0.3	0.1	0.2	2.2	2.2	2.6	0.5
ALP	-0.2	0.0	0.6	-0.1	1.1	1.1	1.3	0.1
Wgt_Reg_MI	0.2	0.0	0.6	0.3	1.2	1.1	1.3	0.3
DR_wgt	0.2	0.0	0.6	0.3	1.2	1.1	1.3	0.3
HD_MI	0.2	0.2	0.5	0.0	1.4	1.4	6.5	0.5
Dual-frame design								
SP	0.0	0.0	0.0	0.0	0.5	0.5	2.5	0.2
SP_Cal	0.0	0.0	0.0	0.0	0.3	0.3	2.5	0.2
Cut-off design								
CO+BD	-6.2	-6.3	-6.5	0.0	6.3	6.4	8.7	0.4
CO_Cal+KWFr	0.1	0.0	0.1	0.2	0.6	0.5	5.6	0.4
Big data only								
AuxDiv	-2.9	-2.8	-3.3	-0.1	2.9	2.8	3.3	0.1
KWFr	0.4	0.1	0.6	0.3	0.4	0.1	0.7	0.3

Note: Definitions for the different estimators are given in Tables 2.4-2.8.

When there is no measurement error, the RRMSE for the QR MA estimator is slightly higher than that for the RDI estimator, reflecting a penalty due to the fact that it estimates y_k for all units in the large dataset B, while the RDI is able to utilise the real values of $y_k, k \in B$. When measurement error exists in B, however, the QR MA estimator provides slightly better RRMSE outcomes for Earn and Emp compared with the RDI which uses the mis-measured y_i^* from B in its benchmarks.

A few variants of propensity score estimators were examined in our study, applicable under a Type II data structure. The KW estimator does not perform as well as other Type II methods, although its RRMSE values are still reasonably low in terms of magnitude. This

result appears to agree with the findings from the simulation study in Liu et al. (2022), where under smaller sampling fractions in the probability sample KW did not perform as well as other approaches, including the ALP method also compared in this study. On the other hand, the KW-Cal estimator performs very well in the ideal scenario of SAR and no measurement error (see Table 2.11). In general, the calibration to frame employment helps to reduce the variance of the KW estimates.

The ALP estimator performs favourably as an alternative to the KW estimator, achieving a lower RRMSE in the four scenarios tested. In the simulation study, it was found that some of the resulting ALP propensities were larger than 1 when using this method, leading to IPW weights of less than 1. This, along with other conceptual issues with the pooled approach noted by Wu (2022), means the survey practitioner will need to consider how applicable this approach will be for their case. In this study, though, these issues did not hinder the ability of the estimator to produce relatively efficient, unbiased estimates of population totals.

In the SAR case, the inclusion of earnings in the KW-Earn estimator did not produce reductions in RRMSE over the KW estimator, which is expected since in the SAR case missingness in B does not depend on earnings (see Table 2.11). On the other hand, in the SNAR case where the propensity of selection into B is also influenced by earnings, the inclusion of earnings in the model yielded lower RB and RRMSE (see Table 2.13) compared with the KW estimator.

The performance of the mass imputation approach Wgt_Reg_MI in this study was better than the KW estimator in all scenarios, although both suffered under measurement error. The inclusion of estimated propensity weights in the regression imputation model improved the performance of the regression model, aligning with findings in Castro-Martín et al. (2022). The DR_wgt estimator, which provides protection against mis-specification in one of the IPW or Regression models, tends to have comparable RRMSE compared with the Wgt_Reg_MI estimator.

The HD_MI estimator is a non-parametric mass imputation approach for A. This estimator was chosen as a simpler to implement and less computationally intensive approach

(and hence faster) compared with k Nearest Neighbour imputation. Compared to the parametric Wgt_Reg_MI approach, HD_MI produced comparable RB and RRMSE results for 3 of the four data items of interest in all scenarios except the ideal SAR without measurement error case. The Ovt data item was the one item where the hot deck imputation did not perform as well as the regression model. The results suggest that, when suitable imputation classes are formed using the covariates $\boldsymbol{x}$, the hot-deck method can perform well.

Table 2.12: Monte Carlo Bias and RRMSE of estimators based on 2000 samples - SAR, With Measurement Error

Estimator	RB ($\times 10^2$)				RRMSE ($\times 10^2$)			
	Earn	Emp	Ovt	AWE	Earn	Emp	Ovt	AWE
Single-frame design								
RDI	0.2	0.1	0.3	0.0	1.4	1.3	6.4	0.4
GREG	0.0	0.0	0.2	0.0	0.8	0.7	6.3	0.4
QR MA	0.0	0.0	0.2	0.0	1.3	1.2	6.4	0.4
KW	-12.9	-12.8	-13.0	-0.2	13.0	12.9	13.2	0.5
KW-Cal	-12.8	-12.6	-12.9	-0.2	12.8	12.6	13.0	0.5
KW-Cor	-0.1	-0.4	-0.1	0.3	3.2	3.6	5.3	1.2
KW-Cal-Cor	0.1	-0.3	0.1	0.3	2.5	3.0	4.8	1.2
KW-Earn	-13.0	-12.9	-13.1	-0.2	13.2	13.1	13.3	0.5
ALP	-12.9	-12.4	-12.5	-0.6	13.0	12.5	12.6	0.6
Wgt_Reg_MI	-12.8	-12.6	-12.8	-0.1	12.8	12.7	12.8	0.2
DR_wgt	-12.8	-12.6	-12.8	-0.1	12.8	12.7	12.8	0.2
HD_MI	-12.8	-12.4	-12.7	-0.4	12.9	12.6	14.8	1.0
Dual-frame design								
SP	-8.2	-7.9	-8.4	-0.4	8.3	7.9	8.8	0.4
SP_Cal	-8.2	-7.9	-8.4	-0.4	8.2	7.9	8.8	0.4
Cut-off design								
CO+BD	-7.2	-7.1	-7.5	-0.1	7.3	7.2	9.4	0.4
CO_Cal+KWFr	-1.7	-1.7	-1.8	0.0	1.9	1.8	5.9	0.4
CO_Cal+KWFr-Cor	0.2	0.1	0.3	0.0	0.9	0.8	5.7	0.6
Big data only								
AuxDiv	-15.1	-14.7	-15.8	-0.4	15.1	14.7	15.8	0.5
KWFr	-12.7	-12.5	-12.7	-0.1	12.7	12.5	12.7	0.2
KWFr-Cor	0.2	-0.1	0.3	0.3	2.5	2.9	4.6	1.1

Note: Definitions for the different estimators are given in Tables 2.4-2.8.

It is worth pointing out the RRMSE results for the Overtime variable when no measurement error is present (Tables 2.11 and 2.13). The Overtime variable has a larger population

variance and lower correlation with the benchmarking variable Frame Employment compared with Earnings or Reported Employment. The total value of Overtime in B was not included as a calibration constraint for the RDI and QR MA estimators, and as a result the RRMSE results for the RDI and QR MA estimators tend to be large. The B totals could of course have been included for all the variables of interest in the calibration, but as noted above this may lead to more variable weights and hence higher RRMSE, or an infeasible calibration process. This suggests that in a multi-purpose survey with a potentially large number of disparate variables, these estimators may not be suitable as it will be infeasible to include all of the different variables in the calibration process.

More generally for Overtime when no measurement error is present, the estimators based on the probability sample A - RDI, GREG, QR MA and HD_MI - tend to have a large RRMSE. On the other hand, the KW-based, ALP, Wgt_Reg_MI and DR_wgt methods achieve a significantly lower RRMSE; these estimators all base their inference on the much larger dataset B which helps in lowering the variance of the Overtime estimates.

From Tables 2.12 and 2.14, the uncorrected KW-based, ALP, HD_MI, Wgt_Reg_MI and DR_wgt approaches applicable under a Type II data structure all yield poor results under measurement error due to their reliance on the data in B. This was true under both a SAR and SNAR setting. None of these approaches included any provision for erroneously measured data items. Measurement error correction successfully negates the bias from measurement error when the measurement error model holds. However, the cost is a much higher overall RRMSE.

2.3.2 Dual-frame design results

In the without measurement error scenarios, the split-population estimators are unbiased, and their RRMSE results are generally among the lowest of all estimators examined across SAR and SNAR settings (see Tables 2.11 and 2.13). Including a calibration to auxiliary population totals for the dual-frame estimator helps to reduce RRMSE even further for earnings and reported employment.

However, measurement error in B leads to bias in the split-population estimators due to the reliance on B to produce a value for Y_B (see Tables 2.12 and 2.14). The dual-frame samples here don't have any overlap between A and B. Including some overlap between the samples A and B may be desirable to provide some data for correcting the measurement error.

2.3.3 Cut-off design results

The results for the estimators using the cut-off sample design show that simply combining the reference sample and the portion of B under the cut-off threshold did not account for all the contribution below the threshold. This was because there was a significant negative bias due to the contribution of units under the cut-off threshold which were also not in B. This was exacerbated by the fact that small units are less likely to be in B.

The CO_Cal+KWFr estimator aims to account explicitly for the contribution of the excluded part of the population $\mathcal{U}_E$ using the KW estimator and available frame information. In the without-measurement error scenarios, this effectively corrects for the selection bias, and also results in a low RRMSE for all items except overtime. When measurement error exists, using the big data to account for the contribution of $\mathcal{U}_E$ means that there will be some bias in the estimates due to measurement error. The estimator CO_Cal+KWFr-Cor which applies a measurement error model using available auxiliary information (in this case, from the frame) assists in reducing the RB, with an overall RRMSE that ranks well compared with the other estimators.

In the SAR without measurement error scenario (Table 2.11) there is little reason to opt for the CO_Cal+KWFr estimator rather than the KWFr estimator. In this case, where the quality of the data in B is reliable, the reference sample is less necessary for producing efficient estimates. However, when SNAR or measurement error exists, the reference sample provides a safety net, providing a means to help correct the measurement error and reduce the impact from SNAR missingness in the big data.

For the worst-case SNAR with measurement error scenario, the CO_Cal+KWFr-Cor esti-

Table 2.13: Monte Carlo Bias and RRMSE of estimators based on 2000 samples - SNAR, No Measurement Error

Estimator	RB ($\times 10^2$)				RRMSE ($\times 10^2$)			
	Earn	Emp	Ovt	AWE	Earn	Emp	Ovt	AWE
Single-frame design								
RDI	0.0	0.0	-0.1	0.0	1.1	1.0	6.5	0.4
GREG	0.0	0.0	-0.1	0.0	0.8	0.7	6.4	0.4
QR MA	0.0	0.0	-0.1	0.0	1.3	1.2	6.5	0.4
KW	-3.0	-3.1	-2.9	0.1	3.6	3.6	3.9	0.5
KW-Cal	-1.5	-1.6	-1.6	0.1	1.6	1.6	2.3	0.5
KW-Earn	-0.2	-0.2	-0.1	0.0	1.9	1.8	2.5	0.5
ALP	-1.7	-1.6	-1.5	-0.1	2.1	2.0	1.9	0.1
Wgt_Reg_MI	-1.4	-1.6	-1.5	0.2	1.9	2.0	1.9	0.2
DR_wgt	-1.4	-1.6	-1.5	0.2	1.8	1.9	1.9	0.2
HD_MI	-1.3	-1.2	-1.2	-0.1	1.9	1.8	6.8	0.5
Dual-frame design								
SP	0.0	0.0	0.0	0.0	0.6	0.6	2.9	0.2
SP_Cal	0.0	0.0	0.0	0.0	0.4	0.3	2.8	0.2
Cut-off design								
CO+BD	-6.6	-6.6	-6.8	0.0	6.7	6.7	8.9	0.4
CO_Cal+KWFr	-0.8	-0.9	-0.8	0.1	1.0	1.0	5.7	0.4
Big data only								
AuxDiv	-3.3	-3.2	-3.9	-0.2	3.3	3.2	4.0	0.2
KWFr	-2.7	-2.8	-2.6	0.1	2.7	2.8	2.7	0.1

Note: Definitions for the different estimators are given in Tables 2.4-2.8.

mator yielded close to the lowest overall RRMSE for the data items of interest (see Table 2.14). This was the case even though the contribution from the KW estimator did not fully reflect the non-response mechanism since that estimator did not include earnings (hence there is still a negative bias in the estimates obtained). This demonstrates the benefit of not relying solely on B to produce estimates (such as in the single-frame KW-based estimators). Additionally, the exclusion of the smallest units from the cut-off sample helps to reduce the overall sample variance.

Table 2.14: Monte Carlo Bias and RRMSE of estimators based on 2000 samples - SNAR, With Measurement Error

Estimator	RB ($\times 10^2$)				RRMSE ($\times 10^2$)			
	Earn	Emp	Ovt	AWE	Earn	Emp	Ovt	AWE
Single-frame design								
RDI	0.2	0.2	0.1	0.0	1.4	1.3	6.5	0.4
GREG	0.0	0.0	-0.1	0.0	0.8	0.7	6.4	0.4
QR MA	0.0	0.0	-0.1	0.0	1.3	1.2	6.5	0.4
KW	-15.4	-15.1	-15.6	-0.3	15.5	15.2	15.8	0.6
KW-Cal	-14.1	-13.8	-14.5	-0.3	14.1	13.8	14.6	0.5
KW-Cor	-3.0	-3.2	-3.1	0.2	4.4	4.8	6.3	1.4
KW-Cal-Cor	-1.5	-1.7	-1.8	0.2	2.9	3.5	5.4	1.4
KW-Earn	-13.1	-12.8	-13.4	-0.4	13.2	12.9	13.6	0.6
ALP	-14.2	-13.7	-14.2	-0.5	14.2	13.8	14.3	0.5
Wgt_Reg_MI	-14.0	-13.8	-14.4	-0.3	14.1	13.8	14.4	0.3
DR_wgt	-14.0	-13.8	-14.4	-0.3	14.1	13.8	14.4	0.3
HD_MI	-14.0	-13.6	-14.2	-0.5	14.2	13.8	16.0	1.0
Dual-frame design								
SP	-7.6	-7.3	-7.7	-0.3	7.6	7.3	8.2	0.4
SP_Cal	-7.6	-7.3	-7.7	-0.3	7.6	7.3	8.2	0.4
Cut-off design								
CO+BD	-7.5	-7.4	-7.7	-0.1	7.6	7.5	9.6	0.4
CO_Cal+KWFr	-2.5	-2.5	-2.6	0.0	2.6	2.6	6.2	0.4
CO_Cal+KWFr-Cor	-0.7	-0.8	-0.6	0.1	1.2	1.2	5.8	0.6
Big data only								
AuxDiv	-15.5	-15.0	-16.4	-0.6	15.5	15.0	16.4	0.6
KWFr	-15.1	-14.9	-15.4	-0.3	15.1	14.9	15.4	0.3
KWFr-Cor	-2.7	-2.9	-2.8	0.2	3.6	4.2	5.6	1.3

Note: Definitions for the different estimators are given in Tables 2.4-2.8.

2.3.4 Big data only results

When aggregate auxiliary information is available from the population it can be used to help reduce selection bias, as evidenced by the performance of the AuxDiv estimator. The AuxDiv estimator does not remove all bias as it does not include all of the $\boldsymbol{x}$ variables related to the probability of inclusion in B. The presence of measurement error increases the bias of the estimate.

The KWFr estimator performs very well in the SAR without measurement error scenario

(see Table 2.11), achieving close to the lowest RRMSE for all data items. This estimator uses unit-level auxiliary information for the full population to estimate the propensity scores for the large dataset B. Most of the contribution to RRMSE comes from bias rather than variance.

Measurement error degrades the performance of the KWFr estimator (see Tables 2.12 and 2.14). Similarly, in the SNAR scenarios the estimator does not perform as well since it is not able to include the earnings variable in the propensity model. The results for the KWFr and AuxDiv estimators demonstrate that relying on the non-probability dataset itself for inference is likely to lead to biased estimates if there are inherent issues with either the measurement of its data items or incompleteness of available auxiliary information from the population frame.

2.4 Concluding remarks

A variety of approaches, ranging from weighting-based methods, to model-based or imputation approaches, to combinations of the two, have been developed to address the faults of non-probability data. The objective of this chapter is to compare how a range of these methods perform in a business survey context assuming different missingness and measurement error settings for a large non-probability dataset, and a reference probability sample available to assist. The results provide valuable insight into the usefulness of these methods under various conditions, and are important for increasing the efficiency of statistics produced while reducing respondent burden and sample sizes.

When auxiliary information, either related to y or which effectively describe the missingness mechanism in B, is available and used in the estimation process, the methods examined can effectively account for selection bias. In the most ideal scenario of SAR missingness and no measurement error in the non-probability dataset, it is not imperative to have a reference sample which overlaps with the non-probability data. When the non-probability data are big and can be linked to the population frame the use of a calibrated split population estimator provides a beneficial combination of a low reference sample size

and the best accuracy of the estimators that were considered. The KW estimator calibrated to population totals also performs very well in this scenario, assuming the working model for π_i^B holds.

When there is SNAR missingness in the non-probability dataset but no measurement error, the calibrated split population estimator still provides the best results. This approach is robust to the selection mechanism at play in the non-probability dataset, since we use the non-probability data as-is and supplement it with data from the reference sample to cover the contribution for the population not in the non-probability dataset. The estimator is more efficient for large non-probability datasets. In situations when the non-probability dataset is a small fraction of the population, for example a small web panel survey, the performance of the estimator will be closer to a GREG calculated using just the data in the reference sample.

The presence of measurement error in y on the non-probability data source affects the performance of the estimators, such that the best estimator tended not to heavily rely on the non-probability data. In our study, the GREG, RDI and QR MA estimators - all of which rely on the reference sample as the basis for estimation - tended to perform the best in the with measurement error scenarios. An advantage of the RDI estimator over the measurement error corrected estimators listed in Table 2.8 is that we can continue to use the y values in B as-is - the form of the estimator does not need to change, and remains unbiased.

The corrected cut-off estimator, which combines the cut-off sample contribution with an estimate for the excluded part of the population, also performs well under measurement error. Further research could be beneficial for determining appropriate cut-off thresholds for the reference sample which take advantage of the ability to model or estimate the contribution from the excluded part of the population based on the available non-probability data.

Collecting information on the data item(s) of interest in the probability sample would assist with developing a model to correct for measurement error. One factor to be wary of is that the measurement error model may introduce additional variability into the estimates,

so that one may choose instead to use the probability sample as the basis for inference and the non-probability sample being employed as auxiliary information only, for example using the RDI estimator.

Although the RDI estimator is never the best-performing estimator, it is (asymptotically) unbiased, and has a reasonably low RRMSE in all scenarios. The estimator is also robust to SNAR situations as well as the presence of measurement error, but its efficiency can suffer under those less-than-ideal scenarios. When data for y is available in both A and B (Type III data structure) then the use of the RDI estimator, with the addition of population level benchmarks in the calibration, could be a relatively safe, low-risk approach which yields some gains while potentially not maximising them. I note the need to ensure the total number of benchmarks applied does not become excessively large. One could also combine the RDI estimator with a split-population approach in the design of A to obtain a more efficient reference sample.

Many of the methods that have been developed make two assumptions: *ignorability*, which implies a SAR scenario in the non-probability data source, and *common support*, which implies that the probability of being in the non-probability dataset B is non-zero for $i \in \mathcal{U}$. These assumptions may not necessarily hold. Methods are being developed to address the SNAR missingness scenario - see for example Marella (2023) and Kim and Morikawa (2023). Chen et al. (2023) consider approaches to estimation when the common support assumption is violated and some part of the population does not have any chance of being selected into the non-probability sample. This may occur, for instance, with some social media datasets. In Chapter 5 I further explore the issue of SNAR in B.

Including all y data items in the propensity score model is not necessarily helpful as a catch-all approach to include potentially helpful covariates to explain SNAR non-response. Testing for the presence of a SNAR mechanism and determining which y data item(s) need to be included in a SNAR model remain important issues. Beaumont et al. (2024b) have recently proposed a variable selection procedure for non-probability IPW methods based on a modified Akaike Information Criterion (AIC). Such an approach could potentially be used to identify the y variables to be included in a model. Some other potential

directions include work by Meng (2018, 2022), who has put forward the notion of *data defect correlation* and suggested to miniaturise this quantity to eliminate bias, and the index proposed by Andridge et al. (2019) to assess the magnitude of nonignorable selection bias.

Chapter 3

Overlap-Excluded Estimation of Participation Probabilities

3.1 Introduction

As a byproduct of the ongoing digitalisation of society, there is an increasing amount of data being collected about individuals, entities and the environment. Some examples of this data include mobile device location information, satellite imagery, commercial transactions, administrative byproducts, and web panel surveys (see for example Tam and Van Halderen 2020, Beaumont 2020, and Tam and Clarke 2015). Many of these datasets are primarily generated for non-statistical purposes, and so may not accurately represent the population of interest. In particular, they may suffer from selection bias where some subgroups of the target population are less likely to be included in the dataset. When this occurs, they may be better considered non-probability samples. Using these datasets directly without adjustment to produce inferences about population parameters such as totals and means is likely to result in biased estimates (see for example Bethlehem 2010 and Meng 2018).

Various approaches have been developed to address selection bias in non-probability data. Chapter 1 briefly reviewed the methods that have been developed, while Chapter 2 com-

pared the performance of a cross-section of these methods under various data quality assumptions.

The focus of this chapter is on the quasi-randomisation approach to inference for non-probability samples, in the case when the data item of interest is available on the non-probability dataset only, and auxiliary data items are available on both the non-probability dataset and a separate reference probability sample. This differs from the usual notion of an auxiliary variable in survey sampling, where data for these variables are available for the full population.

The quasi-randomisation approach assumes that the missingness in the non-probability sample is governed by an unknown selection mechanism that may be estimated via a model using available auxiliary information. The estimated model is applied to the units in the non-probability sample to produce estimated propensities of selection, also known as participation probabilities. The inverse of these propensities are used as weights to form an IPW estimator for the non-probability sample to estimate finite population quantities.

Valliant and Dever (2011), Elliott and Valliant (2017) and Wang et al. (2021a) pooled the auxiliary data from the non-probability and probability samples to perform their model fitting. Chen et al. (2020) proposed a pseudo-likelihood approach which does not require the samples to be pooled together for the model fitting. Burakauskaitė and Čiginas (2023) looked at the Chen et al. (2020) approach for the case when auxiliary data for the full population is available to assist with estimating the participation probabilities. These methods do not require any linkage of the data from the two samples.

In some scenarios, it may be possible to identify the overlap between the reference sample and the non-probability dataset. This would occur, for example, if a common unit identifier is available (for example, a unique business or person number), through the addition of questions on one sample asking about membership in the other, or through linking using common matching variables. Some approaches take advantage of this information to estimate the propensities. By linking the indicator of membership in the non-probability dataset to the probability sample, Kim and Wang (2019) maximised a pseudo-likelihood based on the probability sample alone to estimate the parameters of the participation

probability model. Liu et al. (2022) extended the Elliott and Valliant (2017) method by using the units present in one of the two samples but not both. Beaumont et al. (2024a) proposed a sample likelihood approach which uses the pooled sample but excludes duplicate units from the overlapping sample to estimate the participation probability model parameters. Kim and Kwon (2024) proposed a similar approach which they referred to as conditional likelihood, again using the pooled sample but excluding the duplicate units from the overlapping sample.

In this context, this chapter makes a contribution to the literature on inference for non-probability samples by proposing an Overlap-Excluded (OE) pseudo-likelihood approach to estimating participation probabilities when the overlap with a reference sample is known. The other major contributions in this chapter include a theoretical and empirical comparison of the efficiency of population estimates using the OE approach relative to existing methods. It is shown theoretically that the OE approach is more efficient compared to the methods of Chen et al. (2020) and Kim and Wang (2019). Further, a real-world application using real business taxation data illustrates the behaviour of competing methods under skewed populations and different reference sample sampling schemes. When it is not possible to identify the overlapping units, an Expectation-Maximisation (EM) algorithm is constructed to simultaneously estimate the overlap and produce the maximum pseudo-likelihood estimator of the participation probability model parameters. Empirical comparisons of these two approaches against other methods under a simulation setting and a real-data scenario demonstrate the superiority of the OE method especially when the non-probability dataset is large, and when the inclusion probabilities of units into the reference sample cannot be considered mutually independent.

While the method presented in this chapter places importance on the non-overlapping probability sample units for constructing the participation probabilities, it is noted that the overlapping units may also hold potential value for inferential purposes. One case is when measurement error is present in some of the auxiliary data items for one of the samples. This may occur, for example, in the probability sample if an out-of-date sample is being used. The overlapping sample may thus be used to correct the measurement error in the affected items. This aspect is out of scope of the chapter. Neither does this chapter

consider the potential value of an overlapping sample when estimates of change from one time period to another are of interest.

The remainder of the chapter is structured as follows. Given the basic setup provided in Section 1.1, Section 3.2 introduces the Overlap-Excluded propensity score estimator which is applicable when the overlap between the probability and non-probability samples can be determined. Section 3.3 provides details on the variance of the estimator, and compares its efficiency with several alternative methods under the assumption that the reference sample is a Poisson sample. Section 3.4 describes the use of the EM algorithm to simultaneously estimate the overlapping sample and participation probabilities when the overlap between the non-probabiliy sample and reference sample is not known. In Section 3.5 the performance of the proposed estimators is compared through a simulation study. The methods are applied using real-world data in Section 3.6. Section 3.7 provides some summary remarks.

3.2 Overlap-Excluded Propensity Score Estimation

In this chapter, the end goal is to produce an estimate of the population mean using B. This can be achieved by using an IPW estimator such as that suggested by Chen et al. (2020):

$$\hat{\mu}_{IPW} = \frac{1}{\hat{N}_B} \sum_{i \in B} \frac{y_i}{\hat{\pi}_i^B} \tag{3.1}$$

where $\hat{N}_B = \sum_{i \in B} (\hat{\pi}_i^B)^{-1}$. The problem addressed in this chapter is how best to produce $\hat{\pi}_i^B$ for feeding into (3.1).

It is assumed that the π_i^B are governed by a parametric model, so that $\pi_i^B = \pi(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ are the unknown (true) model parameters. The population likelihood function of $\boldsymbol{\theta}$ over $\mathcal{U}$ is

$$L(\boldsymbol{\theta}) = \prod_{i \in \mathcal{U}} (\pi(\boldsymbol{x}_i, \boldsymbol{\theta}))^{\delta_i} (1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta}))^{(1-\delta_i)} \tag{3.2}$$

and the corresponding log-likelihood function is

$$l(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \delta_i \log(\pi(\boldsymbol{x}_i, \boldsymbol{\theta})) + \sum_{i \in \mathcal{U}} (1 - \delta_i) \log(1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})). \tag{3.3}$$

If $\boldsymbol{x}$ is available for the full population, for example from a census of the population or from a population frame, then $\hat{\boldsymbol{\theta}}$ can be produced directly by maximising $l(\boldsymbol{\theta})$ to yield the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. These model parameter estimates can then be used to produce estimated participation probabilities $\hat{\pi}_i^B = \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$.

The focus in this thesis is on the case when $\boldsymbol{x}$ is not available for the full population. The second term in (3.3) is then not observed. However, we can expand the term, which yields an alternate expression for the log-likelihood:

$$l(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \delta_i \log(\pi(\boldsymbol{x}_i, \boldsymbol{\theta})) + \sum_{i \in \mathcal{U}} \log(1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})) - \sum_{i \in \mathcal{U}} \delta_i \log(1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})). \tag{3.4}$$

It can be seen from (3.4) that the first and third terms of the log-likelihood can be obtained directly from B. Chen et al. (2020) estimate the second term using the reference sample A. Their approach, which I label as CLW, has a resulting pseudo-likelihood given by

$$l_{CLW}(\boldsymbol{\theta}) = \sum_{i \in B} \log \frac{\pi(\boldsymbol{x}_i, \boldsymbol{\theta})}{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})} + \sum_{i \in A} d_i^A \log(1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})). \tag{3.5}$$

Note that the second term in (3.3) is actually a sum over the population units that are not in B. Under the assumption that it is possible to accurately identify which units in A are not in B, and further assuming that $A \setminus B$ is a representative sample from $\mathcal{U} \setminus B$, this term can be estimated using $A \setminus B$. The resulting pseudo-likelihood is given by

$$l_{OE}(\boldsymbol{\theta}) = \sum_{i \in B} \log(\pi(\boldsymbol{x}_i, \boldsymbol{\theta})) + \sum_{i \in A \setminus B} d_i^A \log(1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})). \tag{3.6}$$

The design expectation of (3.6) conditional on B will be equal to (3.3) if $\pi_i^A = P(i \in A | B)$, or if the two samples A and B are conditionally independent, that is, $P(i \in A | B) =$

$P(i \in A|\boldsymbol{q})$. Conditional independence means that knowledge of B contains no further information about the inclusion probabilities for A once the design variables $\boldsymbol{q}$ are taken into account.

In this thesis, $l_{OE}(\boldsymbol{\theta})$ will be referred to as the Overlap-Excluded (OE) pseudo-likelihood. In a similar manner, the estimated participation probabilities $\pi(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$ will be referred to as the OE probabilities or propensities, and the IPW estimates created using the OE probabilities as OE IPW estimates.

It is worth pointing out the connection between the CLW and OE approaches. Chen et al. (2020) estimate the quantity $\sum_{i\in\mathcal{U}} \log(1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta}))$ in (3.4) by the Horvitz-Thompson estimator $\sum_{i\in A} d_i^A \log(1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta}))$. However, it can be seen from (3.6) that the OE approach is essentially estimating that same quantity using

$$\sum_{i\in B} \log(1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta})) + \sum_{i\in A\setminus B} d_i^A \log(1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta})).$$

This is the dual-frame estimator from Kim and Tam (2021), which is expected to be more efficient than the Horvitz-Thompson estimator, especially when B is large and when the overlap between A and B is significant.

Remark 3.1. l_{OE} bears similarities to a number of other approaches proposed in the literature which also exclude the overlapping units. Kim and Wang (2019) propose a data integration propensity score estimator which has a pseudo-likelihood of the form

$$l_{KW}(\boldsymbol{\theta}) = \sum_{i\in A} d_i^A \delta_i \log(\pi(\boldsymbol{x}_i,\boldsymbol{\theta})) + \sum_{i\in A} d_i^A (1-\delta_i)\log(1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta})) \tag{3.7}$$

where the first term of (3.7) is estimated using A instead of being computed directly from B, as I have proposed. Their approach also requires knowledge of $\delta_i, i \in A$. The sample likelihood (SL) approach described in Beaumont et al. (2024a) utilises the sample $A \cup B$ containing the unique elements from each sample - duplicates from the overlapping sample are removed. The approach makes use of the conditional probability $p_{si} = P(\delta_i = 1|i \in$

$A \cup B, \boldsymbol{x}_i) = \pi_i^B/(\pi_i^B + \pi_i^A - \pi_i^B\pi_i^A)$. The log-likelihood of $\boldsymbol{\theta}$ for this approach is

$$\begin{aligned} l_{SL}(\boldsymbol{\theta}) &= \sum_{i \in A \cup B} \Big[\delta_i \log(p_s(\boldsymbol{x}_i;\boldsymbol{\theta})) + (1-\delta_i)\log(1-p_s(\boldsymbol{x}_i;\boldsymbol{\theta}))\Big] \\ &= \sum_{i \in B} \log(p_s(\boldsymbol{x}_i;\boldsymbol{\theta})) + \sum_{i \in A} \log(1-p_s(\boldsymbol{x}_i;\boldsymbol{\theta})) - \sum_{i \in A \cap B} \log(1-p_s(\boldsymbol{x}_i;\boldsymbol{\theta})). \end{aligned} \tag{3.8}$$

The SL approach estimates the parameters $\boldsymbol{\theta}$ through $p_s = \pi(\boldsymbol{x}_i,\boldsymbol{\theta})/(\pi(\boldsymbol{x}_i,\boldsymbol{\theta}) + \pi_i^A - \pi(\boldsymbol{x}_i,\boldsymbol{\theta})\pi_i^A)$. It is also unweighted and does not include a d_i^A term. The SL method assumes that $A \perp B$.

Remark 3.2. In contrast to the OE, CLW, KW and SL approaches, the Adjusted Logistic Propensity (ALP) approach described in Wang et al. (2021a) takes a "stacked sample" approach whereby all the sample units in A and B, including overlapping sample, are stacked together. Each element of the stacked sample is assumed to be distinct even though the overlapping sample will be present twice, once in B and once in A. The approach has the pseudo log-likelihood

$$l_{ALP}(\boldsymbol{\theta}) = \sum_{i \in B} \log(p_i) + \sum_{i \in A} d_i^A \log(1-p_i) \tag{3.9}$$

where $p_i = \pi(\boldsymbol{x}_i,\boldsymbol{\theta})/(1+\pi(\boldsymbol{x}_i,\boldsymbol{\theta}))$ is the modelled probability of being in the sample B given the unit is in the stacked sample.

Remark 3.3. Both $\hat{\boldsymbol{\theta}}$ and $\pi(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}})$ depend on having accurate measurements of $\boldsymbol{x}$ in both A and B. However, covariate measurement error may occur in either of A or B, through reasons such as out-of-date data, data collected using different definitions, and data collected using varying quality standards. The presence of measurement error in some or all of the covariates $\boldsymbol{x}$ in A or B may result in biased parameter estimates of $\hat{\boldsymbol{\theta}}$ and estimated participation probabilities. The correction of measurement error issues is left to Chapter 5, but it is noted here the potential usefulness of having some overlapping sample $A \cap B$ to assist with addressing these issues.

Under a logistic regression model for the propensity scores, $\pi_i^B = \exp(\boldsymbol{x}_i^T\boldsymbol{\theta}_0)/(1+\exp(\boldsymbol{x}_i^T\boldsymbol{\theta}_0))$.

Then the pseudo-likelihood function (3.6) becomes

$$\begin{aligned} l_{OE}(\boldsymbol{\theta}) &= \sum_{i\in B} \log\Big\{\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\theta})}\Big\} + \sum_{i\in A\setminus B} d_i^A \log\Big\{1-\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\theta})}\Big\} \\ &= \sum_{i\in B}\Big\{\boldsymbol{x}_i^T\boldsymbol{\theta} - \log(1+\exp(\boldsymbol{x}_i^T\boldsymbol{\theta}))\Big\} - \sum_{i\in A\setminus B} d_i^A \log(1+\exp(\boldsymbol{x}_i^T\boldsymbol{\theta})). \end{aligned} \tag{3.10}$$

The score function $S(\boldsymbol{\theta})$ is then given by

$$\begin{aligned} S(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}} l_{OE}(\boldsymbol{\theta}) \\ &= \sum_{i\in B}\boldsymbol{x}_i - \sum_{i\in B}\pi(\boldsymbol{x}_i,\boldsymbol{\theta})\boldsymbol{x}_i - \sum_{i\in A\setminus B} d_i^A\pi(\boldsymbol{x}_i,\boldsymbol{\theta})\boldsymbol{x}_i. \end{aligned}$$

The maximum pseudo-likelihood estimator $\hat{\boldsymbol{\theta}}$ can be obtained by solving $S(\boldsymbol{\theta}) = \mathbf{0}$. A solution can be found by applying an iterative procedure such as Newton-Raphson:

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \Big\{H(\boldsymbol{\theta}^{(m)})\Big\}^{-1} S(\boldsymbol{\theta}^{(m)})$$

where

$$\begin{aligned} H(\boldsymbol{\theta}^{(m)}) &= -\frac{\partial^2}{\partial\boldsymbol{\theta}^2} l_{OE}(\boldsymbol{\theta}) \\ &= \sum_{i\in B}\pi(\boldsymbol{x}_i,\boldsymbol{\theta})(1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta}))\boldsymbol{x}_i\boldsymbol{x}_i^T + \sum_{i\in A\setminus B} d_i^A\pi(\boldsymbol{x}_i,\boldsymbol{\theta})(1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta}))\boldsymbol{x}_i\boldsymbol{x}_i^T. \end{aligned}$$

The OE participation probabilities for $i \in B$ are obtained from the maximum pseudo-likelihood estimates of $\hat{\boldsymbol{\theta}}$ through

$$\hat{\pi}_i^B = \frac{\exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\theta}})}{(1+\exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}))} \tag{3.11}$$

These participation probabilities can be fed into (3.1) to produce an estimate of the population mean. Alternatively, the participation probabilities may be used in a doubly robust manner such as described in Chen et al. (2020), or in multiply robust approaches such as those described in Chen and Haziza (2023).

3.3 Consistency and variance of the OE IPW estimator

In order to examine the consistency and variance of the OE IPW estimator, an asymptotic framework used by other authors (see for example Chen et al. 2020, Wang et al. 2021a and Scholtus et al. 2025) is first adopted. Within this framework, suppose there is a sequence of finite populations $\mathcal{U}_t$ of size N_t indexed by t. Associated with each $\mathcal{U}_t$ are a probability sample A_t of size $n_{A,t}$ and non-probability sample B_t of size $N_{B,t}$. The population size $N_t \to \infty$, while the sample sizes $n_{A,t} \to \infty$ and $N_{B,t} \to \infty$ as $t \to \infty$. In what follows, the index t is omitted, and the limiting process is represented by $N \to \infty$. The properties of the estimator under the model for the propensity scores and the probability sample design are described.

Theorem 3.1. *When assumptions **A1** to **A3** from Section 1.1 and conditions **C1** to **C4** in Appendix B.1 hold, I_i and δ_i are independent for all i, and a logistic participation probability model holds, $E(\hat{\mu}_{OE}) \to \mu_y$ and $|\hat{\mu}_{OE} - \mu_y| \xrightarrow{p} 0$ as the sizes of the population and samples tend toward infinity. Further, the variance of the OE IPW estimator, denoted by $Var(\hat{\mu}_{OE})$, is given as*

$$Var(\hat{\mu}_{OE}) = \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-\pi_i^B)\pi_i^B\Big[\frac{(y_i-\mu_y)}{(\pi_i^B)} - \boldsymbol{b}^T\boldsymbol{x}_i\Big]^2 + \tag{3.12}$$

$$\boldsymbol{b}^T\boldsymbol{V}\boldsymbol{b} + \tag{3.13}$$

$$\frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big[(\pi_i^B)^3(1-\pi_i^B)(d_i^A-1)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big] \tag{3.14}$$

where $\pi_i^B = \pi(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$, $\boldsymbol{b}^T = \sum_{i\in\mathcal{U}}(1-\pi_i^B)(y_i-\mu_y)\boldsymbol{x}_i^T\Big[\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1}$, $\boldsymbol{V} = \frac{1}{N^2}Var_d\Big(\sum_{i\in A}d_i^A\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\Big)$, *and Var_d refers to the design-based variance under the probability sample design for A.*

Details of the derivations for consistency and variance can be found in Appendix B.2.

A plug-in estimator can be constructed to estimate the variance of the OE IPW estimator

(3.1):

$$\widehat{Var}(\hat{\mu}_{OE}) = \frac{1}{N^2}\sum_{i\in B}(1-\hat{\pi}_i^B)\Big[\frac{(y_i-\hat{\mu}_{IPW})}{(\hat{\pi}_i^B)} - \hat{\boldsymbol{b}}^T\boldsymbol{x}_i\Big]^2 + \tag{3.15}$$

$$\hat{\boldsymbol{b}}^T\hat{\boldsymbol{V}}\hat{\boldsymbol{b}} + \tag{3.16}$$

$$\frac{1}{N^2}\sum_{i\in B}\Big[(\hat{\pi}_i^B)^2(1-\hat{\pi}_i^B)(d_i^A-1)\hat{\boldsymbol{b}}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\hat{\boldsymbol{b}}\Big] \tag{3.17}$$

where $\hat{\boldsymbol{b}}^T = \sum_{i\in B}(1/\hat{\pi}_i^B-1)(y_i-\hat{\mu}_{IPW})\boldsymbol{x}_i^T\Big[\sum_{i\in B}(1-\hat{\pi}_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1}$ is an estimate of $\boldsymbol{b}^T$, $\hat{\pi}_i^B$ is the estimated participation probability for unit i obtained using $\hat{\boldsymbol{\theta}}$ which is "plugged in" for each unit i as required in A and B, $\hat{\mu}_{IPW}$ is the IPW estimate of mean for the variable of interest given by (3.1), and $\hat{\boldsymbol{V}} = \frac{1}{N^2}\widehat{Var}_d\Big(\sum_{i\in A} d_i^A\hat{\pi}_i^B(1-\hat{\pi}_i^B)\boldsymbol{x}_i\Big)$ is the design-based variance of the quantity $\sum_{i\in\mathcal{U}}\hat{\pi}_i^B(1-\hat{\pi}_i^B)\boldsymbol{x}_i$ in A, estimated using standard methods. When B is small relative to A, then $\hat{\boldsymbol{b}}^T = \sum_{i\in B}(1/\hat{\pi}_i^B-1)(y_i-\hat{\mu}_{IPW})\boldsymbol{x}_i^T\Big[\sum_{i\in A} d_i^A\hat{\pi}_i^B(1-\hat{\pi}_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1}$ may be used instead (see Chen et al. 2020).

Component (3.17) requires knowledge of the reference sample design weights of units in B, and this may not be available. In this case, an alternative plug-in estimator for (3.17) is

$$\frac{1}{N^2}\sum_{i\in A}(d_i^A)^2\Big[(\hat{\pi}_i^B)^3(1-\hat{\pi}_i^B)\hat{\boldsymbol{b}}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\hat{\boldsymbol{b}}\Big] - \frac{1}{N^2}\sum_{i\in B}\Big[(\hat{\pi}_i^B)^2(1-\hat{\pi}_i^B)\hat{\boldsymbol{b}}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\hat{\boldsymbol{b}}\Big]. \tag{3.18}$$

3.3.1 Comparison of variance against alternative methods

The OE pseudo-likelihood aims to take advantage of knowledge about $\delta_i, i \in A$ to use only those units in A that are required. As such, it may be expected to result in a lower variance compared to other methods which use the full A sample. In this subsection a comparison is made between the variance of the OE IPW estimator with the variances from the Hájek-like IPW2 estimator (which I denote by $Var(\hat{\mu}_{CLW})$) in Chen et al. (2020), the data integration IPW estimator ($Var(\hat{\mu}_{KW})$) in Kim and Wang (2019), and the ALP method ($Var(\hat{\mu}_{ALP})$) in Wang et al. (2021a).

Unlike the OE, ALP, Chen et al. (2020) and Kim and Wang (2019) methods which employ a pseudo-likelihood approach to analysis, the SL approach proposed by Beaumont et al. (2024a) uses a likelihood conditional on the observed combined sample. This makes it more difficult to directly compare the variances of the OE and SL methods. In this chapter, I limit myself to the following remarks. Firstly, by avoiding the need to use the reference sample weights in the sample likelihood, the SL approach should lead to more efficient estimation of $\boldsymbol{\theta}$ compared with the pseudo-likelihood approaches. However, the approach also requires more restrictive assumptions to hold. In particular, it is assumed that the samples A and B are mutually independent so that $P(i \in A \cap B) = P(i \in A)P(i \in B) = \pi_i^A \pi_i^B$. In practice, this may be achieved by ensuring that there is independence in the sample design and the selection mechanism (for example if both A and B are drawn as independent Poisson samples), or when the inclusion probabilities into the two samples do not depend on shared (observed or unobserved) information. These assumptions are explored further in the empirical study in Section 3.6.

To simplify variance comparisons, it is assumed that the reference sample A has been drawn using Poisson sampling. When this is the case, $\boldsymbol{V}$ in Theorem 3.1 simplifies, so that the terms (3.13) and (3.14) can be combined, giving

$$
\begin{aligned}
Var(\hat{\mu}_{OE}) = & \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-\pi_i^B)\pi_i^B\Big(\frac{y_i-\mu_y}{\pi_i^B}-\boldsymbol{b}^T\boldsymbol{x}_i\Big)^2+ \\
& \boldsymbol{b}^T\boldsymbol{V}\boldsymbol{b}+\frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big(\frac{1}{\pi_i^A}-1\Big)(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \\
= & \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-\pi_i^B)\pi_i^B\Big(\frac{y_i-\mu_y}{\pi_i^B}-\boldsymbol{b}^T\boldsymbol{x}_i\Big)^2+ \\
& \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big(\frac{1}{\pi_i^A}-1\Big)(\pi_i^B)^2(1-\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}+ \\
& \frac{1}{N^2}\sum_{i\in\mathcal{U}}(\frac{1}{\pi_i^A}-1)(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \\
= & \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-\pi_i^B)\pi_i^B\Big(\frac{y_i-\mu_y}{\pi_i^B}-\boldsymbol{b}^T\boldsymbol{x}_i\Big)^2+ \\
& \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big(\frac{1}{\pi_i^A}-1\Big)(\pi_i^B)^2(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}
\end{aligned}
\tag{3.19}
$$

Corollary 3.1. *Under Poisson sampling for A, $Var(\hat{\mu}_{CLW}) \geq Var(\hat{\mu}_{OE})$ holds.*

Proof. Under Poisson sampling for A and using equation (10) of Chen et al. (2020), the variance of their Hájek-like estimator IPW2 becomes

$$Var(\hat{\mu}_{CLW}) = \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-\pi_i^B)\pi_i^B\Big(\frac{y_i-\mu_y}{\pi_i^B} - \boldsymbol{b}^T\boldsymbol{x}_i\Big)^2 + \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big(\frac{1}{\pi_i^A}-1\Big)(\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \tag{3.20}$$

Note that $\boldsymbol{b}$ is defined the same way in both methods and so the first terms of (3.20) and (3.19) are identical. When taking the difference between $Var(\hat{\mu}_{CLW})$ and $Var(\hat{\mu}_{OE})$, these terms thus cancel out, giving

$$\begin{aligned}
&Var(\hat{\mu}_{CLW}) - Var(\hat{\mu}_{OE}) \\
&= \frac{1}{N^2}\sum_{i\in\mathcal{U}}(\frac{1}{\pi_i^A}-1)(\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} - \frac{1}{N^2}\sum_{i\in\mathcal{U}}(\frac{1}{\pi_i^A}-1)(\pi_i^B)^2(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \\
&= \frac{1}{N^2}\sum_{i\in\mathcal{U}}(\frac{1}{\pi_i^A}-1)(\pi_i^B)^3\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \geq 0
\end{aligned}$$

□

The OE IPW estimator is thus always at least as efficient as the Chen et al. (2020) estimator when the reference sample is a Poisson sample. Its efficiency increases when $\{\pi_i^A\}$ are smaller, and when $\{\pi_i^B\}$ are larger. When A is a census of the population, the two methods achieve the same variance.

In order to compare the OE IPW estimator with the data integration estimator in Kim and Wang (2019), an explicit variance expression for their data integration estimator under a logistic regression model is needed. Theorem 3.2 provides an asymptotic variance expression for their estimator. The proof of the theorem is included in Appendix B.3.

Theorem 3.2. *When assumptions **A1** to **A3** from Section 1.1 hold, δ_i and I_i are independent, and a logistic regression model for the participation probabilities holds, the variance of the data integration estimator from Kim and Wang (2019) is given by*

$$Var(\hat{\mu}_{KW}) = \frac{1}{N^2}\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\Big(\frac{y_i-\mu}{\pi_i^B} - \boldsymbol{b}^T\boldsymbol{x}_i\Big)^2 + \frac{1}{N^2}\sum_{i\in\mathcal{U}}(d_i^A-1)\pi_i^B(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \tag{3.21}$$

where $\boldsymbol{b}^T$ *is defined as*

$$\boldsymbol{b}^T = \sum_{i\in\mathcal{U}}(1-\pi_i^B)(y_i-\mu)\boldsymbol{x}_i^T\Big[\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1}$$

Corollary 3.2. *Under Poisson sampling for A, the inequality* $Var(\hat{\mu}_{KW}) \geq Var(\hat{\mu}_{OE})$ *holds.*

Proof. Noting that $\boldsymbol{b}^T$ is defined identically in $Var(\hat{\mu}_{KW})$ and $Var(\hat{\mu}_{OE})$ (see Theorem 3.1), the first terms in (3.19) and (3.21) are the same, and will cancel out when $Var(\hat{\mu}_{OE})$ is subtracted from $Var(\hat{\mu}_{KW})$. Then,

$$\begin{aligned}
&Var(\hat{\mu}_{KW}) - Var(\hat{\mu}_{OE}) \\
&= \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big(\frac{1}{\pi_i^A}-1\Big)\pi_i^B(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} - \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big(\frac{1}{\pi_i^A}-1\Big)(\pi_i^B)^2(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \\
&= \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big(\frac{1}{\pi_i^A}-1\Big)\pi_i^B(1-\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \\
&\geq 0
\end{aligned}$$

□

From Wang et al. (2021b), the variance of the ALP estimator under Poisson sampling in A is

$$\begin{aligned}
Var(\hat{\mu}_{ALP}) = \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-2p_i)p_i\Big(\frac{y_i-\mu_y}{p_i} - \boldsymbol{b}_W^T\boldsymbol{x}_i\Big)^2 + \\
\frac{1}{(N+N_B)^2}\sum_{i\in\mathcal{U}}(\frac{1}{\pi_i^A}-1)p_i^2\boldsymbol{b}_W^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}_W
\end{aligned} \tag{3.22}$$

where $p_i = \pi_i^B/(1+\pi_i^B)$ and $\boldsymbol{b}_W^T = \sum_{i\in\mathcal{U}}(y_i-\mu)\boldsymbol{x}_i^T\Big[\sum_{i\in\mathcal{U}}p_i\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1}$.

Making a direct comparison between the variance of the OE and ALP methods using (3.22) is complicated by the use of p_i instead of π_i^B and $\boldsymbol{b}_W^T$ instead of $\boldsymbol{b}^T$. To get an initial indication of the relative efficiency of the two approaches, I consider the variance of the score functions of the two methods conditional on the sample B.

For an arbitrary parametric model for π_i^B, the ALP pseudo-estimating function can be expressed as (see Beaumont et al. 2024a)

$$S_{ALP}(\boldsymbol{\theta}) = \sum_{i \in B} \frac{1}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} \frac{\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{1+\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} - \sum_{i \in A} d_i^A \frac{\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{1+\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} \tag{3.23}$$

where $\pi'(\boldsymbol{x}_i;\boldsymbol{\theta}) = \partial\pi(\boldsymbol{x}_i;\boldsymbol{\theta})/\partial\theta$. In Wang et al. (2021a) the model $\pi(\boldsymbol{x}_i,\boldsymbol{\theta}) = \exp(\boldsymbol{x}_i^T\boldsymbol{\theta})$ is considered. It can be seen that (3.23) is the difference of two independent Horvitz-Thompson estimators of the same quantity. Assuming that both A and B are drawn independently of each other using Poisson sampling, the variance of (3.23) conditional on the sample B can be expressed as

$$Var(S_{ALP}(\boldsymbol{\theta})|B) = \sum_{i \in \mathcal{U}} \frac{1-\pi_i^A}{\pi_i^A} \Big(\frac{\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{1+\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} \Big)^2. \tag{3.24}$$

Now, the estimating function for the OE approach can be expressed as

$$S_{OE}(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \Big(\frac{\delta_i}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} \Big) \pi'(\boldsymbol{x}_i;\boldsymbol{\theta}) - \sum_{i \in \mathcal{U}} \frac{d_i^A I_i (1-\delta_i) \pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} \tag{3.25}$$

$$= \sum_{i \in B} \frac{\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))} - \sum_{i \in \mathcal{U}} \delta_i \frac{\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} - \left[\sum_{i \in A} d_i^A \frac{\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} - \sum_{i \in A} d_i^A \delta_i \frac{\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} \right]. \tag{3.26}$$

The last three terms on the right hand side of (3.26) are in the form of a difference estimator where the "auxiliary information" available for the population is $\delta_i \pi'(\boldsymbol{x}_i;\boldsymbol{\theta})/(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))$, and the "variable of interest" is $\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})/(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))$. As the size of B increases the correlation between these variables should increase and the estimating function should become more efficient. The sampling variance of the OE estimating function, conditional on B, is

$$Var(S_{OE}(\boldsymbol{\theta})|B) = \sum_{i \in \mathcal{U} \setminus B} \frac{1-\pi_i^A}{\pi_i^A} \Big(\frac{\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} \Big)^2. \tag{3.27}$$

In comparing (3.24) and (3.27), one can see that

$$\begin{aligned} &Var(S_{OE}(\boldsymbol{\theta})|B) - Var(S_{ALP}(\boldsymbol{\theta})|B) \\ &= \sum_{i\in\mathcal{U}} \frac{1-\pi_i^A}{\pi_i^A}\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})^2\Big(\frac{1-\delta_i}{(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))^2} - \frac{1}{(1+\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))^2}\Big). \end{aligned} \tag{3.28}$$

For unit $i \in \mathcal{U}$, the term in brackets will be negative when $\delta_i = 1$, that is, when the unit i is part of the non-probability sample B. When i is not part of the non-probability sample, then $\delta_i = 0$ and the term in brackets will be positive. As such, when B is large enough, (3.28) will tend to be negative and so the conditional variance $Var(S_{OE}(\boldsymbol{\theta})|B)$ will be smaller than $Var(S_{ALP}(\boldsymbol{\theta})|B)$.

A fuller comparison of the relative efficiency of the OE and ALP approaches should take into account both selection mechanisms and consider the variance of the estimates of mean $\hat{\mu}_{OE}$ and $\hat{\mu}_{ALP}$. To facilitate this, I present Theorem 3.3, which provides a general expression for $Var(\hat{\mu}_{IPW})$ in terms of the unit-level terms in the score function, under Poisson sampling for A and B.

Theorem 3.3. *Under the regularity conditions in Appendix B.1, with $-E(\partial S(\boldsymbol{\theta})/\partial\boldsymbol{\theta})$ being positive definite, and A and B being selected independently using Poisson sampling,*

$$Var(\hat{\mu}(\hat{\boldsymbol{\theta}})) \approx \sum_{i\in\mathcal{U}} Var\Big[\frac{\delta_i}{\pi_i^B}\Big(\frac{y_i-\mu_y}{N}\Big) - \boldsymbol{\beta}^T\boldsymbol{s}_i\Big] \tag{3.29}$$

where $\hat{\mu}(\hat{\boldsymbol{\theta}})$ is an IPW estimator of the form described in (3.1), $\boldsymbol{\beta}^T = 1/N\sum_{i\in\mathcal{U}}(\pi_i'^B)^T/\pi_i^B\Big(y_i - \mu_y\Big)\boldsymbol{G}^{-1}$, $\boldsymbol{G}$ being $-E(\partial S(\boldsymbol{\theta})/\partial\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$, $\pi_i^B = \pi(\boldsymbol{x}_i;\boldsymbol{\theta}_0)$, $\pi_i'^B$ is the first derivative of $\pi(\boldsymbol{x}_i;\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$, and $\boldsymbol{s}_i$ are the individual unit i terms in $S(\boldsymbol{\theta}_0)$.

Details of the derivation are provided in Appendix B.4.

Corollary 3.3. *Under the conditions specified in Theorem 3.3,*

$$Var(\hat{\mu}_{ALP}) \approx \sum_{i\in\mathcal{U}}\Big[\pi_i^B(1-\pi_i^B)\Big(\frac{y_i-\mu_y}{\pi_i^B N} - \boldsymbol{\beta}_A^T\boldsymbol{t}_i\Big)^2 + \frac{1-\pi_i^A}{\pi_i^A}(\pi_i^B)^2\Big(\boldsymbol{\beta}_A^T\boldsymbol{t}_i\Big)^2\Big] \tag{3.30}$$

where $\boldsymbol{\beta}_A^T$ is defined similarly to $\boldsymbol{\beta}^T$ except that it uses $\boldsymbol{G}_A^{-1}$, with $\boldsymbol{G}_A$ being $-E(\partial S_{ALP}(\boldsymbol{\theta})/\partial\boldsymbol{\theta})$

evaluated at $\boldsymbol{\theta}_0$, *and* $\boldsymbol{t}_i = \pi_i'^B/[\pi_i^B(1+\pi_i^B)]$.

Proof. From (3.23), $\boldsymbol{s}_i = \delta_i/\pi_i^B \boldsymbol{u}_i - I_i d_i^A \boldsymbol{u}_i$ where $\boldsymbol{u}_i = \pi_i'^B/(1+\pi_i^B)$. Then substituting this into Theorem 3.3 gives

$$\begin{aligned}
Var(\hat{\mu}_{ALP}) &\approx \sum_{i\in\mathcal{U}} Var\Big[\frac{\delta_i}{\pi_i^B}\Big(\frac{y_i-\mu_y}{N}\Big) - \boldsymbol{\beta}_A^T(\frac{\delta_i}{\pi_i^B}\boldsymbol{u}_i - I_i d_i^A \boldsymbol{u}_i)\Big] \\
&= \sum_{i\in\mathcal{U}} Var\Big[\frac{\delta_i}{\pi_i^B}\Big(\frac{y_i-\mu_y}{N} - \boldsymbol{\beta}_A^T\boldsymbol{u}_i\Big) + I_i d_i^A \boldsymbol{\beta}_A^T\boldsymbol{u}_i\Big] \\
&= \sum_{i\in\mathcal{U}} \Big[\frac{1-\pi_i^B}{\pi_i^B}\Big(\frac{y_i-\mu_y}{N} - \boldsymbol{\beta}_A^T\boldsymbol{u}_i\Big)^2 + \frac{1-\pi_i^A}{\pi_i^A}\Big(\boldsymbol{\beta}_A^T\boldsymbol{u}_i\Big)^2\Big] \\
&= \sum_{i\in\mathcal{U}} \Big[\pi_i^B(1-\pi_i^B)\Big(\frac{y_i-\mu_y}{\pi_i^B N} - \frac{\boldsymbol{\beta}_A^T\boldsymbol{u}_i}{\pi_i^B}\Big)^2 + \frac{1-\pi_i^A}{\pi_i^A}\Big(\boldsymbol{\beta}_A^T\boldsymbol{u}_i\Big)^2\Big] \\
&= \sum_{i\in\mathcal{U}} \Big[\pi_i^B(1-\pi_i^B)\Big(\frac{y_i-\mu_y}{\pi_i^B N} - \boldsymbol{\beta}_A^T\boldsymbol{t}_i\Big)^2 + \frac{1-\pi_i^A}{\pi_i^A}(\pi_i^B)^2\Big(\boldsymbol{\beta}_A^T\boldsymbol{t}_i\Big)^2\Big]
\end{aligned}$$

□

Corollary 3.4. *Under the conditions specified in Theorem 3.3,*

$$Var(\hat{\mu}_{OE}) \approx \sum_{i\in\mathcal{U}} \Big[\pi_i^B(1-\pi_i^B)\Big(\frac{y_i-\mu_y}{\pi_i^B N} - \boldsymbol{\beta}_O^T\boldsymbol{k}_i\Big)^2 + \frac{1-\pi_i^A}{\pi_i^A}(1-\pi_i^B)(\pi_i^B)^2\Big(\boldsymbol{\beta}_O^T\boldsymbol{k}_i\Big)^2\Big] \tag{3.31}$$

where $\boldsymbol{\beta}_O^T$ *is defined similarly to* $\boldsymbol{\beta}^T$ *except that it uses* $\boldsymbol{G}_O^{-1}$, *with* $\boldsymbol{G}_O$ *being* $-E(\partial S_{OE}(\boldsymbol{\theta})/\partial\boldsymbol{\theta})$ *evaluated at* $\boldsymbol{\theta}_0$, *and* $\boldsymbol{k}_i = \pi_i'^B/[\pi_i^B(1-\pi_i^B)]$.

Proof. From (3.25), $\boldsymbol{s}_i = \delta_i\boldsymbol{v}_i - d_i^A I_i(1-\delta_i)\boldsymbol{c}_i$, where $\boldsymbol{v}_i = \pi_i'^B/\pi_i^B$ and $\boldsymbol{c}_i = \pi_i'^B/(1-\pi_i^B)$. Then substituting this into Theorem 3.3 gives

$$\begin{aligned}
Var(\hat{\mu}_{OE}) &\approx \sum_{i\in\mathcal{U}} Var\Big[\frac{\delta_i}{\pi_i^B}\Big(\frac{y_i-\mu_y}{N}\Big) - \boldsymbol{\beta}_O^T(\delta_i\boldsymbol{v}_i - d_i^A I_i(1-\delta_i)\boldsymbol{c}_i)\Big] \\
&= \sum_{i\in\mathcal{U}} Var(e_i)
\end{aligned}$$

From the law of total variance $Var(e_i) = E(Var(e_i|\delta_i)) + Var(E(e_i|\delta_i))$. This gives

$$\begin{aligned} Var(E(e_i|\delta_i)) &= \pi_i^B(1-\pi_i^B)\Big(\frac{y_i-\mu_y}{\pi_i^B N} - \boldsymbol{\beta}_O^T\boldsymbol{v}_i - \boldsymbol{\beta}_O^T\boldsymbol{c}_i\Big)^2 \\ &= \pi_i^B(1-\pi_i^B)\Big(\frac{y_i-\mu_y}{\pi_i^B N} - \boldsymbol{\beta}_O^T\boldsymbol{k}_i\Big)^2 \end{aligned}$$

and

$$\begin{aligned} E(Var(e_i|\delta_i)) &= \frac{1-\pi_i^A}{\pi_i^A}(1-\pi_i^B)(\boldsymbol{\beta}_O^T\boldsymbol{c}_i)^2 \\ &= \frac{1-\pi_i^A}{\pi_i^A}(1-\pi_i^B)(\pi_i^B)^2(\boldsymbol{\beta}_O^T\boldsymbol{k}_i)^2 \end{aligned}$$

So

$$Var(\hat{\mu}_{OE}) \approx \sum_{i\in\mathcal{U}}\Big[\pi_i^B(1-\pi_i^B)\Big(\frac{y_i-\mu_y}{\pi_i^B N} - \boldsymbol{\beta}_O^T\boldsymbol{k}_i\Big)^2 + \frac{1-\pi_i^A}{\pi_i^A}(1-\pi_i^B)(\pi_i^B)^2(\boldsymbol{\beta}_O^T\boldsymbol{k}_i)^2\Big]$$

□

Comparing (3.31) with (3.30), it can be seen that the two variance expressions have a similar form, with two differences. The first is that ALP has a weighted least squares "fitted value" for $(y_i - \mu_y)/(\pi_i^B N)$ of $\beta_A^T\boldsymbol{t}_i$, compared with $\beta_O^T\boldsymbol{k}_i$ for OE. The second difference is that the second term of $Var(\hat{\mu}_{OE})$ includes an additional $(1-\pi_i^B)$ factor. When the auxiliary variables $\boldsymbol{x}$ explain a large part of the variation in $(y_i - \mu_y)/(\pi_i^B N)$, and A is small relative to B, the second variance term will be dominant. That term will tend to be smaller for the OE approach when the π_i^B values are larger, or in other words, when B consists of a large proportion of the population.

Remark 3.4. There is equivalence between $Var(\hat{\mu}_{CLW})$, $Var(\hat{\mu}_{KW})$, $Var(\hat{\mu}_{SL})$ and $Var(\hat{\mu}_{OE})$ when $\pi_i^A = 1$, that is, when the reference sample A is in fact the full population and $\boldsymbol{x}_i$ is available for all $i \in \mathcal{U}$. In this case, knowledge about $\delta_i, i \in \mathcal{U}$ does not give any advantages for the SL, OE and Kim and Wang (2019) approaches over the Chen et al. (2020) approach.

3.4 OE propensity score estimation when data integration is not possible

The proposed OE pseudo log-likelihood (3.6) requires knowledge of δ_i for each unit in A, and this information may not be available in some situations. To address unknown overlap when implementing their SL approach, Beaumont et al. (2024a) proposed to estimate the term involving the overlapping sample in their likelihood function (3.8) using a best linear unbiased prediction approach. In this section I describe an alternative approach to the problem by treating the unknown δ_i as a missing-data problem along the lines of Kim and Tam (2021). A two-step algorithm with similarities to the EM algorithm of Dempster et al. (1977) is employed to simultaneously estimate the missing δ_i values and obtain the maximum likelihood estimator for (3.6). I first re-write (3.6) as

$$l_{OE}(\boldsymbol{\theta}) = \sum_{i \in B} \log\pi(\boldsymbol{x}_i, \boldsymbol{\theta}) + \sum_{i \in A} d_i^A \log(1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})) - \sum_{i \in A} d_i^A \delta_i \log(1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})). \tag{3.32}$$

Under a logistic regression model for π_i^B, (3.32) becomes

$$\begin{aligned} l_{OE}(\boldsymbol{\theta}) = &\sum_{i \in B} \left\{ \boldsymbol{x}_i^T \theta - \log(1 + \exp(\boldsymbol{x}_i^T \theta)) \right\} - \sum_{i \in A} d_i^A \log(1 + \exp(\boldsymbol{x}_i^T \theta)) + \\ &\sum_{i \in A} d_i^A \delta_i \log(1 + \exp(\boldsymbol{x}_i^T \theta)), \end{aligned} \tag{3.33}$$

and the score function $S(\boldsymbol{\theta})$ is given as

$$\begin{aligned} S(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} l_{OE}(\boldsymbol{\theta}) \\ &= \sum_{i \in B} \boldsymbol{x}_i - \sum_{i \in B} \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i - \sum_{i \in A} d_i^A \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i + \sum_{i \in A} d_i^A \delta_i \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i. \end{aligned}$$

The EM-like procedure proceeds as follows:

Step 1: Calculate initial values of the parameters, $\boldsymbol{\theta}^{(1)}$, based on the information available. These can be obtained using a method like Chen et al. (2020) which does not require

knowledge of δ_i. The values $\boldsymbol{\theta}^{(1)}$ are used to obtain $\pi_i^{B,(1)}$.

Step 2: For the Expectation step, we calculate the expected value of $l_{OE}(\boldsymbol{\theta})$ with respect to $\delta_i \sim P(\cdot|\boldsymbol{x}, \boldsymbol{\theta}^{(j)})$, holding the sample A as fixed. The expectation is given by

$$E_{\delta_i \sim P(\cdot|\boldsymbol{x},\boldsymbol{\theta}^{(j)})}[l_{OE}(\boldsymbol{\theta})] = \sum_{i \in B} \left\{ \boldsymbol{x}_i^T \boldsymbol{\theta} - \log(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\theta})) \right\} - \sum_{i \in A} d_i^A \log(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\theta})) + \sum_{i \in A} d_i^A E(\delta_i | \boldsymbol{x}, \boldsymbol{\theta}^{(j)}) \log(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\theta})). \tag{3.34}$$

Note that $E(\delta_i|\boldsymbol{x}, \boldsymbol{\theta}^{(j)})$ can be interpreted as the conditional probability of unit i being included in the non-probability sample B, given the auxiliary variables $\boldsymbol{x}$ and $\boldsymbol{\theta}^{(j)}$. Under a logistic model this expectation can be estimated for each unit using

$$E(\delta_i|\boldsymbol{x}, \boldsymbol{\theta}^{(j)}) = \hat{\delta}_i^{(j)} = \pi_i^{B,(j)} = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\theta}^{(j)})}{(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\theta}^{(j)}))}$$

where $\pi_i^{B,(j)}$ is the value of π_i^B for the j'th iteration.

Step 3: In the Maximisation step, the log-likelihood (3.34) is maximised with respect to $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{(j+1)}$. The maximisation can generally be accomplished using an iterative procedure such as the Newton-Raphson method described in Section 3.2.

Step 4: Set $j = j + 1$, and go back to Step 2. Repeat until convergence - that is, $\max(|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(j-1)}|) < \epsilon$, where $\max(\boldsymbol{x})$ returns the component from vector $\boldsymbol{x}$ that is the largest, and ϵ is a user-specified threshold.

The final set of parameters $\hat{\boldsymbol{\theta}}$ are used to obtain $\hat{\pi}_i^B$ and subsequently may be fed into (3.1) to produce the IPW estimate of population mean.

Remark 3.5. In the special case where A is a Census, the OE, KW and SL approaches reduce to the full likelihood estimate like CLW. From empirical testing, the proposed EM-like procedure does not appear to converge exactly to the full likelihood estimate. However, it was found to produce results very close to the full likelihood estimate, with parameter estimates for $\hat{\boldsymbol{\theta}}$ generally equal to those from the CLW approach up to the second decimal place, and with relative bias and relative root mean square error results

essentially equivalent to those obtained from the CLW approach.

Variance estimation for the IPW estimator which uses $\hat{\pi}_i^B$ from the EM-like approach may be undertaken using the bootstrap method. For larger sample sizes the combination of the EM-like iterative procedure along with a bootstrap procedure was found to be slightly more computationally intensive compared with the OE approach, but not unduly so.

Algorithm 1 Bootstrap variance estimation for EM-like approach

Require: A, B, n_A, N_B, b_{rep}

1: Set $r := 1$
2: **while** $r \leq b_{rep}$ **do**
3: Select a Simple Random Sample With Replacement A_r from A of size n_A
4: Select a Simple Random Sample With Replacement B_r from B with size N_B
5: Calculate $\hat{\boldsymbol{\theta}}^r$ using A_r and B_r based on the EM-like approach, where the weight used for unit i in A_r is its original sample weight, d_i^A.
6: Calculate $\hat{\pi}_i^{B_r}$ for each unit $i \in B_r$ using $\hat{\boldsymbol{\theta}}^r$
7: Calculate $\hat{Y}^r$ based on $\hat{\pi}_i^{B_r}$ and B_r
8: $r := r + 1$
9: **end while**
10: Return the variance of the bootstrap estimates $\hat{Y}^r$, where $r = \{1, .., b_{rep}\}$

Algorithm 1 presents a relatively simple-to-implement procedure for the bootstrap variance estimation in the case when A is a single-stage Probability Proportional to Size (PPS) sample with small sampling fractions (see for example Chen et al. 2020), where b_{rep} is the number of bootstrap replications used. The algorithm would need to be altered to account for the case when A is a stratified and/or multistage design. Note also that the use of with-replacement sampling for B in the algorithm may not be so appropriate when B is large and has a non-negligible sampling fraction. In this case, assuming Poisson sampling for B and employing a bootstrap procedure which reflects that would be better. Options include the approach described in Berg et al. (2025) and originally proposed in Antal and Tillé (2011), or the pseudo-population bootstrap approach proposed in Algorithm 2 of Chapter 5.

3.5 Simulation Study

A simulation study was undertaken to examine the consistency of the proposed OE approach, and to compare its efficiency with alternative methods. The setup of the simulation is similar to what was used in Chen et al. (2020). Consider a finite population of size $N = 200,000$ that includes a data item of interest y and auxiliary variables $\boldsymbol{x}$. The data items y and $\boldsymbol{x}$ are related by the model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma \epsilon_i$$

where $\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, x_{3i}, x_{4i})$, $\boldsymbol{\beta} = (2, 1, 1, 1, 1)$, $x_{1i} = z_{1i}, x_{2i} = z_{2i} + 0.3x_{1i}, x_{3i} = z_{3i} + 0.2(x_{1i} + x_{2i}), x_{4i} = z_{4i} + 0.1(x_{1i} + x_{2i} + x_{3i})$, and $z_{1i} \sim$ Bernoulli$(0.5), z_{2i} \sim$ Uniform$(0, 2), z_{3i} \sim$ Exponential(1), and $z_{4i} \sim \chi^2(4)$. The error terms ϵ_i are independent and identically distributed as Normal$(0, 1)$. In the simulations, I utilise two values of σ, chosen such that the correlation coefficient ρ between y and $\boldsymbol{x}^T\boldsymbol{\beta}$ is either 0.3 (moderately weak) or 0.7 (moderately strong).

The true propensity scores π_i^B are given by

$$\log\Big(\frac{\pi_i^B}{1 - \pi_i^B}\Big) = \theta_0 + 0.1x_{1i} + 0.2x_{2i} + 0.3x_{3i} + 0.4x_{4i}$$

where the intercept θ_0 was chosen to ensure the target sample size for N_B is achieved. Poisson sampling was used to select the sample B, while randomised systematic probability proportional to size sampling (PPS) was used to select the reference sample A. In general, the inclusion probabilities for π_i^A were proportional to $z_i = c + x_{3i}$, where c was chosen to ensure the ratio $\max(z_i)/\min(z_i)$ was equal to 10.

The simulation exercise aimed to examine the performance of the OE IPW estimators relative to a number of other approaches, and to assess the effectiveness of the proposed variance estimators. For the comparison of IPW estimators and the plug-in variance estimator, I used four different sample sizes for N_B of 750, 2,000, 50,000 and 140,000. To assess the bootstrap estimator for the EM-like algorithm, I used the following sample sizes

for N_B - 750, 2,000, 50,000 and 90,000. The use of 90,000 instead of 140,000 for the largest sample size was due to runtime considerations under repeated simulation iterations. The sample size for A was kept fixed at $n_A = 5,000$ for all simulation runs.

$R = 2,000$ simulations were undertaken for each combination of ρ and sample size. In early empirical tests it was found that this number of simulations was more than sufficient to produce stable results. For the EM algorithm approach, 200 bootstrap iterations were used to calculate the bootstrap variances. The performance of the different estimators was evaluated by calculating the Monte Carlo Relative Bias (RB) and Relative Root Mean Squared Error (RRMSE):

$$\text{RB} = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\mu}_{y,r} - \mu_y}{\mu_y}, \ \text{RRMSE} = \frac{\sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\mu}_{y,r} - \bar{\hat{\mu}}_y)^2 + (\bar{\hat{\mu}}_y - \mu_y)^2}}{\mu_y}$$

where $\hat{\mu}_{y,r}$ is the estimate of the population mean computed for the r'th simulation run, $\bar{\hat{\mu}}_y = 1/R\sum_{r=1}^{R}\hat{\mu}_{y,r}$ and μ_y is the true population mean.

Results were produced for the following estimators:

- Naive - Simple sample mean from B
- OE - The OE IPW estimator
- OE-EM - The IPW estimator produced by applying the EM algorithm to estimate $\hat{\delta}_i, i \in A$ in the OE pseudo-likelihood
- SL - The sample likelihood approach described in Beaumont et al. (2024a) and Kim and Kwon (2024)
- CLW - The IPW2 estimator proposed by Chen et al. (2020)
- ALP - The ALP estimator proposed by Wang et al. (2021a)
- KW - The estimator proposed by Kim and Wang (2019)

A logistic model was used for π_i^B when implementing most of the methods. The exception was the ALP method which, as described in Wang et al. (2021a), assumes a logistic model

for $p_i = \pi(\boldsymbol{x}_i, \boldsymbol{\theta})/(1 + \pi(\boldsymbol{x}_i, \boldsymbol{\theta}))$. This essentially means that the π_i^B in the ALP method are implicitly modelled as $\pi(\boldsymbol{x}_i, \boldsymbol{\theta}) = \exp(\boldsymbol{x}_i^T \boldsymbol{\theta})$.

3.5.1 Simulation Results

Tables 3.1 and 3.2 compare RB and RRMSE values for each of the estimators under the different sample sizes N_B and correlations between y and $\boldsymbol{x}^T\beta$. The results for $\rho = 0.3$ and $\rho = 0.7$ are similar, and so the following discussion applies to both tables unless stated otherwise.

Table 3.1: Comparative performance of estimators - $\rho(y, \boldsymbol{x}^T\beta) = 0.3$

	$N_B = 750$		$N_B = 2,000$		$N_B = 50,000$		$N_B = 140,000$	
	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE
Naive	85.09	85.20	73.52	73.57	27.55	27.56	8.11	8.11
OE	-0.13	9.84	-0.06	5.80	-0.01	0.77	-0.02	0.32
OE-EM	0.00	9.77	0.03	5.81	0.01	0.82	0.00	0.42
SL	0.09	9.63	0.11	5.63	0.01	0.71	-0.01	0.31
CLW	-0.33	10.21	-0.17	6.16	0.25	1.15	-0.06	0.74
ALP	0.89	9.47	1.10	5.59	0.47	0.88	0.04	0.53
KW	0.27	12.56	0.01	6.88	-0.04	1.13	-0.04	0.42

The Naive estimate based solely on the sample mean of B is very biased for all N_B sizes, as expected given the selection bias present in B. The proposed OE and OE-EM estimators, along with the CLW, KW and SL estimators, are all approximately unbiased. There is a hint of bias in the ALP estimator for the smaller N_B sample sizes, but the bias looks to dissipate as N_B gets larger.

Table 3.2: Comparative performance of estimators - $\rho(y, \boldsymbol{x}^T\beta) = 0.7$

	$N_B = 750$		$N_B = 2,000$		$N_B = 50,000$		$N_B = 140,000$	
	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE
Naive	85.06	85.09	73.70	73.71	27.62	27.62	8.13	8.13
OE	-0.20	4.44	-0.16	2.65	-0.02	0.52	-0.01	0.27
OE-EM	-0.06	4.35	-0.06	2.64	0.00	0.58	0.00	0.38
SL	0.04	3.97	0.02	2.26	0.00	0.40	-0.01	0.26
CLW	-0.41	5.13	-0.26	3.33	0.24	0.97	-0.06	0.72
ALP	0.84	3.97	1.01	2.43	0.45	0.69	0.04	0.50
KW	0.15	7.85	-0.10	4.46	-0.05	0.97	-0.03	0.38

In terms of the RRMSE performance of the estimators, when $\rho = 0.3$ the ALP estimator

provided the most efficient estimates for the smallest N_B sizes of 750 and 2,000. When N_B was 50,000 or 140,000 the SL estimator provided the lowest RRMSE results. The proposed OE and OE-EM estimators tend not to perform as well as the SL and ALP estimators for the smaller N_B sizes, but provide the second-lowest RRMSE results (behind only the SL approach) when N_B was larger. The RRMSE results for $\rho = 0.7$ are similar, however it is noted that for this higher correlation value the SL estimator provides the lowest RRMSE values even for the lowest N_B values.

The KW estimator did not perform as well compared with most of the other estimators when N_B was small. Very small values of N_B, coupled with small to moderate reference sample sizes n_A, may result in a very small number of units in the domain $A \cap B$, and in the extreme case the domain may be empty. Under larger values of N_B, the performance of the KW estimator improved.

The results from Tables 3.1 and 3.2 also show that estimating the values of δ_i in A using the EM algorithm produces consistent estimates of the population mean. The RRMSE tends to be slightly higher compared to when we have δ_i available for larger N_B samples. However, the OE-EM RRMSE values were similar to or lower than those for the OE approach for the smaller N_B values.

I next examined the coverage probabilities (CP) and average length (AL) of the 95% confidence intervals for the OE plug-in variance estimator, and the OE-EM bootstrap variance approach. In this comparison, the design-based variance term (3.16) was estimated by assuming A was drawn using with-replacement sampling. This is a reasonable approximation to make for a without-replacement PPS sample when N is large and the sampling fraction is small (Wu and Thompson 2020, pp. 80-81).

Table 3.3 provides the CP and AL values for the 95% confidence intervals using the plug-in variance estimator for the OE IPW estimator. Results are shown for both $\rho = 0.3$ and $\rho = 0.7$. The coverage probabilities are all close to the nominal value of 0.95. Given the population mean of y is about 9.3, the average lengths of the confidence intervals are also not excessively large. For example, when $\rho = 0.3$ and $N_B = 750$, the 95% confidence interval for $\hat{\mu}_y$ is $\hat{\mu}_y \pm 1.85$ and the Relative Standard Error (RSE) of the estimate is about

10%. When $\rho = 0.7$ for the same value of N_B, the RSE drops to about 4.6%.

Table 3.3: Coverage Probabilities (CP) and Average Length (AL) of 95% Confidence Intervals for OE plug-in variance estimator

	$\rho(Y, X\beta) = 0.3$		$\rho(Y, X\beta) = 0.7$	
N_B	CP	AL	CP	AL
750	0.945	3.71	0.951	1.66
2,000	0.940	2.12	0.956	0.99
50,000	0.955	0.29	0.952	0.19
140,000	0.951	0.11	0.946	0.09

Table 3.4 provides the CPs and ALs for the 95% confidence intervals when the bootstrap approach is used to estimate variances for the OE-EM approach. Due to the computational time involved in running 2,000 simulations, results were only produced for $\rho = 0.3$. The CPs for the bootstrap approach are again close to the nominal value of 0.95, although it appears that there is a slight over-coverage when N_B is larger. A possible cause for this slight over-coverage for the larger N_B is that the bootstrap procedure uses with-replacement sampling for B. The AL again appears to be reasonable, and reduces as sample size increases.

Table 3.4: Coverage Probabilities (CP) and Average Length (AL) of 95% Confidence Intervals - Bootstrap variance for OE-EM approach

	$\rho(Y, X\beta) = 0.3$	
N_B	CP	AL
750	0.947	3.59
2,000	0.941	2.09
50,000	0.969	0.32
90,000	0.970	0.23

3.6 Real Data Application

In this section the proposed OE and OE-EM methods are compared alongside the SL, CLW and ALP approaches using real data, where the reference sample may also be drawn using sampling approaches other than PPS sampling.

The data used for this exercise is based on Australian Business Activity Statement (BAS) data from the July-September quarter of the 2022-2023 financial year. The BAS data was

sourced from the Business Longitudinal Analysis Data Environment (BLADE) standard module and accessed using the Australian Bureau of Statistics (ABS) DataLab. Records designated as live businesses, in the private or public sector, and with an active Goods and Services Tax (GST) payer role, were extracted. Pay-As-You-Go (PAYG) data relating to the number of employees in each business was also obtained for the 2022-2023 financial year. The complete records with no missing data item values - 721,769 records in total - were designated to be the population dataset.

Table 3.5: Population distribution across state

	NSW	VIC	QLD	SA	WA	TAS	NT	ACT
Business Count	249,089	187,065	142,994	43,229	68,727	13,830	5,529	11,306

Note: Derived population counts created using BLADE indicative items and BAS data from Quarter 1 2022-2023, as well as PAYG data from 2022-2023 (Business Longitudinal Analysis Data Environment (BLADE) 2022) and BAS data from Quarter 1 2022-2023, accessed through the ABS DataLab December 2024.

The data item of interest for this exercise was total sales excluding GST (turnover), while available auxiliary items included total salary, wages and other payments (wages), geographic location of the business (state), and a derived industry division classification. Population distributions across state and industry, as well as population quantiles for turnover and wages, are provided for the interested reader in Tables 3.5, 3.6 and 3.7. In particular, it is noted that the turnover and wages data items are very right skewed, a typical scenario for business surveys (see for example Rivière 2002). The total turnover for the full population during the July-September quarter was about $611,102.1 million, which equated to a mean turnover per business of $846,673.

Table 3.6: Selected quantiles for turnover and wages

	Quantile				
	1%	25%	50%	75%	99%
Turnover	$0	$59,408	$147,527	$403,741	$10,101,149
Wages	$0	$15,150	$39,108	$102,078	$2,074,381

Note: Derived population counts created using BLADE indicative items and BAS data from Quarter 1 2022-2023, as well as PAYG data from 2022-2023 (Business Longitudinal Analysis Data Environment (BLADE) 2022) and BAS data from Quarter 1 2022-2023, accessed through the ABS DataLab December 2024.

The selected estimation methods were assessed by conducting a simulation study using the data from this population. In each iteration of the simulation, a non-probability dataset

Table 3.7: Population distribution across industry division

Industry Division	Business Count
Agriculture, Forestry and Fishing	35,081
Mining	2,504
Manufacturing	37,218
Electricity, Gas, Water and Waste Services	2,716
Construction	128,297
Wholesale Trade	30,133
Retail Trade	59,945
Accommodation and Food Services	57,862
Transport, Postal and Warehousing	27,727
Information Media and Telecommunications	6,324
Financial and Insurance Services	24,108
Rental, Hiring and Real Estate Services	27,096
Professional, Scientific and Technical Services	102,844
Administrative and Support Services	32,477
Public Administration and Safety	3,141
Education and Training	16,322
Health Care and Social Assistance	58,132
Arts and Recreation Services	12,275
Other Services	57,567

Note: Derived population counts created using BLADE indicative items and BAS data from Quarter 1 2022-2023, as well as PAYG data from 2022-2023 (Business Longitudinal Analysis Data Environment (BLADE) 2022) and BAS data from Quarter 1 2022-2023, accessed through the ABS DataLab December 2024.

was selected using Poisson sampling. The selection mechanism into the non-probability dataset was driven by a logistic model with one continuous covariate (a standardised wage variable) and one categorical variable (industry division):

$$\log\Big(\frac{\pi_i^B}{1-\pi_i^B}\Big) = \theta_0 + 0.05x_{\text{stdwages}} + \boldsymbol{\theta}_2^T \boldsymbol{x}_{\text{industry}}$$

where $\boldsymbol{\theta}_2^T = (0, 0, 0, 0, 0, -0.4, -0.1, 0, 0, 0, 0, -0.2, 0, 0, 0, 0, 0, -0.2)$ and $\boldsymbol{x}_{\text{industry}}$ is the industry division variable transformed into 18 0-1 indicator variables (one for each division except for Division A). These values were chosen to reflect some industries being less likely to be included in the non-probability dataset. The value θ_0 was chosen so that the expected size of the non-probability dataset was equal to $100,000$.

Of interest was to examine the behaviour of the different approaches when the probability sample was chosen using several different sampling schemes. These included

1. *PPS sampling* (PPS): Wages was used as the size variable, where any business with wages greater than \$500,000 was given a wages value of \$500,000. The inclusion probabilities for π_i^A were proportional to $z_i = c + \text{wages}_i$, where c was chosen to ensure the ratio $\max(z_i)/\min(z_i)$ was equal to 20. This step was taken to control the variability of the weights, and hence increase the stability of estimates. The probability sample was selected using randomised systematic PPS sampling.

2. *PPS sampling with probabilities dependent on membership in B* (PPS Cond): The same setup as for sampling scheme 1, except that i) the initial selection probabilities for units already in the non-probability dataset were first multiplied by 1.3, and then ii) all probabilities were adjusted so that the sum of the probabilities equalled 1. These probabilities were then multiplied by 100,000 to obtain the final selection probability.

3. *Simple Random Sample* (SRSWOR): A Simple Random Sample Without Replacement (SRSWOR) was drawn from the full population.

4. *Stratified SRSWOR* (Strat. SRSWOR): The population was assigned to strata based on state, industry division and a size category based on number of employees in a business. The size categories were: 0 to 4 employees, 5 to 19 employees, 20 to 299 employees, and 300+ employees. An optimal allocation to strata was undertaken using the Bethel-Chromy algorithm (Bethel 1989; Chromy 1987) to ensure accuracy constraints at the national level (3%) and industry division level (5%) were met. A minimum sample size of 6 in each stratum was applied, while all businesses in the 300+ size strata were completely enumerated.

For the first three sample design scenarios, the sample size for the probability sample was 15,000, while for the Stratified SRSWOR scenario the total sample size was 14,929. For each sample design 2,000 simulations were undertaken. In each iteration of the simulation, the OE, OE-EM, SL, ALP and CLW methods were applied to the drawn samples, and the mean turnover was calculated.

Table 3.8 provides the Monte Carlo RB and RRMSE values for each estimator under the various sample design scenarios for A. The OE and SL approaches are generally the best performing approaches in all sample design scenarios. Under the PPS and Stratified SRSWOR design scenarios, the OE and SL RRMSEs are about equal. However, under the SRSWOR and PPS with conditional probabilities scenarios, the OE approach provides a lower RRMSE. When the probabilities of selection into A depend on whether or not the unit is in B, the SL estimate becomes slightly biased, while the OE approach remains unbiased.

Table 3.8: Real-data performance of estimators for turnover

	PPS		PPS Cond		SRSWOR		Strat. SRSWOR	
Method	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE
OE	-0.39	8.47	-0.57	8.53	-1.35	8.95	0.25	8.60
OE-EM	0.46	11.36	0.57	11.15	-2.14	12.85	0.21	8.56
SL	-0.24	8.41	2.06	9.38	-0.33	9.19	0.23	8.58
CLW	-1.30	13.87	-0.9	13.34	-7.4	16.23	0.23	8.60
ALP	3.06	11.36	3.22	11.38	1.41	12.32	3.96	10.97

The OE-EM approach tends to yield higher RRMSEs compared to the OE and SL approaches, but still tends to outperform the CLW and ALP methods in all four sample design scenarios. Compared to the other estimators, the ALP approach tends to yield biased results in the simulation under all sample design scenarios. This bias appears to be due to the fact that the ALP method models a probability $\hat{p}_i = \hat{\pi}_i^B/(1+\hat{\pi}_i^B)$ (see equation (3.9)), and it turns out that for this population and simulation scenarios many of the estimates of $\hat{p}_i$ tend to work out to be greater than 0.5. This means that many of the resulting $\hat{\pi}_i^B = \hat{p}_i/(1-\hat{p}_i)$ values, where $\hat{p}_i$ is modelled using logistic regression, end up being greater than one. It is possible in the ALP approach to assume a different model for the π_i^B, such as the logistic model, to avoid this issue; see for example Gershunskaya and Beresovsky (2024). However, this alternative formulation was not explored in this thesis.

There is some residual bias in the estimators under the SRSWOR scenario, but this bias is small relative to the RRMSE. The exception is the CLW approach which yielded a RB of -7.4%. The underlying cause for the residual bias appears to be the highly skewed nature of the standardised wages item. When the (relatively rare) units with the larger wage values are not selected in the simple random sample, the selection bias model coefficient

for standardised wages tends to be underestimated, leading to a downward bias in the resulting population estimates. The OE and SL approaches may be less affected by this due to the fact that they only use the data from the reference sample which does not overlap with the non-probability sample. To a lesser extent, the OE-EM approach also aims to do this by estimating and then removing the contribution from the overlapping sample. This result demonstrates the importance of ensuring the reference sample is of sufficient quality so as to effectively assist in correcting selection bias.

3.7 Concluding Remarks

In this chapter knowledge about which reference sample units are also in the non-probability sample was used to help estimate the participation probabilities of the non-probability units. The proposed OE pseudo-likelihood excludes from the likelihood expression the probability sample units which are also in the non-probability sample.

The performance of the OE pseudo-likelihood approach was compared against several alternative methods in terms of their efficiency when making statistical inferences based on the non-probability sample. This comparison was undertaken both theoretically, and empirically within both a simulated data setting and a real-data scenario. In the first simulation exercise the OE approach was found to out-perform the other pseudo-likelihood based methods when the non-probability datasets were a larger fraction of the population. Theoretical comparisons corroborated this empirical finding. In particular, under the assumption of Poisson sampling for the reference sample, the OE approach was found to always have an asymptotic variance no larger than the KW and CLW methods.

The SL approach was the most efficient approach in the first simulation exercise. However, it was found that in the real-data simulation the efficiency of the SL approach tended to be comparable or slightly worse than the OE approach. In particular, when the probability of inclusion into the reference sample depends on its presence in the non-probability dataset, the SL estimate tended to have a slight bias and larger RRMSE, while the OE estimate remained unbiased. These results suggest that in practical situations the OE approach

may be preferred as the safer, more robust option compared to the SL approach which relies on tighter assumptions.

When it is not possible to identify which units in the reference sample are also on the non-probability dataset, the EM algorithm can be used in conjunction with the OE pseudo-likelihood to estimate the non-probability membership status of reference sample units and the model parameters for the propensity scores simultaneously. Estimating the δ_i links tends to require a cost in terms of a slightly higher RRMSE, however the empirical studies showed that it can still be preferred compared to other alternatives.

This chapter has assumed that a SAR missingness mechanism exists in B, so that participation probabilities only depend on the auxiliary data items. The auxiliary information is also assumed to be captured without measurement error. In reality, these assumptions may not hold. Tests can be utilised (see for example Andridge et al. 2019) to assess whether a SNAR situation exists in the non-probability data. The reference sample units that are in common with the non-probability sample can be used to examine whether there is measurement error in the auxiliary information. These issues will be explored further in Chapter 5.

Chapter 4

Optimal sample design for the reference sample

4.1 Introduction

In current applied practice, the usual purpose of a probability survey is to collect information about specific topic(s) of interest to the survey statistician; see for example Chapter 1 of Cochran (1977). It is generally designed to provide efficient estimates for key variables of interest, and available auxiliary information may be used to provide efficiency gains for the probability survey estimates. In the context of this thesis, the situation is reversed - the variables that are of interest are now collected via a non-probability sample, and these are different to the variables available on the reference probability survey. In particular, the probability survey is not designed to provide efficient estimates for the variables that are of interest.

There may be opportunity to efficiently design the reference sample A to specifically achieve an optimal result (lowest cost for a given variance, or vice versa) for the estimate $\hat{\mu}_{IPW}$ produced using B. Some scenarios when this might be attractive include:

- An existing survey A collects certain information about a population which a new,

large but incomplete administrative dataset also includes. To reduce respondent burden, it is proposed to stop collecting y in A and instead rely on B. The sample A can then be re-purposed to provide the characteristic information $\boldsymbol{x}$ to be used in addressing selection bias.

- There is interest in taking advantage of a newly available data source B but there is not sufficient auxiliary information on $\boldsymbol{x}$ to help with correcting selection bias. In this case, one may choose to develop a probability survey A to assist in collection of the required $\boldsymbol{x}$ variables.

A reasonable question that may be asked is: "If we are designing the reference survey to enable statistics to be produced about y, why not just collect y in the survey?"; Indeed if only a small number of extra data items need to be collected then the additional cost of doing so could be minimal. However, if any additional data items are so substantial that significant increases in cost and respondent burden are introduced, then there is a savings trade-off to be made. Some variables may be impractical to collect in A, for example because of high cost of collecting the data items and/or response burden. Another consideration would be that A can only be run much less frequently than data from the non-probability sample B is updated.

The savings increase further if A can be used to help address selection bias for more than one non-probability dataset. This leads to the potential for a National Statistical Office (NSO) to reduce the number, and frequency, of surveys they run. Before, the NSO may have conducted a suite of surveys, each of which collected data on a particular set of topics. Now, the NSO would rely more heavily on a suite of non-probability data, supplemented by auxiliary information from a "Population Profile Survey" A to help address selection bias issues.

This chapter extends the previous chapter on OE participation probabilities. In particular, it explores the problem of optimally designing a reference sample which will be combined with available non-probability data to form OE IPW estimates. The optimal selection probabilities are derived under several different sampling schemes, and a procedure for how these selection probabilities may be estimated in practice is described. Under the

scenario where it is possible to link records from a non-probability sample with the population frame, a split-population approach is presented whereby a reference sample is taken from the complement $C = \mathcal{U} \setminus B$ and participation probabilities are then estimated by combining that sample with the non-probability sample. The results from an empirical study demonstrate the efficiency gains that are possible.

The rest of the chapter is outlined as follows. In Section 4.2 optimal selection probabilities for A under Poisson, PPS, and stratified SRSWOR sampling schemes are derived, assuming that the sample will be combined with B using the OE IPW estimator defined in Section 3.2. Section 4.3 presents the split-population approach for forming participation probabilities. Section 4.4 examines the potential efficiency gains from these reference sample design options through a simulation study. Finally, Section 4.5 provides some concluding remarks.

4.2 Optimal selection probabilities for OE IPW estimator

Recall that Section 3.2 introduced the OE IPW estimator $\hat{\mu}_{OE}$

$$\hat{\mu}_{OE} = \frac{1}{\hat{N}_B} \sum_{i \in B} \frac{y_i}{\hat{\pi}_i^B} \tag{4.1}$$

where $\hat{N}_B = \sum_{i \in B} (\hat{\pi}_i^B)^{-1}$, the $\hat{\pi}_i^B$ are estimated by a logistic regression model, and the parameters of that model are estimated by solving the OE score equation

$$S(\boldsymbol{\theta}) = \sum_{i \in B} \boldsymbol{x}_i - \sum_{i \in B} \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i - \sum_{i \in A \setminus B} d_i^A \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i = 0. \tag{4.2}$$

The variance of the OE IPW estimator generally depends on several factors. This includes the model that is assumed for π_i^B (for example a logistic model), and the sampling scheme used for A (for example, Poisson sampling). In this chapter, I describe optimal sample design as the process of finding selection probabilities that will minimise the variance of

the estimator given the expected cost of taking the sample, given as

$$C_t = \sum_{i=1}^{N} \pi_i^A c_i \tag{4.3}$$

where c_i is the cost of enumerating unit i and π_i^A is the probability of selection of unit i into A. An alternative goal might be to minimise the overall cost C_t for a desired variance outcome for $\hat{\mu}$.

Recall from Theorem 3.1 in Section 3.3 that $Var(\hat{\mu}_{OE})$ includes the terms (3.12), (3.13) and (3.14), of which the last two depend on the sample design of A. These last two terms are the relevant terms when deriving the optimal selection probabilities, since the term (3.12) does not involve π_i^A. For (3.13), $\boldsymbol{b}$ can be brought inside the summation to obtain

$$\begin{aligned}
\boldsymbol{b}^T \boldsymbol{V} \boldsymbol{b} &= \frac{1}{N^2} Var_d\Big(\sum_{i \in A} d_i^A \pi_i^B (1 - \pi_i^B) \boldsymbol{b}^T \boldsymbol{x}_i\Big) \\
&= \frac{1}{N^2} Var_d\Big(\sum_{i \in A} d_i^A v_i\Big) \\
&= \frac{1}{N^2} Var_d(\hat{v})
\end{aligned}$$

where $v_i = \pi_i^B(1 - \pi_i^B)\boldsymbol{b}^T \boldsymbol{x}_i$ and $\hat{v}$ is the Horvitz-Thompson estimator of the population total of $v = \sum_{i \in \mathcal{U}} \pi_i^B(1 - \pi_i^B)\boldsymbol{b}^T \boldsymbol{x}_i$. Now, the variance of a Horvitz-Thompson estimate $\hat{v}$ can be expressed as (see for example Section 4.2 of Wu and Thompson 2020)

$$Var_d(\hat{v}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij}^A - \pi_i^A \pi_j^A}{\pi_i^A \pi_j^A} v_i v_j.$$

So

$$\boldsymbol{b}^T \boldsymbol{V} \boldsymbol{b} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij}^A - \pi_i^A \pi_j^A}{\pi_i^A \pi_j^A} \pi_i^B \pi_j^B (1 - \pi_i^B)(1 - \pi_j^B) \boldsymbol{b}^T \boldsymbol{x}_i \boldsymbol{x}_j^T \boldsymbol{b}. \tag{4.4}$$

Re-writing $d_i^A = 1/\pi_i^A$, expression (3.14) (which is the third term of $Var(\hat{\mu}_{OE})$ in Theorem

3.1), can be re-written as

$$\frac{1}{N^2}\sum_{i=1}^{N}\Big[\frac{1}{\pi_i^A}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}-(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big] \tag{4.5}$$

where only the first term in the sum is of relevance for the optimal design since the other term does not involve π_i^A.

The following subsections describe the derivation of the optimal selection probabilities when the sample A is selected using three different sampling schemes: Poisson sampling, PPS sampling, and stratified SRSWOR sampling.

4.2.1 Poisson sampling

Under Poisson sampling, $\pi_{ij}^A = \pi_i^A\pi_j^A, i \neq j$, and $\pi_{ii}^A = \pi_i^A$, and so (4.4) becomes

$$\boldsymbol{b}^T\boldsymbol{V}\boldsymbol{b} = \frac{1}{N^2}\sum_{i=1}^{N}\Big(\frac{1}{\pi_i^A}-1\Big)(\pi_i^B)^2(1-\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}. \tag{4.6}$$

The relevant OE IPW variance terms to obtain optimal selection probabilities under Poisson sampling are thus the first term of (4.5) and the term (4.6). Denoting the sum of these two terms as M_{Poi} gives

$$\begin{aligned}
M_{Poi} &= \frac{1}{N^2}\left\{\sum_{i=1}^{N}\Big[\frac{1}{\pi_i^A}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big]+\sum_{i=1}^{N}\Big(\frac{1}{\pi_i^A}-1\Big)(\pi_i^B)^2(1-\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\right\} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{\pi_i^A}\Big[(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}+(\pi_i^B)^2(1-\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big]- \\
&\quad \frac{1}{N^2}\sum_{i=1}^{N}(\pi_i^B)^2(1-\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \qquad (4.7) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{\pi_i^A}\Big[(\pi_i^B)^2(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big]-\frac{1}{N^2}\sum_{i=1}^{N}(\pi_i^B)^2(1-\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}. \qquad (4.8)
\end{aligned}$$

The goal is to minimise (4.8) given the cost constraint (4.3). The Lagrange equation is

set up as

$$\mathcal{L} = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{\pi_i^A}\left[(\pi_i^B)^2(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\right] - \frac{1}{N^2}\sum_{i=1}^{N}(\pi_i^B)^2(1-\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} + \lambda\left\{\sum_{i=1}^{N}\pi_i^A c_i - C_t\right\}$$

Elementary application of Lagrange multipliers gives

$$\begin{aligned}\pi_i^A &= \sqrt{\frac{(\pi_i^B)^2(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}}{\lambda c_i N^2}} \\ &= \frac{\pi_i^B\sqrt{(1-\pi_i^B)}|\boldsymbol{b}^T\boldsymbol{x}_i|}{\sqrt{\lambda c_i}N}\end{aligned} \tag{4.9}$$

Substituting (4.9) into the cost constraint(4.3) yields

$$\frac{1}{\sqrt{\lambda}} = \frac{C_t N}{\sum_{i=1}^{N}\pi_i^B\sqrt{c_i(1-\pi_i^B)}|\boldsymbol{b}^T\boldsymbol{x}_i|} \tag{4.10}$$

The optimal selection probabilities can now be obtained by feeding (4.10) into (4.9), giving

$$\begin{aligned}\pi_i^A &= \frac{\pi_i^B\sqrt{(1-\pi_i^B)}|\boldsymbol{b}^T\boldsymbol{x}_i|}{\sqrt{c_i}N} \times \frac{C_t N}{\sum_{i=1}^{N}\pi_i^B\sqrt{c_i(1-\pi_i^B)}|\boldsymbol{b}^T\boldsymbol{x}_i|} \\ &= \frac{C_t\pi_i^B\sqrt{(1-\pi_i^B)}|\boldsymbol{b}^T\boldsymbol{x}_i|}{\sqrt{c_i}\sum_{i=1}^{N}\pi_i^B\sqrt{c_i(1-\pi_i^B)}|\boldsymbol{b}^T\boldsymbol{x}_i|} \\ \implies \pi_{i,opt}^A &\propto \frac{\pi_i^B\sqrt{(1-\pi_i^B)}|\boldsymbol{b}^T\boldsymbol{x}_i|}{\sqrt{c_i}}\end{aligned} \tag{4.11}$$

Remark 4.1. From 4.11, the optimal probabilities of selection can be seen to be greater when

- c_i is smaller. As the cost of sampling unit i increases, it is less desirable to include it in the sample.
- $|\boldsymbol{b}^T\boldsymbol{x}_i|$ is larger. This term can be thought of as the predicted/modelled value of $(y_i - \mu)/\pi_i^B$ when fitting a weighted linear regression using $\boldsymbol{x}$ as the explanatory

variables, $\boldsymbol{b}$ as the model parameters, and $\pi_i^B(1-\pi_i^B)$ as the model parameter weights. Units with values of $\boldsymbol{x}_i$ associated with more extreme values of y relative to μ have greater chance of selection.

- $\pi_i^B\sqrt{(1-\pi_i^B)}$ is larger. This will occur when π_i^B is not too close to 0 or 1.

Instead of using monetary values, the unit costs c_i may instead be specified to provide a rough measure of control over the level of overlap between A and B. For example, in order to reduce respondent burden it may be desirable to reduce the number of units that are selected in A while also being in B. If the unit cost is given by $c_i = (\pi_i^B)^2$, then units will be more likely to be selected into A if they have smaller values of π_i^B. Alternatively, one may wish to increase the level of overlap between A and B in particular segments of the population, for example if there is reason to doubt the quality of the data in B within those segments. In this case setting a lower c_i value will increase the number of overlapping units in those segments and these units can be used to produce a measurement error model to correct the quality issues in B.

When A is an existing probability sample that has been made available to be combined with B in estimation, then it is possible to make use of the data from A and B to design a more efficient sample A for future use in combining with B. On the other hand, A may not already exist, and so an optimal sample for A can be designed from the ground up. In this case, one would need to obtain the required design information from another source, such as a pilot study or historical census data.

4.2.2 Probability Proportional to Size Sampling with Fixed Sample Size

Poisson sampling is a form of probability proportional to size (PPS) sampling, however the resulting sample size is random. To find optimal selection probabilities for general fixed sample size PPS designs, I derive the optimal probabilities for the simpler with-replacement case. This is reasonable under the assumption of large N and a small sampling fraction n_A/N (see Section 4.5 of Wu and Thompson 2020), and allows the simplification of the variance expression for the optimisation.

From Matei and Tillé (2005), the variance of $\hat{Y} = \sum_{i=1}^{n} y_i/\pi_i$ in a with-replacement design is approximated by

$$
\begin{aligned}
Var_{WR}(\hat{Y}) &\approx \sum_{i=1}^{N} \frac{b_i}{\pi_i^2}(y_i - y_i^*)^2, \text{ where} \\
b_i &= \pi_i \frac{N}{N-1} \text{ and} \\
y_i^* &= \pi_i \frac{\sum_{l=1}^{N} b_l y_l/\pi_l}{\sum_{l=1}^{N} b_l} \\
&= \pi_i \frac{\sum_{l=1}^{N} \pi_l \frac{N}{N-1} y_l/\pi_l}{\sum_{l=1}^{N} \pi_l \frac{N}{N-1}} = \pi_i \frac{Y}{n}, \text{ where } Y = \sum_{l=1}^{N} y_l \text{ and } n = \sum_{l=1}^{N} \pi_l
\end{aligned}
$$

And so

$$
\begin{aligned}
Var_{WR}(\hat{Y}) &\approx \sum_{i=1}^{N} \frac{\pi_i}{\pi_i^2} \frac{N}{N-1} \Big(y_i - \pi_i \frac{Y}{n}\Big)^2 \\
&= \frac{N}{N-1} \sum_{i=1}^{N} \frac{1}{\pi_i} \Big(y_i^2 - 2y_i \pi_i \frac{Y}{n} + \pi_i^2 \frac{Y^2}{n^2}\Big) \\
&= \frac{N}{N-1} \Big\{ \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} - 2\frac{Y}{n} \sum_{i=1}^{N} y_i + \frac{Y^2}{n^2} \sum_{i=1}^{N} \pi_i \Big\} \\
&= \frac{N}{N-1} \Big\{ \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} - 2\frac{Y^2}{n} + \frac{Y^2}{n} \Big\} \\
&= \frac{N}{N-1} \Big\{ \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} - \frac{Y^2}{n} \Big\} \qquad (4.12)
\end{aligned}
$$

Using (4.12), substituting $v_i = \pi_i^B(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i$ and denoting v as the population total for the v_i, (3.13) becomes

$$
\begin{aligned}
\boldsymbol{b}^T \boldsymbol{V} \boldsymbol{b} &= \frac{1}{N^2} Var_d(\hat{v}) \\
&= \frac{1}{N^2} \frac{N}{N-1} \Big\{ \sum_{i=1}^{N} \frac{(\pi_i^B)^2 (1-\pi_i^B)^2 \boldsymbol{b}^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{b}}{\pi_i^A} - \frac{v^2}{n_A} \Big\} \qquad (4.13)
\end{aligned}
$$

Denoting M_{PPS} as the sum of (4.13) and the first component of (4.5), one obtains

$$M_{PPS} = \frac{1}{N^2}\Big\{\frac{N}{N-1}\Big[\sum_{i=1}^{N}\frac{(\pi_i^B)^2(1-\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}}{\pi_i^A} - \frac{v^2}{n_A}\Big] + \sum_{i=1}^{N}\frac{1}{\pi_i^A}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big\}$$
$$\approx \frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{\pi_i^A}\Big[(\pi_i^B)^2(1-\pi_i^B)^2\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} + (\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big] - \frac{v^2}{n_A N^2} \tag{4.14}$$

The goal is to minimise M_{PPS} subject to the cost constraint. Note that the last term of (4.14) does not include π_i^A, while the first term which does include π_i^A is equivalent to the first term in (4.7). As such, when the Lagrange equation is set up to solve for π_i^A, the resulting optimal selection probabilities $\pi_{i,opt}^A$ will be the same as those obtained for Poisson sampling in Section 4.2.1.

4.2.3 Stratified Simple Random Sampling

For a stratified simple random sample, the population is divided into $h = 1, ..., H$ disjoint strata, with the sampling probabilities $\pi_i = \pi_j$ for $i, j \in h$. Samples are taken independently from each stratum, and so $\pi_{ij} = \pi_i\pi_j$ when i and j are in different strata. In particular, the overall variance is calculated as the sum of the variances from each stratum.

To obtain an expression for the design variance $Var(\hat{Y})$ of $\hat{Y}$ under the stratified simple random sampling case, first define y_{ih} as the value of y for the i'th unit in stratum h, π_h as the probability of selection for units in stratum h, N_h as the population size of stratum h, and $\bar{y}_h = \sum_{i=1}^{N_h} y_{ih}/N_h$ as the population mean of y in stratum h. Then

$$Var(\hat{Y}) = \sum_{h=1}^{H}\Big(\frac{1}{\pi_h} - 1\Big)N_h S_{hy}^2, \text{ where } S_{hy}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{ih} - \bar{y}_h)^2 \tag{4.15}$$

is a well-known result; see for example Theorem 3.1 in Wu and Thompson (2020) which provides an expression for the variance of an estimate of population mean, and noting that $Var(\hat{\bar{Y}}) = Var(\hat{Y})/N^2$.

Using (4.15), equation (3.13) in the stratified SRSWOR case becomes

$$\begin{aligned}
\boldsymbol{b}^T\boldsymbol{V}\boldsymbol{b} &= \frac{1}{N^2}Var_d(\hat{v}) \\
&= \frac{1}{N^2}\sum_{h=1}^{H} N_h\Big(\frac{1}{\pi_h^A}-1\Big)S_{hv}^2, \text{ where} \\
S_{hv}^2 &= \frac{1}{N_h-1}\sum_{i=1}^{N_h}(v_i-\bar{v}_h)^2, \\
v_i &= \pi_i^B(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i, \text{ and} \\
\bar{v}_h &= \frac{1}{N_h}\sum_{i=1}^{N_h} v_i
\end{aligned} \tag{4.16}$$

Under stratified simple random sampling the first component of (4.5) becomes

$$\begin{aligned}
&\frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{\pi_i^A}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \\
&= \frac{1}{N^2}\sum_{h=1}^{H}\frac{1}{\pi_h^A}\sum_{i=1}^{N_h}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \\
&= \frac{1}{N^2}\sum_{h=1}^{H}\frac{1}{\pi_h^A}\sum_{i=1}^{N_h}\frac{\pi_i^B}{1-\pi_i^B}v_i^2 \text{ where } v_i = \pi_i^B(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i
\end{aligned} \tag{4.17}$$

Denoting M_{strat} as the sum of (4.16) and (4.17), the variance expression for the optimisation can be obtained as

$$\begin{aligned}
M_{strat} &= \frac{1}{N^2}\sum_{h=1}^{H} N_h\Big(\frac{1}{\pi_h^A}-1\Big)S_{hv}^2 + \frac{1}{N^2}\sum_{h=1}^{H}\frac{1}{\pi_h^A}\sum_{h=1}^{N_h}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \\
&= \frac{1}{N^2}\sum_{h=1}^{H}\frac{1}{\pi_h^A}\Big(N_hS_{hv}^2 + \sum_{i=1}^{N_h}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big) - \frac{1}{N^2}\sum_{h=1}^{H}N_hS_{hv}^2 \\
&= \frac{1}{N^2}\sum_{h=1}^{H}\frac{1}{\pi_h^A}D_h - \frac{1}{N^2}\sum_{h=1}^{H}N_hS_{hv}^2 \text{ where } D_h = N_hS_{hv}^2 + \sum_{i=1}^{N_h}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}
\end{aligned} \tag{4.18}$$

The aim is to minimise M_{strat} subject to the cost constraint $C_t = \sum_{h=1}^{H}\pi_h^A c_h$ where

$c_h = \sum_{i=1}^{N_h} c_i$. The Lagrangian is

$$\mathcal{L} = \frac{1}{N^2}\sum_{h=1}^{H}\frac{1}{\pi_h^A}D_h - \frac{1}{N^2}\sum_{h=1}^{H}N_h S_{hv}^2 + \lambda\Big[\sum_{h=1}^{H}\pi_h^A c_h - C_t\Big]$$

Application of the Lagrange multiplier approach gives

$$\pi_h^A = \frac{1}{\sqrt{\lambda}}\sqrt{\frac{D_h}{N^2 c_h}}$$

and substituting this expression for π_h^A into the cost constraint $C_t = \sum_{h=1}^{H}\pi_h^A c_h$ results in

$$\frac{1}{\sqrt{\lambda}} = \frac{C_t}{\sum_{h=1}^{H}\sqrt{\frac{D_h c_h}{N^2}}}.$$

The optimal π_h^A values are thus derived as

$$\begin{aligned}
\pi_h^A &= \frac{C_t}{\sum_{h=1}^{H}\sqrt{\frac{D_h c_h}{N^2}}}\sqrt{\frac{D_h}{N^2 c_h}} \\
&= \frac{C_t\sqrt{D_h/c_h}}{\sum_{h=1}^{H}\sqrt{D_h c_h}} \\
&= \frac{C_t\sqrt{\Big(N_h S_{hv}^2 + \sum_{i=1}^{N_h}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big)c_h^{-1}}}{\sum_{h=1}^{H}\sqrt{\Big(N_h S_{hv}^2 + \sum_{i=1}^{N_h}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big)c_h}} \\
\implies \pi_{h,opt}^A &\propto \sqrt{\frac{N_h S_{hv}^2 + \sum_{i=1}^{N_h}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}}{c_h}}. \qquad (4.19)
\end{aligned}$$

Remark 4.2. From (4.19) the optimal selection probabilities can be seen to be greater when

- c_h is smaller. As stratum cost increases, it becomes less desirable to include it in the sample.
- N_h or S_{hv}^2 are greater. Strata with a larger population size or larger variation in v should be given a greater sample size.
- $\sum_{i=1}^{N_h}(\pi_i^B)^3(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}$ is larger. We can interpret this as a weighted sum of

the squared predictions for $(y_i - \mu)/\pi_i^B$ within stratum h. Strata that tend to have values of $\boldsymbol{x}$ associated with more extreme values of y relative to μ should have a greater sample size. At the same time, the contribution of units with π_i^B very close to zero will be downweighted given the weights $(\pi_i^B)^3(1 - \pi_i^B)$ in the sum.

4.2.4 Optimal sample design in practice

The expressions for the optimal sampling probabilities (4.11) and (4.19) require knowledge of the terms $\boldsymbol{x}_i$, $\pi_i^B = \pi(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$, and $\boldsymbol{b}$ for the population $\mathcal{U}$. We may estimate $\boldsymbol{b}$ and $\boldsymbol{\theta}_0$ using data from a sample A that is already available (for example, a historical sample, or a pilot study), and the non-probability sample B. However, the value of $\boldsymbol{x}_i$ will generally not be known for all units in the population. Instead, it is assumed that there is information about a set of design variables $\boldsymbol{q}_i$ for $i \in \mathcal{U}$. These design variables may be obtained from sources such as a recent population census or administrative data such as business tax data, and may have some overlap with the set of variables $\boldsymbol{x}$.

How might one implement an optimal sample design in practice when $\boldsymbol{x}$ is not available for the whole population, or only partially available? Clark and Steel (2022) proposed to minimise the anticipated variance of $\hat{\mu}_{IPW}$, conditional on the available information $\boldsymbol{q}$. In this thesis, I suggest an approach which is loosely motivated by their idea. Specifically, a multiple imputation-type approach (see Rubin 1987) is adopted to estimate $\hat{\pi}_{i,opt}^A$. In practice, the following steps may be followed:

1. Use available records i which have both $\boldsymbol{x}_i$ and $\boldsymbol{q}_i$ to impute $\hat{\boldsymbol{x}}_i$ for the rest of the population
2. Produce $\hat{\pi}_{i,opt}^A$ using $\boldsymbol{x}_i$ and, where needed, the imputed $\hat{\boldsymbol{x}}_i$
3. Repeat Steps 1 and 2 m times
4. Calculate the mean of the m values of $\hat{\pi}_{i,opt}^A$ and use that in the sample design

One possible approach for creating the estimated values is to apply the k-Nearest Neighbour (kNN) method to mass impute values $\hat{\boldsymbol{x}}_i$ where they are missing, using $\boldsymbol{q}_i$ as the

distance measure. The idea of mass imputation using kNN has been explored previously, for example in Rivers (2007) and Yang et al. (2021). An alternative that may be less computationally intensive is to use hot deck imputation, where the hot deck classes are devised based on $\boldsymbol{q}$.

Data for the imputation may be obtained from an existing survey A which collects $\boldsymbol{x}_i$, or from a pilot study. Additionally, if we can link records in B to the population frame, then the linked dataset will contain both $\boldsymbol{x}_i$ and $\boldsymbol{q}_i$ for each unit $i \in B$. The linked dataset may then supply $\boldsymbol{x}_i$ for $i \in B$, and imputed values $\hat{\boldsymbol{x}}_i$ for $i \in \mathcal{U} \setminus B$ can be obtained from the existing survey or pilot study data.

The proposed approach for finding optimal selection probabilities will be evaluated in the simulation study described in Section 4.4.

4.3 Split-Population Sample Design for A

In Chapter 3 it was assumed that it was possible to identify which units in A are also in B. In this section, a stronger assumption is made whereby it is possible to identify which units on the population frame are not in B. This could be possible, for example, when a common unit identifier exists on the frame and on B. If this is the case, one may decide to design the reference sample A based only on the sub-population $\mathcal{U} \setminus B$. This approach has similarities to the "split-population" scenario (see Zhang 2019) where $\mathcal{U}$ is divided into incomplete "frames" B and $C = \mathcal{U} \setminus B$, B is fully sampled, and A is a sample from C (see also Lohr 2021). The difference is that in this case the variable y is not collected in A.

The split-population OE (SPOE) pseudo-likelihood is given by

$$\begin{aligned} l_{SPOE}(\boldsymbol{\theta}) &= \sum_{i \in B} \log(\pi(\boldsymbol{x}_i, \boldsymbol{\theta})) + \sum_{i \in A} d_i^A (1 - \delta_i) \log(1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})) \\ &\equiv \sum_{i \in B} \log(\pi(\boldsymbol{x}_i, \boldsymbol{\theta})) + \sum_{i \in A} d_i^A \log(1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})) \end{aligned} \tag{4.20}$$

where δ_i is always zero for units in A and the d_i^A are weights that now sum to $(N - N_B)$.

Assuming that a logistic model holds for the propensity scores, the pseudo-likelihood becomes

$$l_{SPOE}(\boldsymbol{\theta}) = \sum_{i \in B} \left\{ \boldsymbol{x}_i^T \boldsymbol{\theta} - \log(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\theta})) \right\} - \sum_{i \in A} d_i^A \log(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\theta})) \tag{4.21}$$

and the resulting score function is given as

$$S_{SPOE}(\boldsymbol{\theta}) = \sum_{i \in B} \boldsymbol{x}_i - \sum_{i \in B} \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i - \sum_{i \in A} d_i^A \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i \tag{4.22}$$

The score equation $S_{SPOE} = \mathbf{0}$ may be solved for $\boldsymbol{\theta}$ using an iterative procedure such as the Newton-Raphson approach.

Remark 4.3. The split-population approach described above would be appropriate if the data in B is of high quality. When the variable of interest y in B or some of the auxiliary variables $\boldsymbol{x}$ in A or B have measurement error, or if it is suspected that the missingness in B also depends on y, then having some overlap between A and B would be preferable. The overlapping sample would facilitate building a model to correct for the measurement error issues, or to address SNAR issues in B. The question of how to address measurement error in a subset of the auxiliary variables in A is addressed in Chapter 5. The question of how best to design an overlapping sample between A and B when there are quality issues in B is left for future research.

I_i and δ_i are no longer independent since the sample A is drawn specifically from $\mathcal{U} \setminus B$. The probabilities of selection π_i^A are now dependent on whether or not unit i is in B. In other words, $\pi_i^A = P(I_i = 1 | \delta_i)$. However, it can be shown that the IPW estimator (3.1) which uses $\hat{\pi}_i^B$ obtained by maximising the split-population pseudo-likelihood (4.20) is still consistent.

Theorem 4.1. *Under assumptions **A1** and **A2** from Section 1.1 and the conditions listed in Appendix B.5, and given that a logistic participation probability model holds,* $E(\hat{\mu}_{SPOE}) \to \mu_y$ *and* $|\hat{\mu}_{SPOE} - \mu_y| \xrightarrow{p} 0$ *as the sizes of the population and samples tend toward infinity.*

The proof for Theorem 4.1 is given in Appendix B.6.

4.4 Simulation Study

A simulation study was conducted to examine the performance of the IPW estimates formed when the reference sample A can be designed to supplement the non-probability data. The general setup of the simulation, including creation of the population data items and the selection mechanism for B, is identical to the simulation study in Section 3.5, and so the interested reader is pointed to that section for further information.

The following is noted to aid the reader. $R = 2,000$ simulation iterations were conducted for various N_B sample sizes of 750, 2,000, 50,000 and 140,000. The reference sample size A was kept fixed at $n_A = 5,000$, and the sample design used was randomised systematic PPS sampling. For the optimal designs in this simulation, the cost to enumerate a unit c_i was set at a constant value of 1 for all units in the population. The performance of the different estimators was evaluated using %RB and %RRMSE values, as defined in Section 3.5.

Results were produced for the following estimators:

- OE - IPW estimates using the OE propensity score estimator. Selection probabilities in A were based on the auxiliary variable x_3.
- Optimal OE (Opt. OE) - Optimal selection probabilities were produced for A using (4.11), assuming that all auxiliary information was available for the population.
- "Anticipated" Optimal OE (Est. Opt) - Optimal selection probabilities were estimated using the multiple imputation approach described in Section 4.2.4, assuming all auxiliary information was missing from the frame except for the variable x_3.
- Split-Population OE (SPOE) - Sample A is drawn from $\mathcal{U} \setminus B$, and used to estimate propensities using the OE approach. Selection probabilities in A are based on the auxiliary variable x_3.

For the anticipated optimal OE scenario, random hot-deck imputation was used in a multiple imputation manner to estimate optimal π_i^B values based on (4.11). Ten sets of hot deck imputed values were produced, using the deciles of the variable x_3 to identify the domains within which imputes would be selected. For this simulation setup, it was assumed that the variable x_3 was available on a population-level dataset. These values were used to impute values for the variables x_1, x_2 and x_4 in the segment $C = \mathcal{U} \setminus B$.

Tables 4.1 and 4.2 contain the results from the simulation study for values of $\rho(y, \boldsymbol{x}^T\beta) = 0.3$ and $\rho(y, \boldsymbol{x}^T\beta) = 0.7$. The use of optimal selection probabilities for A reduces the RRMSE in the Opt. OE case compared with the OE case, as would be expected. Designing A so that it is drawn only from $\mathcal{U} \setminus B$ results in further gains in accuracy for larger N_B - in the largest N_B scenario, this approach results in the lowest RRMSE of all the options. The SPOE estimator is also unbiased.

Table 4.1: Performance of different sample designs for A - $\rho(y, \boldsymbol{x}^T\beta) = 0.3$

	$N_B = 750$		$N_B = 2,000$		$N_B = 50,000$		$N_B = 140,000$	
	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE
OE	-0.13	9.84	-0.06	5.80	-0.01	0.77	-0.02	0.32
Opt. OE	0.05	9.69	0.10	5.70	0.00	0.72	-0.01	0.27
Est. Opt	-0.16	9.87	-0.02	5.83	-0.02	0.78	0.00	0.29
SPOE	-0.09	9.92	-0.14	5.84	0.00	0.75	-0.01	0.22

It is encouraging to see that the use of "anticipated" optimal selection probabilities can also be effective. In the simulation, for the largest sample size of N_B it was possible to achieve a reduction in RRMSE when using these estimated optimal probabilities over the un-optimised design for A.

Table 4.2: Performance of different sample designs for A - $\rho(y, \boldsymbol{x}^T\beta) = 0.7$

	$N_B = 750$		$N_B = 2,000$		$N_B = 50,000$		$N_B = 140,000$	
	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE
OE	-0.20	4.44	-0.16	2.65	-0.02	0.52	-0.01	0.27
Opt. OE	0.00	4.19	-0.01	2.43	-0.01	0.42	0.00	0.21
Est. Opt	-0.25	4.48	-0.12	2.72	-0.01	0.52	0.00	0.24
SPOE	-0.15	4.49	-0.22	2.69	0.00	0.46	0.00	0.13

4.5 Concluding remarks

In this chapter, several ideas have been proposed for more efficiently designing the reference sample in conjunction with an accompanying IPW estimator which uses estimated propensities produced through the OE pseudo-likelihood. Optimal design of the reference sample A to collect $\boldsymbol{x}$ presents an opportunity for the NSO to repurpose their survey program, facilitating the use of multiple non-probability (especially big) datasets as the primary source for statistics, supplemented by a small auxiliary reference sample.

These design ideas were examined empirically through a simulation study. The results indicate that when data for $\boldsymbol{x}$ are available on the population frame, using optimal selection probabilities for the design provides gains in efficiency. In the simulation study, the gains ranged from about a 1.5% reduction in RRMSE for small N_B to about a 15.6% reduction for the largest N_B size when ρ was 0.3. When $\boldsymbol{x}$ is not fully available on the frame, but B is large, the use of "anticipated" optimal selection probabilities based on predicted values $\hat{\boldsymbol{x}}$ can also lead to gains, especially if N_B is large. For the simulation study, a reduction of about 10% in RRMSE was obtained for the largest N_B sample size. Finally, if it is possible to identify which units on the population frame are not on B, selecting the sample A from $\mathcal{U} \setminus B$ using a split-population approach can also result in more efficient estimates, especially for larger N_B. In the simulation study, a reduction in RRMSE of about 31% was obtained under the split-population approach for the largest N_B sample when $\rho = 0.3$.

How best to design A using split-population design principles remains an open question. This chapter has not explored how to obtain optimal selection probabilities for the split-population sample. Additionally, the split-population approach requires the data from B to be of high quality. When this is not the case, it will be preferable to include some overlapping sample between A and B. There is a question of how best to design this overlap to facilitate the production of fit-for-purpose statistics, taking into account possible measurement error issues, SNAR issues, and so on. One possible direction for how to design the overlap may be to utilise the audit or validation sampling ideas proposed in Zhang (2021) and Zhang (2023).

Chapter 5

Addressing Measurement Error and Selection-Not-At-Random

5.1 Introduction

The methods that were considered in previous chapters for addressing selection bias generally make two simplifying assumptions: the data contained in the reference and non-probability samples are without error, and the missingness in the non-probability sample satisfies the ignorability assumption described in Section 1.1. In reality, one may not be confident that these assumptions hold. The without-error assumption may be violated, for example, if the reference sample is from a year or two ago, rendering some auxiliary data items from that sample out of date. An administrative dataset in which businesses with low turnover are less likely to report their data may violate the ignorability assumption if the goal is to obtain statistics about turnover from that dataset.

This chapter will consider methods for addressing selection bias in the non-probability sample while relaxing these two assumptions. As in Chapters 3 and 4, the focus is on a Type II data structure. Recall that in this scenario, the auxiliary variables are measured in both the non-probability sample and a reference probability sample, while the target variable is measured only in the non-probability sample.

For the case of measurement error, the specific scenario of interest is when measurement error exists in some of the auxiliary variables $\boldsymbol{x}$ present on the reference sample (that is, the probability sample A). In this chapter, this type of measurement error will sometimes be referred to as covariate measurement error given that the variables $\boldsymbol{x}$ are used as covariates in a model to predict participation probabilities.

For the missingness in the non-probability sample, the scenario of interest is when a nonignorable selection mechanism holds. In this case, participation in the non-probability sample depends on the variable of interest y as well as the auxiliary variables $\boldsymbol{x}$. Note that the reference sample is still assumed to have full response in this scenario.

5.1.1 Measurement Error

In this thesis measurement error has been defined broadly to be a misalignment between the true value of a variable and the value actually captured by the sample. In addition to the use of out-of-date data, other examples of measurement error may include, for example, the use of concept definitions which differ from what is intended, input error, or rounding error. This error may be present in the target or auxiliary data items, and on the reference sample or the non-probability sample. This chapter focuses on the case where the auxiliary data items $\boldsymbol{x}$ are measured in both A and B, and measurement error is present in a subset of $\boldsymbol{x}$ on A. On the other hand, the target variable y is measured only in B, and without error. At the end of the chapter, I will briefly touch on the other case where measurement error is present in the non-probability sample.

In Carroll et al. (2006) three effects of covariate measurement error are discussed: it can lead to a loss of statistical power in detecting relationships among variables, it masks the features of the data, and more pertinently for this thesis it introduces bias in the estimation of model parameters. Since the error-prone auxiliary variables from A are used to construct the participation probability model, this so-called attenuation bias can lead to inaccurately estimated participation probabilities for the non-probability sample units, and thus ineffective correction of selection bias.

There has been a large body of work undertaken on the topic of measurement error. Fuller (1987) and Carroll et al. (2006) produced monographs examining the effect of measurement error on linear and non-linear models, while the monograph published by Yi (2017) looked at analysing measurement error and misclassification problems arising in the medical research and epidemiological fields. Stefanski (2000) provides a brief introduction to measurement error problems and reviews some key contributions to the topic.

The literature on addressing measurement error issues specifically for non-probability samples is still relatively sparse. Valliant and Dever (2011) pointed out that covariate measurement error may lead to biased estimation of the participation probabilities for B. Schenkel and Zhang (2022) developed a modelling approach to adjust for misclassification in a target variable in an administrative dataset, using a second classifier from another source that is also subject to misclassification. Dharma et al. (2025) examined the case when a data item collected in the full reference sample is used to identify a subgroup of interest corresponding to the target population of a non-probability sample, and that data item could be misclassified - specifically, that there are some units that should be included in the subgroup but are not. The authors proposed to estimate the probability of misclassification and then generate a predicted classification to be used to define the subgroup.

The calibration approach proposed in Kim and Tam (2021) for integrating probability and non-probability samples allows for measurement error in the target variable on either of the two samples. Given that their approach applies to a Type III data structure, it is possible to fit a measurement error model like (1.1) using the overlapping units in the reference and non-probability samples. The fitted model can then be used to adjust the mis-measured data for the target variable on the reference sample. In Chapter 2, an empirical study was conducted to examine the impact of measurement error in the target variable on estimates of totals, and the effectiveness of using a measurement error model to correct for those errors.

5.1.2 Selection-Not-At-Random

The ignorability assumption is often made when dealing with selection bias; see for example Wu (2022) and Rueda-Sánchez et al. (2025). Ignorability implies that selection into the non-probability dataset does not depend on the target variable once the auxiliary variables $\boldsymbol{x}$ have been taken into account. Recall that Andridge et al. (2019) and Boonstra et al. (2021) refer to this type of selection process as Selection At Random (SAR). When selection into the non-probability dataset does depend on the target variable y, then we instead have a nonignorable, or Selection Not At Random (SNAR), selection mechanism. Chapter 2 explored empirically how various estimators perform under the SAR versus SNAR situations. This chapter will examine how to address SNAR under the Type II data scenario, where the target variable is available on the non-probability sample only, and auxiliary data items are available on both the non-probability and reference samples.

Recently, researchers are beginning to turn their attention towards the problem of addressing SNAR in non-probability samples. Andridge et al. (2019) and Little et al. (2020) developed an index to assess the magnitude of nonignorable selection bias in estimated means from a non-probability dataset. West and Andridge (2023) applied the index to polling data. Hammon and Zinn (2024) applied the method to data on political attitudes from a non-probability sample, and derived some guidelines for using the method in practice.

Kim and Morikawa (2023) employed the sample empirical likelihood approach to address selection bias under a SNAR selection mechanism using a two-step process. First, an initial participation probability is estimated using a parametric model. The empirical likelihood method is then used to incorporate a bias-calibration constraint and calibration constraints relating to known population means of auxiliary variables.

Marella (2023) also considered a sample empirical likelihood method for addressing SNAR bias, combining it with the profile likelihood approach in an iterative fashion to obtain the multinomial probabilities for the units in the non-probability sample. Under a Type II data structure, sample estimates for relevant auxiliary variables may be used as calibration

constraints in the empirical likelihood formulation.

Liu et al. (2024) considered the case of a non-probability sample with an ignorable SAR selection mechanism, but with nonignorable item nonresponse for a variable of interest in the non-probability sample. The authors employed logistic models to estimate participation probabilities and the response probabilities for the variable of interest, and a superpopulation model to explain the relationship between the variable of interest and covariates. A Bayesian approach was used to estimate model parameters.

Liu et al. (2025) proposed a superpopulation model approach combined with a model for the informative sampling mechanism. They formulated the sample likelihood using these two models and maximised this to estimate the parameters of the superpopulation model. The fitted superpopulation model was then used to impute the target variable for units not in the non-probability sample. Their method assumed a linear relationship between the target variable and the covariates, and that the covariates were available for the units not in the non-probability sample.

Given the absence of the target variable on the reference sample in a Type II data scenario, it is not possible to use it directly to fit a participation probability model such as those proposed or examined in previous chapters of this thesis. Xu and Nandram (2024) proposed using Classification and Regression Tree (CART) imputation to get around this problem. The authors used the values of the target variable and auxiliary items on the non-probability sample to fit the CART, and it was then applied to the auxiliary items on the reference sample to obtain imputed values for the target variable. The imputed values were then used in the participation probability model parameter estimation process. Since the CART model for the target variable is fitted using the non-probability sample, and the data suffers from SNAR in the target variable, this approach may not lead to an accurate model.

5.1.3 Work undertaken in this chapter

A common thread in addressing the covariate measurement error and SNAR issues is a need to formulate two models: a model for the selection mechanism, and an "outcome" model which relates an error-prone or missing data item to some other data items. In the case of measurement error in the auxiliary variables, the outcome model links a mis-measured covariate to its correctly measured version, and potentially along with other variables related to it. For the SNAR case, the outcome model would link a target variable y not captured on the reference sample with variables $\boldsymbol{x}$ which are measured in both the reference sample and the non-probability sample. This dissertation examines these two cases separately; the case where both SNAR in B and covariate measurement error in A is present is not considered.

In this chapter a semiparametric approach is presented to simultaneously address covariate measurement error and selection bias, when it is possible to integrate the reference sample and the non-probability sample. It is assumed that the mis-measured auxiliary variables do not influence participation probabilities over and above the correctly measured versions of those variables. Under this assumption, a parametric model for the participation probabilities is combined with a nonparametric model for the outcome to obtain a semiparametric maximum likelihood estimator for the parameters of the participation probability model. Of note, it is shown that the proposed framework can also be used to correct selection bias in a non-probability sample that has SNAR issues.

The rest of the chapter is structured as follows. Section 5.2 outlines the basic setup for this chapter. Section 5.3 introduces the proposed semiparametric framework for correcting measurement error and selection bias. I consider the case when covariate measurement error is present in the reference sample and show how the framework can be applied to the case of SNAR. In Section 5.4 it is shown that the proposed semi-parametric estimating function is consistent, and a bootstrap approach for estimating variance is outlined. In Section 5.5 the performance of the proposed approach is examined through a simulation study. The approach is applied using real-world data in Section 5.6. Section 5.7 discusses how a similar approach may be employed to address measurement error in the auxiliary

variables in B. Finally, Section 5.8 provides some concluding remarks.

5.2 Basic Setup

Let $\mathcal{U}$ be a finite population of size N. For each unit i in the population, define values $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)^T$ for a data item of interest y and auxiliary data items $(\boldsymbol{x}, \boldsymbol{z})$. $\boldsymbol{x}$ may suffer from measurement error, while $\boldsymbol{z}$ is always free from measurement error. Denote the mis-measured version of $\boldsymbol{x}$ using $\boldsymbol{x}^*$. This chapter focuses on estimating the population mean $\mu_y = \bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$, but the approach also applies to population totals.

A reference probability sample A of size n_A is drawn from the population. A contains information for $\boldsymbol{z}$ and $\boldsymbol{x}^*$, and does not contain information for y. Define $\pi_i^A = P(i \in A|\mathcal{U})$ as the inclusion probability for unit i being in the probability sample, and $d_i^A = 1/\pi_i^A$ as the survey design weight for $i \in A$. The π_i^A and d_i^A values are known from the sample design.

A non-probability sample B is also available from the population. B contains information for $\boldsymbol{z}$ and $\boldsymbol{x}$, as well as the variable of interest y. Let $\delta_i = I(i \in B)$ be an indicator variable for unit i being included in the sample B. The non-probability sample size is $N_B = \sum_{i=1}^{N} \delta_i$. In contrast to π_i^A in the reference probability sample, $\pi_i^B = P(\delta_i = 1|\mathcal{U})$ is unknown and needs to be estimated. Define $C = \mathcal{U} \setminus B$ as the units in the population not included in B.

A number of assumptions are made regarding the selection mechanism for B. These assumptions are:

A1 *Positivity*: Conditional on $(\boldsymbol{x}_i, \boldsymbol{z}_i)$, $P(\delta_i = 1|\boldsymbol{x}_i, \boldsymbol{z}_i) > 0, \forall i \in \mathcal{U}$

A2 *Independence*: Conditional on $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ and $(\boldsymbol{x}_j, \boldsymbol{z}_j)$, $\delta_i \perp \delta_j, i \neq j, \forall i, j \in \mathcal{U}$

The positivity assumption **A1** states that conditional on $(\boldsymbol{x}, \boldsymbol{z})$ every unit in the population has a non-zero chance of inclusion in B. This may not always hold - for example, in a big

dataset generated by a social media platform, only those persons who are members of the platform have a chance of inclusion in the big dataset. The independence assumption **A2** facilitates forming a likelihood function (see for example Wu 2023).

A third assumption of ignorability will initially be made to construct the measurement error framework. Ignorability implies $P(\delta_i = 1|y, \boldsymbol{x}, \boldsymbol{z}) = P(\delta_i = 1|\boldsymbol{x}, \boldsymbol{z})$. In other words, given the covariates $(\boldsymbol{x}, \boldsymbol{z})$, δ_i and y_i are independent and a SAR scenario is applicable. Later in the chapter, it will be shown that the proposed measurement error framework can also be applied to resolve SNAR issues in the non-probability sample. In other words, I will dispense of the SAR assumption and enable δ_i to depend on y_i. At times this will also be referred to as nonignorable selection.

The following assumption will also be made in relation to the measurement error and selection mechanisms (see for example Yi 2017):

A3 *Non-differential measurement error*: Conditional on the set of covariates $(\boldsymbol{x}_i, \boldsymbol{z}_i)$, δ_i and $\boldsymbol{x}_i^*$ are independent

Assumption **A3** says that the mis-measured $\boldsymbol{x}^*$ does not give any further information for inference about the participation probabilities given the true covariates. In other words, $P(\delta|\boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{z}) = P(\delta|\boldsymbol{x}, \boldsymbol{z})$.

Finally, it is assumed that there is some overlap between A and B, and it is possible to accurately link the units in A and B. The overlap assumption is reasonable if A and B are not too small, or if the sample design of A facilitates some overlap with units in B. The linkage assumption may be satisfied if there exists a common unit identifier available on B and the reference sample A, or if it is possible to accurately link records in A and B using a common set of linking variables.

5.3 Proposed Method

Recall that the end goal in this chapter is to produce an estimate of the population mean using B. This can be achieved by producing $\hat{\pi}_i^B$ and feeding it into an Inverse Probability Weighted (IPW) estimator such as

$$\hat{\mu}_{IPW} = \frac{1}{\hat{N}_B} \sum_{i \in B} \frac{y_i}{\hat{\pi}_i^B} \tag{5.1}$$

where $\hat{N}_B = \sum_{i \in B} (\hat{\pi}_i^B)^{-1}$. The problem addressed in this chapter is how best to produce $\hat{\pi}_i^B$.

It is assumed that the π_i^B are governed by a parametric model, so that $\pi^B = P(\delta = 1|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_0) = \pi(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ are the true but unknown model parameters for the selection mechanism. I first consider the case when A corresponds to a population-level dataset, for instance when it is the population frame or data from a population census.

The likelihood function for $\{\delta_i, \boldsymbol{x}_i, \boldsymbol{z}_i; i \in B\}$ and $\{\boldsymbol{x}_i^*, \boldsymbol{z}_i; i \in \mathcal{U}\}$ is

$$L(\boldsymbol{\theta}) = \prod_{i \in \mathcal{U}} \pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})^{\delta_i} \Big\{ \int (1 - \pi(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta})) f(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{x} \Big\}^{1-\delta_i}. \tag{5.2}$$

The corresponding log-likelihood function is

$$l(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \delta_i \log \pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) + \sum_{i \in \mathcal{U}} (1 - \delta_i) \log \Big\{ \int (1 - \pi(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta})) f(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{x} \Big\} \tag{5.3}$$

and, under mild conditions enabling the interchange of differentiation and integration, the score function is given by

$$\begin{aligned} s(\boldsymbol{\theta}) &= \frac{\partial l}{\partial \boldsymbol{\theta}} \\ &= \sum_{i \in \mathcal{U}} \delta_i \frac{\pi'(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})}{\pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})} - \sum_{i \in \mathcal{U}} (1 - \delta_i) \frac{\int \pi'(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta}) f(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{x}}{\int (1 - \pi(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta})) f(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{x}} \end{aligned} \tag{5.4}$$

where $\pi'(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \partial \pi(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) / \partial \theta$ is the derivative of $\pi(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Solving

the score function requires knowledge of $f(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i)$. It may be possible to obtain this information through, for example, taking a validation subsample from the original sample A. However, this is not always possible to do, for example due to budget constraints. Note that data is available on $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ from B, and $(\boldsymbol{x}_i^*, \boldsymbol{z}_i)$ from A. If it is possible to link the data from A and B, then one can obtain $(\boldsymbol{x}_i, \boldsymbol{x}_i^*, \boldsymbol{z}_i)$ for records in $A \cap B$. Thus, one can instead estimate $f(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i, \delta_i = 1) = f_1(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i)$ using $A \cap B$, where $f_1(\cdot) = f(\cdot|\delta_i = 1)$.

Using Bayes' Theorem, we have

$$\begin{aligned} f_1(\boldsymbol{x}|\boldsymbol{x}^*, \boldsymbol{z}) &= f(\boldsymbol{x}|\boldsymbol{x}^*, \boldsymbol{z}, \delta = 1) \\ &= \frac{f(\boldsymbol{x}, \delta = 1|\boldsymbol{x}^*, \boldsymbol{z})}{P(\delta = 1|\boldsymbol{x}^*, \boldsymbol{z})} \\ &= \frac{P(\delta = 1|\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{z}) f(\boldsymbol{x}|\boldsymbol{x}^*, \boldsymbol{z})}{P(\delta = 1|\boldsymbol{x}^*, \boldsymbol{z})} \end{aligned}$$

and assuming that the non-differential error mechanism holds, we have $P(\delta = 1|\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{z}) = P(\delta = 1|\boldsymbol{x}, \boldsymbol{z}) = \pi^B$. So

$$f(\boldsymbol{x}|\boldsymbol{x}^*, \boldsymbol{z}) = f_1(\boldsymbol{x}|\boldsymbol{x}^*, \boldsymbol{z}) \frac{P(\delta = 1|\boldsymbol{x}^*, \boldsymbol{z})}{\pi^B}. \tag{5.5}$$

Substituting (5.5) into the score function (5.4) gives

$$\begin{aligned} s(\boldsymbol{\theta}) =& \sum_{i \in \mathcal{U}} \delta_i \frac{\pi'(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})}{\pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})} - \\ & \sum_{i \in \mathcal{U}} (1 - \delta_i) \frac{\int \pi'(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta}) f_1(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i) P(\delta_i = 1|\boldsymbol{x}_i^*, \boldsymbol{z}_i) / \pi(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}}{\int (1 - \pi(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta})) f_1(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i) P(\delta_i = 1|\boldsymbol{x}_i^*, \boldsymbol{z}_i) / \pi(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}} \\ =& \sum_{i \in \mathcal{U}} \delta_i \frac{\pi'(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})}{\pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})} - \sum_{i \in \mathcal{U}} (1 - \delta_i) \frac{\int \pi'(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta}) (1 - \pi(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta}))^{-1} O(\boldsymbol{x}, \boldsymbol{z}_i) f_1(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{x}}{\int O(\boldsymbol{x}, \boldsymbol{z}_i) f_1(\boldsymbol{x}|\boldsymbol{x}_i^*, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{x}} \end{aligned} \tag{5.6}$$

where $O(\boldsymbol{x}, \boldsymbol{z}_i) = (1 - \pi(\boldsymbol{x}, \boldsymbol{z}_i))/\pi(\boldsymbol{x}, \boldsymbol{z}_i)$.

One may evaluate the integrals by assuming a parametric form for $f_1(\boldsymbol{x}|\boldsymbol{x}^*, \boldsymbol{z})$ (see for example Riddles et al. 2016). However, this approach carries a risk of model misspecification. Instead, in this dissertation I propose to adopt a nonparametric regression approach

similar to that considered in Morikawa et al. (2017). Specifically, I take advantage of the fact that the integral terms in (5.6) are conditional expectations. In particular, they are expectations that are conditional on membership in B. That is,

$$\int (\cdot) f_1(\boldsymbol{x}|\boldsymbol{x}^*, \boldsymbol{z}) \mathrm{d}\boldsymbol{x} = E_1[(\cdot)|\boldsymbol{x}^*, \boldsymbol{z}]$$

If A comprises the full population, then the Nadaraya-Watson (Nadaraya 1964, Watson 1964) kernel regression estimator of $E_1[(\cdot)|\boldsymbol{x}^*, \boldsymbol{z}]$ is given by

$$\hat{E}_1[(\cdot)|\boldsymbol{x}^*, \boldsymbol{z}] = \frac{\sum_{j \in B} K_h(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)(\cdot)}{\sum_{j \in B} K_h(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)} \tag{5.7}$$

where $\tilde{\boldsymbol{x}} = (\boldsymbol{x}^*, \boldsymbol{z})$ is a combined vector of covariates, and K_h relates to a given kernel function with bandwidth h. The score function (5.6) is thus estimated by

$$\begin{aligned}
\hat{s}(\boldsymbol{\theta}) &= \sum_{i \in \mathcal{U}} \delta_i \frac{\pi'(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})}{\pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})} - \\
&\quad \sum_{i \in \mathcal{U}} (1 - \delta_i) \frac{\sum_{j \in B} \pi'(\boldsymbol{x}_j, \boldsymbol{z}_i; \boldsymbol{\theta})(1 - \pi(\boldsymbol{x}_j, \boldsymbol{z}_i; \boldsymbol{\theta}))^{-1} O(\boldsymbol{x}_j, \boldsymbol{z}_i) K_h(\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{x}}_i)}{\sum_{k \in B} O(\boldsymbol{x}_k, \boldsymbol{z}_i) K_h(\tilde{\boldsymbol{x}}_k, \tilde{\boldsymbol{x}}_i)} \\
&= \sum_{i \in \mathcal{U}} \delta_i \frac{\pi'(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})}{\pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})} - \sum_{i \in \mathcal{U}} (1 - \delta_i) \sum_{j \in B} \pi'(\boldsymbol{x}_j, \boldsymbol{z}_i; \boldsymbol{\theta})(1 - \pi(\boldsymbol{x}_j, \boldsymbol{z}_i; \boldsymbol{\theta}))^{-1} w_{ij} \\
&\text{where } w_{ij} = \frac{O(\boldsymbol{x}_j, \boldsymbol{z}_i) K_h(\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{x}}_i)}{\sum_{k \in B} O(\boldsymbol{x}_k, \boldsymbol{z}_i) K_h(\tilde{\boldsymbol{x}}_k, \tilde{\boldsymbol{x}}_i)}.
\end{aligned} \tag{5.8}$$

A solution to the score equation $\hat{s}(\boldsymbol{\theta}) = 0$ may be obtained using an algorithm similar to the EM algorithm, where updated values of w_{ij} and $\hat{\boldsymbol{\theta}}$ are produced in an iterative fashion. The EM-like procedure, inspired from the approach described in Riddles et al. (2016), proceeds as follows:

1. Initialise a starting set of values $\boldsymbol{\theta}^{(0)}$
2. Using $\boldsymbol{\theta}^{(t)}$ for step $t \geq 0$, update $O(\boldsymbol{x}_j, \boldsymbol{z}_i)$ and produce new values of w_{ij}
3. Feed updated w_{ij} into the score equation $\hat{s}(\boldsymbol{\theta}) = 0$, and solve (for example, using the Newton-Raphson algorithm with initial values $\boldsymbol{\theta}^{(t)}$) to obtain $\boldsymbol{\theta}^{(t+1)}$

4. Set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t+1)}$. Go back to Step 2 until convergence. Use the final values of $\boldsymbol{\theta}^{(t)}$ as $\hat{\boldsymbol{\theta}}$

$\hat{\pi}_i^B$ can be obtained using the values $\hat{\boldsymbol{\theta}}$, and these can then be applied in (5.1) to produce an estimate of the population mean.

If the reference sample A is a probability sample taken from $\mathcal{U}$, then the score function (5.8) cannot be used. Assuming that A is drawn using Poisson sampling, one can estimate the sum over $\mathcal{U}$ using the survey-weighted data from A, and the sum over B using the survey-weighted data from $A \cap B$. The resulting score function that needs to be solved for is

$$\hat{s}_p(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \delta_i \frac{\pi'(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})}{\pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})} - \sum_{i \in A} d_i^A (1 - \delta_i) \sum_{j \in A \cap B} \pi'(\boldsymbol{x}_j, \boldsymbol{z}_i; \boldsymbol{\theta})(1 - \pi(\boldsymbol{x}_j, \boldsymbol{z}_i; \boldsymbol{\theta}))^{-1} w_{ij} \tag{5.9}$$

$$\text{where } w_{ij} = \frac{d_j^A O(\boldsymbol{x}_j, \boldsymbol{z}_i) K_h(\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{x}}_i)}{\sum_{k \in A \cap B} d_k^A O(\boldsymbol{x}_k, \boldsymbol{z}_i) K_h(\tilde{\boldsymbol{x}}_k, \tilde{\boldsymbol{x}}_i)}$$

Remark 5.1. The assumption of Poisson sampling for A required in (5.9) will hold approximately for more complex sample designs when A has small sampling fractions and joint inclusion probabilities are small. Examples may include stratified sampling where stratum population sizes are large relative to the stratum sample size, two-stage cluster sampling with small numbers of units chosen within each first-stage cluster, or probability proportional to size (PPS) sampling with small sampling fractions.

Remark 5.2. The framework that has been proposed for addressing covariate measurement error in A is also applicable for the case when the selection mechanism in B is SNAR and a Type II data scenario exists where y is only collected in B. Using the proposed framework, it is possible to treat the missing y variable in A as an extreme case of measurement error where the measurement of y is missing. In other words, using the notation in Section 5.2, $\boldsymbol{x}$ now includes the y variable and $\boldsymbol{x}^*$ omits it. To avoid potential model identifiability problems in the SNAR scenario (see for example Shao and Wang 2016 and Wang et al. 2014 in the MNAR literature), it is assumed that the set of auxiliary variables $\boldsymbol{z}$ can be separated into two groups $(\boldsymbol{z}_1, \boldsymbol{z}_2)$, and the participation probability model only depends

on y and the variables in $\boldsymbol{z}_1$, that is, $\pi^B = P(\delta = 1|y, \boldsymbol{z}_1)$. The variables $\boldsymbol{z}_2$ and δ are conditionally independent given $\boldsymbol{z}_1$ and y, though $\boldsymbol{z}_2$ may still be related to y. Again assuming Poisson sampling for A, the kernel score function for the SNAR case is given by

$$\hat{s}_n(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \delta_i \frac{\pi'(\boldsymbol{z}_{1i}, y_i; \boldsymbol{\theta})}{\pi(\boldsymbol{z}_{1i}, y_i; \boldsymbol{\theta})} - \sum_{i \in A} d_i^A (1 - \delta_i) \sum_{j \in A \cap B} \pi'(\boldsymbol{z}_{1i}, y_j; \boldsymbol{\theta})(1 - \pi(\boldsymbol{z}_{1i}, y_j; \boldsymbol{\theta}))^{-1} w_{ij} \tag{5.10}$$

$$\text{where } w_{ij} = \frac{d_j^A O(\boldsymbol{z}_{1i}, y_j) K_h(\boldsymbol{z}_j, \boldsymbol{z}_i)}{\sum_{k \in A \cap B} d_k^A O(\boldsymbol{z}_{1i}, y_k) K_h(\boldsymbol{z}_k, \boldsymbol{z}_i)}$$

where $d_i^A = 1$ and $A \cap B = B$ if A comprises the full population.

Appendix B.7 provides a detailed derivation of $\hat{s}_n$ from the nonignorable missingness perspective.

Remark 5.3. Using a similar reasoning to Remark 5.2, it is possible to apply the proposed framework to the case when some covariates in the participation probability model are missing from the reference sample A. In this case, suppose that the set of auxiliary variables in the participation probability model should include $(\boldsymbol{x}, \boldsymbol{z})$. The auxiliary variables missing from A but are on B are included in $\boldsymbol{x}$, while $\tilde{\boldsymbol{x}}$ now includes only $\boldsymbol{z}$. Model identifiability issues will again need to be considered when applying our framework in this scenario.

When kernel regression is used, a choice needs to be made regarding what kernel function and bandwidth should be used. Generally, the choice of kernel function is not as crucial as the selection of an appropriate bandwidth (see Wand and Jones 1995, for example). A popular choice of kernel function is the Gaussian kernel, while bandwidth selection may be undertaken in a variety of ways, for example using plug-in, bootstrap or cross-validation approaches (Köhler et al. 2014).

In this chapter, I utilise an approach for selecting an optimal kernel bandwidth which is based on an idea first proposed by Morikawa et al. (2017). The approach was described by those authors as a cross-validation approach. The known data $(\boldsymbol{x}_i^*, \boldsymbol{x}_i, \boldsymbol{z}_i)$ taken from

$A \cap B$ is used as training data to fit kernel regressions for the quantities

$$\begin{bmatrix} \pi'(\boldsymbol{x}_i, \boldsymbol{z}_i; \hat{\boldsymbol{\theta}})(1 - \pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \hat{\boldsymbol{\theta}}))^{-1} O(\hat{\boldsymbol{\theta}}; \boldsymbol{x}_i, \boldsymbol{z}_i) \\ O(\hat{\boldsymbol{\theta}}; \boldsymbol{x}_i, \boldsymbol{z}_i) \end{bmatrix}$$

using a set of candidate bandwidths h. The kernel bandwidth which provides the best, in some sense, fit for these quantities over the data $A \cap B$ is chosen to be used in the full set of data to solve the score equation. In this thesis I choose the bandwidth which minimises the overall sum of squared differences between the fitted and true values.

To enunciate the proposed bandwidth selection approach, first define

$$\hat{C}(\hat{\boldsymbol{\theta}}_h; \boldsymbol{z}_i) = (\tilde{n}h^d)^{-1} \sum_{j \in A \cap B} \pi'(\boldsymbol{x}_j, \boldsymbol{z}_i; \hat{\boldsymbol{\theta}}_h)(1 - \pi(\boldsymbol{x}_j, \boldsymbol{z}_i; \hat{\boldsymbol{\theta}}_h))^{-1} O(\hat{\boldsymbol{\theta}}_h; \boldsymbol{x}_j, \boldsymbol{z}_i) K_h(\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{x}}_i) \tag{5.11}$$

$$\hat{D}(\hat{\boldsymbol{\theta}}_h; \boldsymbol{z}_i) = (\tilde{n}h^d)^{-1} \sum_{k \in A \cap B} O(\hat{\boldsymbol{\theta}}_h; \boldsymbol{x}_k, \boldsymbol{z}_i) K_h(\tilde{\boldsymbol{x}}_k, \tilde{\boldsymbol{x}}_i) \tag{5.12}$$

for the overlapping sample $A \cap B$, where $\tilde{n}$ is the sample size feeding into the kernel estimation, h is a candidate bandwidth and d is the dimension of $\tilde{\boldsymbol{x}}$. When A comprises the full population then $A \cap B$ is equivalent to B.

From (5.11) and (5.12), and using (5.7), one can see that

$$\frac{\hat{C}(\hat{\boldsymbol{\theta}}_h; \boldsymbol{z}_i)}{\hat{f}_1(\tilde{\boldsymbol{x}}_i)} = E_1\Big[\pi'(\boldsymbol{x}_i, \boldsymbol{z}_i; \hat{\boldsymbol{\theta}}_h)(1 - \pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \hat{\boldsymbol{\theta}}_h))^{-1} O(\hat{\boldsymbol{\theta}}_h; \boldsymbol{x}_i, \boldsymbol{z}_i) | \tilde{\boldsymbol{x}}_i\Big] \text{ and}$$
$$\frac{\hat{D}(\hat{\boldsymbol{\theta}}_h; \boldsymbol{z}_i)}{\hat{f}_1(\tilde{\boldsymbol{x}}_i)} = E_1\Big[O(\hat{\boldsymbol{\theta}}_h; \boldsymbol{x}_i, \boldsymbol{z}_i) | \tilde{\boldsymbol{x}}_i\Big]$$

where the expectations are kernel regression estimates and $\hat{f}_1(\boldsymbol{x}) = \sum_{i \in A \cap B} K_{h'}(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}})$ for a given bandwidth h'. The bandwidth h' may be chosen using established approaches (see for example Wand and Jones 1995).

The bandwidth selection approach seeks to minimise the overall error in these expectations. Specifically, the following steps are undertaken:

1. For a set of candidate bandwidths $h \in H$, compute $\hat{\boldsymbol{\theta}}_h$, where H is a class of candidates for the bandwidth

2. For each bandwidth candidate $h \in H$, calculate

$$K(h) = \sum_{i \in A} \delta_i \left\| \begin{bmatrix} \hat{C}(\hat{\boldsymbol{\theta}}_h; \boldsymbol{z}_i)/\hat{f}_1(\tilde{\boldsymbol{x}}_i) - \pi'(\boldsymbol{x}_i, \boldsymbol{z}_i; \hat{\boldsymbol{\theta}}_h)(1 - \pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \hat{\boldsymbol{\theta}}_h))^{-1} O(\hat{\boldsymbol{\theta}}_h; \boldsymbol{x}_i, \boldsymbol{z}_i) \\ \hat{D}(\hat{\boldsymbol{\theta}}_h; \boldsymbol{z}_i)/\hat{f}_1(\tilde{\boldsymbol{x}}_i) - O(\hat{\boldsymbol{\theta}}_h; \boldsymbol{x}_i, \boldsymbol{z}_i) \end{bmatrix} \right\|^2$$

   where $\|\cdot\|$ is the Euclidean norm.

3. Choose the bandwidth h^* which minimises $K(h)$

5.4 Consistency and Variance Estimation

This section first shows that under certain regularity conditions the kernel score functions $\hat{s}(\boldsymbol{\theta})$ and $\hat{s}_p(\boldsymbol{\theta})$ converge in probability to the observed score function $s(\boldsymbol{\theta})$. I then provide details of a bootstrap approach to estimate the variance of the IPW estimator when the $\hat{\pi}_i^B$ are obtained using $\hat{\boldsymbol{\theta}}$ estimated from the kernel score function.

To consider consistency, first denote $\partial^j \gamma(\tilde{\boldsymbol{x}})/\partial \tilde{\boldsymbol{x}}^j$ as any vector consisting of all distinct jth-order partial derivatives of all elements of $\gamma(\tilde{\boldsymbol{x}})$, and for some non-negative integer d let

$$\|\gamma\| \equiv \max_{j \leq d} \sup_{\tilde{\boldsymbol{x}} \in \tilde{\mathcal{X}}} \|\partial^j \gamma(\tilde{\boldsymbol{x}})/\partial \tilde{\boldsymbol{x}}^j\|$$

where $\tilde{\mathcal{X}}$ is a set contained in the support of $\tilde{\boldsymbol{x}}$. Newey and McFadden (1994) refer to this as the Sobolev norm. Assume that the regularity conditions in Appendix B.8 and Lemma 8.10 from Chapter 8 of Newey and McFadden (1994) hold. Then for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\|\hat{C}(\boldsymbol{\theta}; \tilde{\boldsymbol{x}}) - C(\boldsymbol{\theta}; \tilde{\boldsymbol{x}})\| = o_p(n^{-1/4}) \text{ and } \|\hat{D}(\boldsymbol{\theta}; \tilde{\boldsymbol{x}}) - D(\boldsymbol{\theta}; \tilde{\boldsymbol{x}})\| = o_p(n^{-1/4}) \tag{5.13}$$

where

$$C(\boldsymbol{\theta};\tilde{\boldsymbol{x}}) = E_1\{\pi'(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})(1-\pi(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta}))^{-1}O(\boldsymbol{\theta};\boldsymbol{x}_i,\boldsymbol{z}_i) \mid \tilde{\boldsymbol{x}}\}f_1(\tilde{\boldsymbol{x}}),$$
$$D(\boldsymbol{\theta};\tilde{\boldsymbol{x}}) = E_1\{O(\boldsymbol{\theta};\boldsymbol{x}_i,\boldsymbol{z}_i) \mid \tilde{\boldsymbol{x}}\}f_1(\tilde{\boldsymbol{x}}),$$

n is the sample size feeding into the kernel estimation, $E_1(\cdot) = E(\cdot|\delta = 1)$ and $\hat{C}(\boldsymbol{\theta};\tilde{\boldsymbol{x}})$ and $\hat{D}(\boldsymbol{\theta};\tilde{\boldsymbol{x}})$ were defined in (5.11) and (5.12). In the following, for ease of notation I will ignore the brackets in $\hat{C}(\boldsymbol{\theta};\tilde{\boldsymbol{x}}), \hat{D}(\boldsymbol{\theta};\tilde{\boldsymbol{x}}), C(\boldsymbol{\theta};\tilde{\boldsymbol{x}})$ and $D(\boldsymbol{\theta};\tilde{\boldsymbol{x}})$.

The second-order Taylor expansion of $\hat{C}/\hat{D}$ around C/D is given by

$$\frac{\hat{C}}{\hat{D}} = \frac{C}{D} + \frac{1}{D}\big(\hat{C}-C\big) - \frac{C}{D^2}\big(\hat{D}-D\big) - \frac{1}{D^2}\big(\hat{C}-C\big)\big(\hat{D}-D\big) + \frac{C}{D^3}\big(\hat{D}-D\big)^2$$

Now,

$$\left\| \frac{1}{D^2}\big(\hat{C}-C\big)\big(\hat{D}-D\big) + \frac{C}{D^3}\big(\hat{D}-D\big)^2 \right\| = o_p(n^{-1/2})$$

since from (5.13), $(\hat{C}-C)(\hat{D}-D)$ and $(\hat{D}-D)^2$ are both $o_p(n^{-1/2})$. Re-arranging the Taylor expansion thus gives

$$\left\| \frac{\hat{C}}{\hat{D}} - \frac{C}{D} - \frac{1}{D}\big(\hat{C}-C\big) + \frac{C}{D^2}\big(\hat{D}-D\big) \right\| = o_p(n^{-1/2}) \tag{5.14}$$

Now,

$$\left\| \frac{\hat{C}}{\hat{D}} - \frac{C}{D} \right\| = \left\| \frac{\hat{C}}{\hat{D}} - \frac{C}{D} - \frac{1}{D}\big(\hat{C}-C\big) + \frac{C}{D^2}\big(\hat{D}-D\big) + \frac{1}{D}\big(\hat{C}-C\big) - \frac{C}{D^2}\big(\hat{D}-D\big) \right\|$$

and so by the triangle inequality,

$$\left\| \frac{\hat{C}}{\hat{D}} - \frac{C}{D} \right\| \leq \left\| \frac{\hat{C}}{\hat{D}} - \frac{C}{D} - \frac{1}{D}\big(\hat{C}-C\big) + \frac{C}{D^2}\big(\hat{D}-D\big) \right\| + \left\| \frac{1}{D}\big(\hat{C}-C\big) \right\| + \left\| \frac{C}{D^2}\big(\hat{D}-D\big) \right\|.$$

Given that the first term on the right is $o_p(n^{-1/2})$ and the second and third terms on the

right are $o_p(n^{-1/4})$,

$$\left\| \frac{\hat{C}}{\hat{D}} - \frac{C}{D} \right\| = o_p(n^{-1/4}). \tag{5.15}$$

For the score functions $\hat{s}(\boldsymbol{\theta})$ and $s(\boldsymbol{\theta})$, adopting the framework of Section 3.3, we examine the mean norm:

$$\begin{aligned} \frac{1}{N} \|\hat{s}(\boldsymbol{\theta}) - s(\boldsymbol{\theta})\| &= \frac{1}{N} \left\| \sum_{i \in \mathcal{U}} (1 - \delta_i) \Big(\frac{\hat{C}}{\hat{D}} - \frac{C}{D} \Big) \right\| \\ &\leq \frac{1}{N} \sum_{i \in \mathcal{U}} (1 - \delta_i) \left\| \frac{\hat{C}}{\hat{D}} - \frac{C}{D} \right\| \qquad (5.16) \\ &\leq o_p(N^{-1/4}) \qquad (5.17) \end{aligned}$$

and so, with D bounded away from 0, $N^{-1} \|\hat{s}(\boldsymbol{\theta}) - s(\boldsymbol{\theta})\|$ converges in probability to 0 with a rate at least as fast as $N^{-1/4}$.

To consider convergence of $\|\hat{s}_p(\boldsymbol{\theta}) - s(\boldsymbol{\theta})\|$, the asymptotic framework of Section 3.3 is again adopted. Further assume that the conditions **C13** to **C15** in Appendix B.8 hold for the probability sample A. I specifically consider the standardised norm, $N^{-1} \|\hat{s}_p(\boldsymbol{\theta}) - s(\boldsymbol{\theta})\|$.

Using the triangle inequality,

$$\begin{aligned} \frac{1}{N} \|\hat{s}_p(\boldsymbol{\theta}) - s(\boldsymbol{\theta})\| &= \frac{1}{N} \| \Big(\hat{s}_p(\boldsymbol{\theta}) - \hat{s}(\boldsymbol{\theta}) \Big) + \Big(\hat{s}(\boldsymbol{\theta}) - s(\boldsymbol{\theta}) \Big) \| \\ &\leq \frac{1}{N} \|\hat{s}_p(\boldsymbol{\theta}) - \hat{s}(\boldsymbol{\theta})\| + \frac{1}{N} \|\hat{s}(\boldsymbol{\theta}) - s(\boldsymbol{\theta})\|. \end{aligned}$$

The first term on the right hand side relates to the use of a probability sample rather than the full population. The second term on the right hand side relates to the use of kernel regression to approximate the integrals in the score function. From **C13**, the sampling fractions n_A/N are bounded, and so the second term converges at least as fast as $o_p(n_A^{-1/4})$ by (5.17). When conditions **C16** and **C17** hold for the sample design of A, the first term can be shown to be $O_p(n^{-1/2})$ (see Appendix B.9 for a sketched proof), giving

$$\frac{1}{N} \|\hat{s}_p(\boldsymbol{\theta}) - s(\boldsymbol{\theta})\| \leq o_p(n^{-1/4}).$$

5.4.1 Pseudo-population bootstrap for variance estimation

To estimate the variance of $\hat{\mu}_y$, a pseudo-population bootstrap procedure is proposed based on the idea of Booth et al. (1994); see also Mashreghi et al. (2016) for a review of bootstrap methods for finite populations. The advantage of the pseudo-population bootstrap approach is that it allows one to conveniently take into account the overlapping sample in the calculation of bootstrap estimates. Liu et al. (2022) employed a similar bootstrap approach to take into account the possible dependency between the non-probability and reference samples.

The bootstrap procedure described in this subsection involves creating a pseudo-population using the data from B. This means that data about $(y, \boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{z})$ is needed for all $i \in B$. When A comprises the full population this requirement will be fulfilled since values for $\boldsymbol{x}^*$ will be available for all units in the population. Similarly, when B is subject to SNAR missingness it will contain all the required data items since in this scenario $\boldsymbol{x}^*$ omits the variable of interest y (recall Remark 5.2).

In the case when there is measurement error in the auxiliary variables in A and A is a sample, $\boldsymbol{x}^*$ is only available for the records in A. In this thesis, I propose to produce imputed values $\hat{\boldsymbol{x}}_i^*$ for $i \in B \setminus A$ to fill the gap in the data for the pseudo-population bootstrap. The imputed values may be obtained by first fitting a model such as (1.1) using data in $A \cap B$. The fitted model, including the normally distributed random error term to enable a stochastic imputation, can then be applied to the records in B to obtain $\hat{\boldsymbol{x}}_i^*, i \in B \setminus A$. The augmented sample dataset B, which now includes $\hat{\boldsymbol{x}}_i^*$ for $i \in B \setminus A$ and $\boldsymbol{x}_i^*$ for $i \in B \cap A$, is used to form the pseudo-population.

The pseudo-population bootstrap proceeds as follows. Once any required imputation to obtain $\hat{\boldsymbol{x}}^*$ is undertaken, a pseudo-population of size N is then produced using the units in B. Like Liu et al. (2022), the $\hat{\pi}_i^B$ are used as the basis for expanding the sample B. The weights $1/\hat{\pi}_i^B$ are first adjusted to ensure they sum to N, and any adjusted weights less than 1 are then set to 1, creating $\tilde{w}_i = \max(1, N(\hat{\pi}_i^B)^{-1} / \sum_{i \in B}(\hat{\pi}_i^B)^{-1})$. The fixed part of the pseudo-population $\mathcal{U}_f$ is then created by replicating each unit i of the original B

sample $k = \lfloor \tilde{w}_i \rfloor$ times. The pseudo-population is completed by drawing $\mathcal{U}_c$ from B using Poisson sampling with probability $\tilde{w}_i - \lfloor \tilde{w}_i \rfloor$ for unit i. The final pseudo-population from which bootstrap samples are taken is $\mathcal{U}^* = \mathcal{U}_c \cup \mathcal{U}_f$.

Algorithm 2 Bootstrap variance estimation for proposed IPW estimator

Require: $A, B, \hat{\pi}_i^B, n_A, N_B, R$

1: Set $r := 1$
2: If required, produce $\hat{\boldsymbol{x}}_i^*, i \in B \setminus A$ using a model fitted using the records in $A \cap B$
3: Produce $\tilde{w}_i = \max(1, N(\hat{\pi}_i^B)^{-1} / \sum_{i \in B}(\hat{\pi}_i^B)^{-1})$
4: Create the fixed part of the pseudo-population $\mathcal{U}_f$ by replicating each unit i of the original B sample $k = \lfloor \tilde{w}_i \rfloor$ times
5: Create $\mathcal{U}_c$ by drawing a sample from B using Poisson sampling with probability $\tilde{w}_i - \lfloor \tilde{w}_i \rfloor$ for unit i. The final pseudo-population is given by combining $\mathcal{U}_f$ and $\mathcal{U}_c$ to form $\mathcal{U}^*$
6: **while** $r \leq R$ **do**
7: Select a sample S_A from $\mathcal{U}^*$ with size n_A, using the original sample design of A
8: Select a sample S_B from $\mathcal{U}^*$ with expected size N_B
9: Calculate $\hat{\boldsymbol{\theta}}^r$ using S_A and S_B using the method outlined in Section 5.3
10: Calculate $\hat{\pi}_i^{Br}$ for each unit $i \in S_B$ using $\hat{\boldsymbol{\theta}}^r$
11: Calculate $\hat{\mu}_y^{(r)}$ using $\hat{\pi}_i^{Br}$
12: $r := r + 1$
13: **end while**
14: Calculate the bootstrap variance using (5.18)

For each bootstrap iteration, bootstrap non-probability samples S_B may be drawn from $\mathcal{U}^*$ using Poisson sampling with inclusion probability $\hat{\pi}_i^B$. An alternative sampling scheme such as a fixed-size PPS sampling scheme may be used based on knowledge about the selection mechanism into B. The bootstrap probability samples S_A are drawn from $\mathcal{U}^*$ based on the sample design of the probability sample, with probabilities of selection π_i^A. The samples S_A and S_B are combined using the proposed method to obtain $\hat{\pi}_i^{Br}$ and hence an estimate of the population quantity of interest, for example population means $\hat{\mu}_y^{(r)}$. The bootstrap sampling procedure is repeated R times to obtain $(\hat{\mu}_y^{(1)}, ..., \hat{\mu}_y^{(R)})$. The bootstrap variance is obtained through

$$\hat{Var}(\hat{\mu}_y) = \frac{1}{R-1} \sum_{r=1}^{R} \left(\hat{\mu}_y^{(r)} - \mu_y^* \right)^2 \tag{5.18}$$

where $\mu_y^* = N^{-1} \sum_{i \in \mathcal{U}^*} y_i$ is the mean of the y variable in the pseudo-population (see Mashreghi et al. 2016).

Algorithm 2 outlines the bootstrap variance procedure.

5.5 Simulation Study

This section describes the implementation and results of two simulation studies to assess the performance of the proposed approach and its bootstrap variance estimator. The first study was undertaken to assess the effectiveness of the approach to correct measurement error in the auxiliary variables in A, while the second assessed its use to address SNAR in the non-probability sample.

5.5.1 Measurement error simulation study

A finite population of size $N = 20,000$ was constructed comprising a data item of interest y and auxiliary variables $\boldsymbol{x}$. The variable y was related to $\boldsymbol{x}$ by the model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma\epsilon_i$$

where $\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, x_{3i}, x_{4i})$, $\boldsymbol{\beta} = (2, 1.2, 1, 1, 1)$, $x_{1i} = z_{1i}, x_{2i} = z_{2i} + 0.3x_{1i}, x_{3i} = z_{3i} + 0.2(x_{1i} + x_{2i}), x_{4i} = z_{4i} + 0.1(x_{1i} + x_{2i} + x_{3i})$, and $z_{1i} \sim$ Bernoulli$(0.5), z_{2i} \sim$ Uniform$(0, 2), z_{3i} \sim$ Exponential(1), and $z_{4i} \sim \chi^2(4)$. The error terms ϵ_i were independent and identically distributed as Normal$(0, 1)$. For these simulations, σ was set such that the correlation coefficient ρ between y and $\boldsymbol{x}^T\boldsymbol{\beta}$ was 0.3. The data item x_4 may be subject to measurement error. The performance of the proposed method was examined under four slightly different measurement error models:

1. $x_4^* = x_4 + \gamma$, where $\gamma \sim$ Normal$(0, 1)$
2. $x_4^* = 0.8x_4 + \gamma$, where $\gamma \sim$ Normal$(0, 1)$
3. $x_4^* = x_4 + \gamma$, where $\gamma \sim$ Normal$(0, 2)$
4. $x_4^* = 10x_4^{1/4} + \gamma$, where $\gamma \sim$ Normal$(0, 0.7)$

Case 1 relates to the classical measurement error scenario where the mis-measured variable is given by the true value of the variable plus some random error. In Case 2, the mis-measured variable tends to have a lower value compared with the true value of the variable. Case 3 is similar to Case 1, but the error term has a larger standard deviation. Finally, Case 4 represents a non-linear measurement error model.

The true participation probabilities into the non-probability sample, π_i^B, were given by

$$\log\Big(\frac{\pi_i^B}{1-\pi_i^B}\Big) = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{3i} + \theta_4 x_{4i}$$

where the intercept θ_0 was chosen to ensure the target sample size for N_B is achieved, and $(\theta_1, \theta_2, \theta_3, \theta_4) = (0.1, 0.2, 0.3, 0.4)$. Poisson sampling was used to select the sample B. Pareto probability proportional to size (PPS) sampling was used to select the probability sample A. In general, the inclusion probabilities for π_i^A were proportional to $z_i = c + x_{3i}$, where c was chosen to ensure the ratio $\max(z_i)/\min(z_i) = 20$.

For each of the measurement error scenarios, $R = 2,000$ simulation iterations were executed, with 50 bootstrap iterations used to calculate a bootstrap variance using the approach outlined in Algorithm 2. This number of bootstrap iterations was found to yield stable variance estimates. Since in this simulation study the samples A do not comprise the full population, imputed values $\hat{x}_{4,i}^*, i \in B \setminus A$ were produced for the bootstrap pseudo-population using a linear model fitted on the common sample $A \cap B$ in each simulation iteration:

$$\hat{x}_{4,i}^* = \hat{\beta}_0 + \hat{\beta}_1 x_{4,i} + \hat{\gamma} \tag{5.19}$$

where $\hat{\beta}_0$, $\hat{\beta}_1$ were the fitted model parameters and $\hat{\gamma}$ was a random normally distributed residual error term with mean 0 and estimated variance obtained from the model fitting process.

A Gaussian kernel was used to implement the kernel smoothing for the proposed approach. An optimal kernel bandwidth for $\hat{C}$ and $\hat{D}$ was chosen using the bandwidth selection method described in Section 5.3. The candidate bandwidths used were of a form based on

the data-based rule in Scott (2014):

$$h(p) = n^{-p}\hat{\sigma}_x \tag{5.20}$$

where n was the sample size feeding in to the kernel estimation, $\hat{\sigma}_x$ was the estimated standard deviation of the variable x, and p was allowed to range from 0.1 to 0.5. To estimate the densities $\hat{f}_1(\tilde{\boldsymbol{x}})$ used in the bandwidth selection approach, kernel smoothing was again utilised, with a Gaussian kernel and bandwidths $n^{-0.4}\hat{\sigma}_x$.

For the simulation exercise, the performance of the proposed approach was assessed for values of N_B of 2,000 and 8,000. The reference sample size n_A was kept at a constant 1,000. Results were produced for the following estimators:

- Naive - Simple sample mean from B
- OE - The Overlap-Excluded (OE) IPW estimator from Chapter 3 without any correction for measurement error
- OE_Model - The OE IPW estimator where measurement error is corrected by fitting a model $x_4^* = \hat{\beta}_0 + \hat{\beta}_1 x_4$ using the common sample $A \cap B$, and then applying it to the sample A to estimate $\hat{x}_4$
- ME - The proposed method for addressing measurement error in some auxiliary variables in A
- OE_True - The OE IPW estimator when there is no measurement error for x_4 in A

The performance of the five estimators was evaluated by calculating the Monte Carlo Relative Bias (RB) and Relative Root Mean Squared Error (RRMSE):

$$\text{RB} = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\mu}_{y,r} - \mu_y}{\mu_y}, \ \text{RRMSE} = \frac{\sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\mu}_{y,r} - \bar{\hat{\mu}}_y)^2 + (\bar{\hat{\mu}}_y - \mu_y)^2}}{\mu_y}$$

where $\hat{\mu}_{y,r}$ is the estimate of the population mean computed for the r'th simulation run, $\bar{\hat{\mu}}_y = 1/R\sum_{r=1}^{R}\hat{\mu}_{y,r}$ and μ_y is the true population mean.

Table 5.1 compares the RB and RRMSE of the estimators for the classical measurement error case, Case 1. The OE_True results represent the benchmark performance under an ideal case where there is no covariate measurement error, and they are unbiased with the lowest RRMSE. The OE and OE_Model estimates have a small bias as a result of attenuation bias in the participation probability model parameters. On the other hand, the resulting estimates of mean from our proposed method are unbiased and have comparable RRMSE values to the true model.

Table 5.1: Comparison of estimators under measurement error - Case 1

	$N_B = 2,000$		$N_B = 8,000$	
	%RB	%RRMSE	%RB	%RRMSE
Naive	42.67	42.73	19.30	19.32
OE	1.55	4.53	1.08	1.84
OE_Model	1.56	4.57	1.09	1.85
ME	-0.13	4.57	0.09	1.59
OE_True	0.00	4.40	-0.01	1.53

In this measurement error scenario, creating a model to correct measurement error does not help at all. To see why this is the case, recall that in the classical measurement error scenario the mis-measured value of a variable is equal to the true value plus some random error. The fitted model would thus result in predicted values that are approximately equal to the original mis-measured values. This is why the OE_Model results are very similar to the OE results. Note that it may seem reasonable to instead employ a measurement error model which includes a fitted error term, for example $x_4^* = \hat{\beta}_0 + \hat{\beta}_1 x_4 + \hat{\sigma}$ where $\hat{\sigma}$ is estimated from the model fitting process. However, in the simulation study this approach was found to actually lead to worse results.

Table 5.2 compares the RB and RRMSE of the various estimators under measurement error for Case 2. The OE estimates are again biased and have a fairly large RRMSE because they have not corrected for measurement error. Fitting a measurement error model, imputing a corrected $\hat{x}_4$ value, and using that in the participation probability model is able to correct much, but not all, of the measurement error bias. The RRMSE results for the measurement error model approach are better than if no action had been taken at all. The proposed approach to correct measurement error is shown to be effective. The estimates produced by the approach are unbiased. In particular, the RRMSE results

for the proposed approach are only slightly higher than the benchmark results for no measurement error.

Table 5.2: Comparison of estimators under measurement error - Case 2

	$N_B = 2,000$		$N_B = 8,000$	
	%RB	%RRMSE	%RB	%RRMSE
Naive	42.67	42.73	19.30	19.32
OE	-7.18	8.92	-2.97	3.41
OE_Model	2.39	4.88	1.63	2.20
ME	0.10	4.64	0.23	1.63
OE_True	0.00	4.40	-0.01	1.53

Table 5.3 provides the RB and RRMSE values of the estimators under Case 3. In this case the mis-measured variable is equal to the true value plus an error that has a larger standard deviation than Case 1. From these results, the OE and OE_Model approaches again produce biased estimates with fairly large RRMSE values. Like Case 1, fitting a model to correct the measurement error does not help with reducing the bias since the predicted true values are approximately equal to the original mis-measured values.

Table 5.3: Comparison of estimators under measurement error - Case 3

	$N_B = 2,000$		$N_B = 8,000$	
	%RB	%RRMSE	%RB	%RRMSE
Naive	42.67	42.73	19.30	19.32
OE	5.66	6.88	3.49	3.76
OE_Model	5.71	7.02	3.51	3.79
ME	2.49	5.31	1.41	2.21
OE_True	0.00	4.40	-0.01	1.53

By including the measurement error correction as part of the likelihood maximisation process, the proposed approach mitigates the bias, and it also has a reduced RRMSE compared to the OE and OE_Model approaches. However, it does not completely remove the bias. On further inspection, the inability to remove all bias seems to stem from the small overlapping sample between A and B. When $N_B = 2,000$ and $n_A = 1,000$, for example, the overlapping sample is around 120. When we used $N_B = 2,000$ and $n_A = 4,000$ the overlapping sample increased to about 500 units, and the RB for our proposed approach dropped to 1.7%. When there is a lot of noise in the measurement error, the proposed approach requires a larger overlapping sample to achieve an accurate result.

Finally, Table 5.4 shows the RB and RRMSE values of the estimators under Case 4. In this scenario, the OE approach severely underestimates the participation probabilities, leading to very large RB and RRMSE results. Fitting a linear measurement error model to the common units and using the predicted true values of x_4 reduces the bias and RRMSE substantially, but does not eliminate it. In contrast, the proposed approach is unbiased and has RRMSE values very close to the gold standard OE_True model.

Table 5.4: Comparison of estimators under measurement error - Case 4

	$N_B = 2,000$		$N_B = 8,000$	
	%RB	%RRMSE	%RB	%RRMSE
Naive	42.67	42.73	19.30	19.32
OE	243.33	243.42	263.83	263.88
OE_Model	6.64	7.71	3.28	3.55
ME	0.08	4.66	0.03	1.59
OE_True	0.00	4.40	-0.01	1.53

The performance of the bootstrap variance estimator described in Section 5.4 is assessed in Table 5.5 for each of the four scenarios. The table contains the coverage probabilities (CP) and average length (AL) of the 95% confidence intervals for the estimates of population mean for two implementations of the ME bootstrap. The first implementation assumes x_4^* is available in the full population, and is used as a benchmark to assess the second implementation which assumes x_4^* is only available in A. The variance estimates for the OE approach are also included for comparison.

It is first noted that in all scenarios the CP and AL values for the ME bootstrap when x_4^* is only available in A are very close to the CP and AL values for the ME boostrap when x_4^* is available for the full population. The imputation used for x_4^* is effective, especially for Case 1, Case 2 and Case 3 where a linear relationship between x_4 and x_4^* exists. In Case 4, where there is a non-linear relationship between x_4 and x_4^*, the imputation leads to a slight over-estimation of the variance leading to slightly higher CP and AL values. This is understandable since from (5.19) the imputation of x_4^* assumes a linear relationship with x_4. The bootstrap variance results could be improved, for example by employing an imputation model more suited to the non-linear measurement error. In the discussion that follows, the focus will be on the ME bootstrap results where x_4^* is assumed to be available only on A.

Table 5.5: Performance of variance estimators under measurement error

	$N_B = 2,000$		$N_B = 8,000$	
	CP	AL	CP	AL
Case 1				
OE	0.924	1.559	0.884	0.554
ME Bootstrap (x_4^* in population)	0.954	1.757	0.950	0.599
ME Bootstrap (x_4^* in A only)	0.957	1.764	0.949	0.599
Case 2				
OE	0.773	2.048	0.584	0.624
ME Bootstrap (x_4^* in population)	0.950	1.764	0.946	0.610
ME Bootstrap (x_4^* in A only)	0.954	1.776	0.948	0.611
Case 3				
OE	0.645	1.361	0.315	0.518
ME Bootstrap (x_4^* in population)	0.918	1.873	0.922	0.766
ME Bootstrap (x_4^* in A only)	0.917	1.881	0.920	0.765
Case 4				
OE	1.000	496.1	1.000	1130.9
ME Bootstrap (x_4^* in population)	0.951	1.765	0.950	0.596
ME Bootstrap (x_4^* in A only)	0.959	1.811	0.961	0.622

Note: CP are the estimated Coverage Probabilities, while AL denotes the average length of the 95% confidence intervals

For Case 1 and Case 2, the CP values for the ME bootstrap variance estimator under both sample sizes for N_B are very close to the nominal value of 0.95. For Case 3, the CP of the bootstrap estimator is still around 0.92; it is slightly lower than the nominal value of 0.95 due to the residual bias in the estimates. In Case 4, as has been noted, the CP values are slightly higher than the nominal value given that the imputation of x_4^* assumes a different measurement error model compared with reality.

The CPs for the OE estimator, which does not have any correction for measurement error, are below the expected 0.95 in Case 1, Case 2 and Case 3. For Case 1, the CPs are still reasonable, hovering at around 0.9. However, for Case 2 and Case 3 they are all far away from 0.95. In Case 4, the CPs are 1.0 because the variance estimate is so large as to be unusable. These results provide further evidence that care is needed to ensure the quality of the covariate information when making inferences using non-probability data combined with a probability sample. Even a small error, when not addressed, may lead to erroneous inferences.

The ALs for the ME bootstrap variance estimator are all reasonable in magnitude relative

to the population mean of around 9.4. In Case 1 and Case 3, the AL for the OE estimator is slightly smaller compared with the AL for the ME bootstrap approach because the OE variance has been under-estimated. For the non-linear measurement error Case 4, the AL values for the OE variance estimator are very much higher compared with the ME bootstrap variance estimator.

5.5.2 SNAR simulation study

The setup for the SNAR simulation study is similar to the setup used in Kim and Morikawa (2023). A population of size $N = 5,000$ was first created. For the population, the outcome model

$$y_i = -4 + x_{1i} + x_{2i} + e_i$$

was used where $x_{1i}, x_{2i} \sim N(2, v)$, $e_i \sim N(0, 1)$, and v took the values 0.3 or 1.

Selection into the non-probability sample is governed by $\delta_i \sim \text{Bernoulli}(\pi_i^B)$, with selection probabilities π_i^B given by

$$\log\Big(\frac{\pi_i^B}{1-\pi_i^B}\Big) = \theta_0 + \theta_1 x_{1i} + \theta_2 y_i$$

where $(\theta_0, \theta_1, \theta_2) = (-2, 1, 0.5)$. In particular, the selection probabilities π_i^B are greater for larger values of the variable of interest y_i. The probability sample A was selected using Pareto PPS sampling. The inclusion probabilities π_i^A were proportional to $z_i = c + x_{1i}$, where c was chosen to ensure the ratio $\max(z_i)/\min(z_i) = 30$.

For this simulation, repeated samples were drawn $R = 1,000$ times, with $B = 50$ bootstrap iterations used to calculate the bootstrap variance. This number of bootstrap iterations was found to yield stable variance estimates. As for the measurement error study, a Gaussian kernel was applied for the kernel smoothing, with optimal kernel bandwidths chosen using the bandwidth selection method described in Section 5.3. Candidate bandwidths were of the same form as (5.20). The densities $\hat{f}_1(\tilde{\boldsymbol{x}})$ were estimated using the Gaussian

kernel and bandwidths $n^{-0.3}\hat{\sigma}_x$.

The sample size n_A for the reference sample was 1,000, while N_B was set at 2,500. The performance of the proposed approach was assessed for the cases when the standard deviation of x_1 took values of 0.3 or 1. Results were produced for the following estimators of population mean:

- Naive - Simple sample mean from B
- OE - The Overlap-Excluded (OE) IPW estimator from Chapter 3 which includes the $\boldsymbol{x}$ variables in the participation probability model
- SNAR - The proposed method applied to address SNAR in B
- SNAR_Cal - The proposed method for addressing SNAR, followed by a calibration weighting step to calibrate to N and population totals (N, X_1, X_2), where X_1 and X_2 are the population totals for those variables
- OE_True - The OE IPW estimator assuming we know the participation probability model

The performance of the five estimators was evaluated by examining the Monte Carlo bias, variance and Mean Squared Error (MSE) values:

$$\text{Bias} = \frac{1}{R}\sum_{r=1}^{R}\hat{\mu}_{y,r} - \mu_y, \ \text{Variance} = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{\mu}_{y,r} - \bar{\hat{\mu}}_y)^2, \ \text{MSE} = \text{Variance} \ + \ \text{Bias}^2$$

where $\hat{\mu}_{y,r}$ is the estimate of the population mean computed for the r'th simulation run, $\bar{\hat{\mu}}_y = 1/R\sum_{r=1}^{R}\hat{\mu}_{y,r}$ and μ_y is the true population mean. Note that these metrics are slightly different to the metrics used in the preceding section. The reason is that in this simulation study the population mean is very close to 0, and so using a relative bias or relative root MSE metric is not as appropriate.

Table 5.6 summarises the simulation results for the various estimators. The incorrectly specified OE model assumes the missingness in B follows a SAR mechanism, and so yields biased estimates of the population mean along with high MSE values. The SNAR

estimates resulting from the proposed approach mitigate the selection bias, and also result in a much lower MSE. In the case when the x_1 variable has lower variability, the bias is well below the level of rounding. When the variability in x_1 is larger, the bias is almost but not completely removed. In this case, applying the calibration step helps to reduce selection bias a little further, and also has the added benefit of improving the efficiency of the estimates. When $v = 1$, the MSE from the calibrated estimates is nearly half the value of the OE_True MSE which comes from the correctly specified model but isn't calibrated.

Table 5.6: Comparison of estimators under SNAR

	$v = 0.3$			$v = 1$		
	Bias	Var(x1000)	MSE(x1000)	Bias	Var(x1000)	MSE(x1000)
Naive	0.49	0.32	243.70	0.84	0.32	713.70
OE	0.25	0.69	60.85	0.25	3.62	66.87
SNAR	0.00	5.71	5.71	0.04	17.17	18.84
SNAR_Cal	0.00	1.85	1.85	0.03	3.62	4.77
OE_True	0.00	0.97	0.97	0.00	8.85	8.85

Note: v represents the standard deviation of the variable x_1

The performance of the bootstrap variance estimator in the simulation is shown in Table 5.7. Like Table 5.5, the CP and AL of the 95% confidence intervals for the estimates of population mean are examined, with the variance estimates for the OE approach also included. The CP values are close to the nominal value of 0.95, albeit with a slight over-coverage. This is because the bootstrap variance estimator appears to be slightly over-estimating the variance.

Table 5.7: Performance of variance estimators under SNAR

	$v = 0.3$		$v = 1$	
	CP	AL	CP	AL
OE	0.000	0.110	0.068	0.228
SNAR Bootstrap	0.960	0.318	0.959	0.527

Note: CP are the estimated Coverage Probabilities, while AL denotes the average length of the 95% confidence intervals

5.6 Real-data Application

In this section the performance of the proposed framework is examined using some real business data. The population dataset created in Chapter 3 using July-September 2022-

2023 quarter BAS data was re-used for this exercise. Recall that the BAS data was sourced using BLADE and accessed through the ABS DataLab. Records designated as live businesses, in the private or public sector, and with an active Goods and Services Tax (GST) payer role, were extracted. Pay-As-You-Go (PAYG) data relating to the number of employees in each business was also obtained for the 2022-2023 financial year. Due to computational considerations, especially with respect to the calculation of the kernel densities, this empirical study was restricted to businesses in the Manufacturing industry division. The resulting population consisted of 37,218 units.

Total sales excluding GST (turnover) was again used as the variable of interest. The auxiliary data items were total salary, wages and other payments (wages) and an employee count (employees) for the business. A version of the wages data item containing measurement error was created by merging onto the population file the wages value for each business, if available, from the July-September 2020-2021 quarter BAS data. This represented an out-of-date frame variable. If a business did not have any wages data from that quarter, the 2022-2023 quarter BAS data was assigned to it. In other words, when the 2020-2021 data was not available, the business was assumed to have no measurement error. Since they were very right-skewed variables, the with- and without-measurement error wages variables were both transformed using a log transformation before being used. To avoid issues with using a log transformation when wages is zero, the specific transformation was to first add 1 to the wages value and then apply the log transformation.

Figure 5.1 is a binned scatterplot which provides an indication of the relationship between true wages and wages with measurement error in the population. For confidentiality reasons, only bins/tiles with 10 or more contributors are included in the plot. In general, there is a good linear relationship between the two variables, albeit with some variability around the general trend. In addition, there are some units in the population for which one of the wages variables has a value of zero while the other wages variable is non-zero.

The performance of the proposed approach for addressing covariate measurement error in A was assessed through a simulation study using the data from this population. For each iteration of the simulation, a non-probability dataset was selected using Poisson

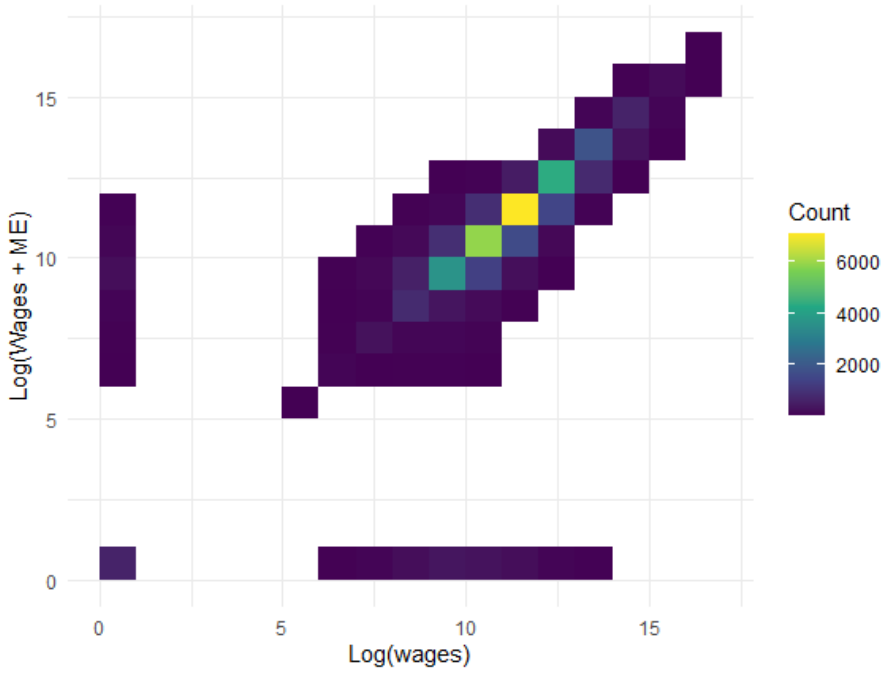

Figure 5.1: Binned scatterplot of Log(wages) vs Log(wages with ME)
Note: Data sourced from BLADE using BAS data from Quarter 1 2022-2023 and BAS data from Quarter 1 2020-2021, accessed through the ABS DataLab August 2025.

sampling. Selection into the non-probability dataset was driven by a logistic model with log(wages+1) and employees as the covariates:

$$\log\Big(\frac{\pi_i^B}{1-\pi_i^B}\Big) = \theta_0 + 0.05x_{\text{employees}} + 0.1x_{\log(\text{wages})}$$

where θ_0 was chosen so that the expected non-probability sample size $N_B = 3,000$. The data items available on this dataset were turnover, employees and true wages.

A reference probability sample was chosen using PPS sampling, with the log of the measurement-error version of wages used as the measure of size. For businesses where this variable had a value greater than 1.5, the value was set to 1.5. The inclusion probabilities π_i^A were proportional to $z_i = c + \log(\text{wages}_{me,i})$, where c was chosen to ensure the ratio $\max(z_i)/\min(z_i)$ was equal to 20 and $\text{wages}_{me,i}$ was the two-year-old wages value for unit i exhibiting measurement error. Using the logarithm of wages rather than the original wages value to form inclusion probabilities helped to make the size measure more symmetric and reduced the influence of more extreme units. For each simulation iteration,

a probability sample of size $n_A = 5,000$ was chosen using Pareto sampling. The data items available on the reference samples were employees and the measurement-error version of wages.

For this simulation using the real-world data, $R = 1,000$ iterations were run. The performance of the following estimators was assessed:

- Naive - Simple sample mean from B
- OE - The Overlap-Excluded (OE) IPW estimator from Chapter 3 without any correction for measurement error
- ME - The proposed method for addressing measurement error in A, where kernel bandwidths were chosen using the bandwidth selection method described in Section 5.3 using candidate bandwidths of the form (5.20), and $\hat{f}_1(\tilde{\boldsymbol{x}})$ estimated using kernel smoothing with a Gaussian kernel and bandwidths $n^{-0.4}\hat{\sigma}_x$
- OE_True - The OE IPW estimator assuming that there is no measurement error for wages in A

Table 5.8 shows the RB and RRMSE values for the estimators. The Naive estimate is severely biased and has a very large RRMSE. The OE estimator corrects for the selection bias, but because the wages variable in A is prone to measurement error, the resulting estimate still shows a large bias and RRMSE. The proposed approach removes most of the bias, although there is still a small residual bias remaining. Its RRMSE is also much smaller than the RRMSE for the OE method, although about 3 percentage points larger than the gold standard OE_True estimator.

Table 5.8: Performance of proposed method using real-world data

	%RB	%RRMSE
Naive	505.57	505.63
OE	-14.54	15.59
ME	1.44	7.00
OE_True	0.02	4.20

5.7 Covariate measurement error in B

The framework that has been proposed in this chapter has focussed on the case where there is measurement error in some of the auxiliary variables present on the probability sample A. This section briefly considers the opposite situation where there is measurement error in some of the auxiliary variables present on the non-probability dataset B. Again, assume that A corresponds to a population-level dataset, and construct the observed likelihood function for π_i^B as

$$L(\boldsymbol{\theta}) = \prod_{i \in \mathcal{U}} \Big\{ \int \pi(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta}) f(\boldsymbol{x} | \boldsymbol{x}_i^*, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{x} \Big\}^{\delta_i} (1 - \pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}))^{1-\delta_i} \tag{5.21}$$

The corresponding log-likelihood is

$$l(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \delta_i \log \Big\{ \int \pi(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta}) f(\boldsymbol{x} | \boldsymbol{x}_i^*, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{x} \Big\} + \sum_{i \in \mathcal{U}} (1 - \delta_i) \log(1 - \pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})) \tag{5.22}$$

and, under mild conditions enabling the interchange of differentiation and integration, the score function is given by

$$\begin{aligned} s(\boldsymbol{\theta}) &= \frac{\partial l}{\partial \boldsymbol{\theta}} \\ &= \sum_{i \in \mathcal{U}} \delta_i \frac{\int \pi'(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta}) f(\boldsymbol{x} | \boldsymbol{x}_i^*, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{x}}{\int \pi(\boldsymbol{x}, \boldsymbol{z}_i; \boldsymbol{\theta}) f(\boldsymbol{x} | \boldsymbol{x}_i^*, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{x}} - \sum_{i \in \mathcal{U}} (1 - \delta_i) \frac{\pi'(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta})}{(1 - \pi(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}))} \end{aligned} \tag{5.23}$$

The density $f(\boldsymbol{x} | \boldsymbol{x}^*, \boldsymbol{z})$ is not known, but the relationship (5.5) can be used to reformulate

the score function as

$$\begin{aligned} s(\boldsymbol{\theta}) &= \sum_{i\in\mathcal{U}} \delta_i \frac{\int \pi'(\boldsymbol{x},\boldsymbol{z}_i;\boldsymbol{\theta}) f_1(\boldsymbol{x}|\boldsymbol{x}_i^*,\boldsymbol{z}_i) P(\delta_i=1|\boldsymbol{x}_i^*,\boldsymbol{z}_i)/\pi(\boldsymbol{x},\boldsymbol{z}_i;\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}}{\int \pi(\boldsymbol{x},\boldsymbol{z}_i;\boldsymbol{\theta}) f_1(\boldsymbol{x}|\boldsymbol{x}_i^*,\boldsymbol{z}_i) P(\delta_i=1|\boldsymbol{x}_i^*,\boldsymbol{z}_i)/\pi(\boldsymbol{x},\boldsymbol{z}_i;\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}} \\ &\quad - \sum_{i\in\mathcal{U}} (1-\delta_i) \frac{\pi'(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})}{(1-\pi(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta}))} \\ &= \sum_{i\in\mathcal{U}} \delta_i \frac{\int \pi'(\boldsymbol{x},\boldsymbol{z}_i;\boldsymbol{\theta})\pi(\boldsymbol{x},\boldsymbol{z}_i;\boldsymbol{\theta})^{-1} f_1(\boldsymbol{x}|\boldsymbol{x}_i^*,\boldsymbol{z}_i)\mathrm{d}\boldsymbol{x}}{\int f_1(\boldsymbol{x}|\boldsymbol{x}_i^*,\boldsymbol{z}_i)\mathrm{d}\boldsymbol{x}} - \sum_{i\in\mathcal{U}} (1-\delta_i) \frac{\pi'(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})}{(1-\pi(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta}))} \\ &= \sum_{i\in\mathcal{U}} \delta_i \int \pi'(\boldsymbol{x},\boldsymbol{z}_i;\boldsymbol{\theta})\pi(\boldsymbol{x},\boldsymbol{z}_i;\boldsymbol{\theta})^{-1} f_1(\boldsymbol{x}|\boldsymbol{x}_i^*,\boldsymbol{z}_i)\mathrm{d}\boldsymbol{x} - \sum_{i\in\mathcal{U}} (1-\delta_i) \frac{\pi'(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})}{(1-\pi(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta}))} \end{aligned} \tag{5.24}$$

Using the kernel regression approach to evaluate the integral, one can obtain

$$\begin{aligned} \hat{s}(\boldsymbol{\theta}) &= \sum_{i\in\mathcal{U}} \delta_i \frac{\sum_{j\in B} \pi'(\boldsymbol{x}_j,\boldsymbol{z}_i;\boldsymbol{\theta})\pi(\boldsymbol{x}_j,\boldsymbol{z}_i;\boldsymbol{\theta})^{-1} K_h(\tilde{\boldsymbol{x}}_j,\tilde{\boldsymbol{x}}_i)}{\sum_{k\in B} K_h(\tilde{\boldsymbol{x}}_k,\tilde{\boldsymbol{x}}_i)} - \sum_{i\in\mathcal{U}} (1-\delta_i) \frac{\pi'(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})}{(1-\pi(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta}))} \\ &= \sum_{i\in\mathcal{U}} \delta_i \sum_{j\in B} \pi'(\boldsymbol{x}_j,\boldsymbol{z}_i;\boldsymbol{\theta})\pi(\boldsymbol{x}_j,\boldsymbol{z}_i;\boldsymbol{\theta})^{-1} w_{ij} - \sum_{i\in\mathcal{U}} (1-\delta_i) \frac{\pi'(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})}{(1-\pi(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta}))} \end{aligned} \tag{5.25}$$

where $w_{ij} = K_h(\tilde{\boldsymbol{x}}_j,\tilde{\boldsymbol{x}}_i)/\sum_{k\in B} K_h(\tilde{\boldsymbol{x}}_k,\tilde{\boldsymbol{x}}_i)$ and $\tilde{\boldsymbol{x}} = (\boldsymbol{x}^*,\boldsymbol{z})$ is a combined vector of covariates.

When A is a subsample of the population, then one can substitute sample-weighted quantities into (5.25) in place of their population counterparts to obtain

$$\hat{s}(\boldsymbol{\theta}) = \sum_{i\in\mathcal{U}} \delta_i \sum_{j\in A\cap B} \pi'(\boldsymbol{x}_j,\boldsymbol{z}_i;\boldsymbol{\theta})\pi(\boldsymbol{x}_j,\boldsymbol{z}_i;\boldsymbol{\theta})^{-1} w_{ij} - \sum_{i\in\mathcal{A}} d_i^A (1-\delta_i) \frac{\pi'(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})}{(1-\pi(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta}))} \tag{5.26}$$

$$\text{where } w_{ij} = \frac{d_j^A K_h(\tilde{\boldsymbol{x}}_j,\tilde{\boldsymbol{x}}_i)}{\sum_{k\in A\cap B} d_k^A K_h(\tilde{\boldsymbol{x}}_k,\tilde{\boldsymbol{x}}_i)}$$

Examining the properties of the estimator that results from solving the score equation arising from (5.26), and assessing its performance empirically, are beyond the scope of this thesis, and are left for future work. It is noted that one possible extension of the framework that has been developed in this section is to help address selection bias in the case of a Type IV data scenario. Recall that in this scenario A collects $(\boldsymbol{x}, y)$ and B collects $\boldsymbol{x}$ only. Within the framework proposed in this section, the missing y on B can

be thought of as an extreme form of measurement error, where y is not measured at all. Assessing the effectiveness of this approach to handle selection bias in the Type IV data scenario is also left for future work.

5.8 Concluding Remarks

The reference sample used to correct selection bias in a non-probability dataset may contain auxiliary (or covariate) information that is subject to measurement error. Using the measurement error-prone auxiliary variables to fit a participation probability model may lead to biased estimation of the model parameters. In this chapter a framework has been proposed to simultaneously correct for both covariate measurement error in the reference sample, as well as selection bias in the non-probability sample. The framework is applicable in the case when it is possible to link the records in the reference and non-probability samples. It turns out that the proposed framework can also be applied in the context of SNAR missingness in the non-probability sample.

In both a simulation and real-world data setting, the proposed framework effectively mitigates the measurement error and selection biases, resulting in unbiased estimates of population mean when the measurement error is not as noisy, with RRMSE values very close to the ideal RRMSE achieved when there is no measurement error. When there was greater variability in the measurement error, the proposed approach was still able to address most of the selection bias.

Two possible reasons for the small residual bias in some of the simulation results are the amount of overlap between the reference and non-probability samples, and the kernel bandwidth chosen. Ensuring there is a reasonable overlap between the two samples is an important consideration for correcting measurement error. How to design an efficient reference sample with sufficient overlap between it and the non-probability sample to correct for measurement error is a future research topic. The bandwidth selection approach used in this chapter to choose the kernel bandwidth was effective in finding an optimal bandwidth for a given set of candidate bandwidths. However, more effective bias reduction

may potentially be obtained by widening the set of candidate bandwidths to choose from. This is an area for future research.

The simulation and real-world data results show that not taking any action to resolve potential measurement error is not advisable as the resulting estimates of population mean remain biased. It was found that fitting a simple linear measurement error model with the common sample, and then imputing corrected values of the covariate using the model, could be beneficial. However, the resulting estimates retained some measure of bias, and did not perform as well as the proposed approach.

The proposed framework may also be applied to handle SNAR issues. In the simulation study, it was found that the approach was able to mitigate the bias due to SNAR. When the variability in the auxiliary variables used in the participation probability model was not large, the approach was found to be more effective at completely removing the bias. Including a subsequent calibration step to adjust the initial weights from the participation probability model was also found to further reduce bias and RRMSE. Further work could be undertaken to explore how to adapt the approach to make it more effective for removing SNAR bias when the auxiliary variables are more variable.

This chapter has briefly outlined a similar approach for addressing measurement error in the auxiliary variables that are on the non-probability dataset, or to address selection bias in a Type IV data scenario. However, an examination of its properties and an empirical assessment of its effectiveness are left for future work.

The proposed framework requires the ability to link the data in the reference and non-probability samples, and this may not be possible to do in reality, or to do it perfectly. One possible extension of this work would be to explore how the approach may be applied when there is imperfect linkage between the two samples, or if linkage is not possible at all.

The proposed approach employs kernel regression to estimate the integrals in the score function. In addition to the bandwidth selection challenges already mentioned, the kernel approach was found to be fairly computationally intensive as sample sizes increased. Part

of the reason is likely the fact that double sums needed to be computed, and the way this was implemented in the code. Computational speed could be increased by a more efficient implementation. A possible extension for this work could be to explore alternative non-parametric approaches to estimate the integrals which are not affected by these issues.

Chapter 6

Concluding remarks

6.1 Summary and Conclusions

The ongoing digitalisation of society has resulted in an increasing amount of data being collected about individuals, entities and the environment. These datasets range from small web panel surveys to administrative records, commercial transactions, satellite imagery data, and so on. These datasets provide opportunities for statistical agencies and other organisations to produce statistics that are less costly, less burdensome on the population, and more accurate. In some cases, high quality measurements of variables of interest are available on these datasets. However, they also tend to have a major flaw, that being nonignorable missingness or selection bias. When this is the case, data is available for only a portion of the population, and the mechanism for selection into these datasets is unknown. These datasets may be considered non-probability samples.

In this dissertation, I have developed new methods by which a probability sample, the bread and butter of national statistical agencies, can be used in combination with the non-probability sample to correct the quality issues in the non-probability sample. I first provided an overview of the main methods and approaches that have been proposed to date to address selection bias. There are several main approaches considered in the literature: weighting approaches, which aim to weight the non-probability sample or adjust weights in

a reference probability sample; imputation approaches, which aim to impute the missing variable of interest in either the reference sample or the non-probability sample; or a combination of both. Methods also differ based on the data available to us: is auxiliary information available for the full population, or a sample? Does the variable of interest exist on a reference probability sample, on the non-probability sample, or on both?

A wide-ranging simulation study was conducted using a realistic business population to examine the effectiveness of a cross-section of the estimation methods that have been developed for non-probability data. The simulation explored how nonignorable non-response and measurement error in the non-probability sample would affect the performance of the various estimators. The simulation study also explored how the sample design for the reference sample could be undertaken in conjunction with these estimators.

The results from the simulation study provided some valuable insights for a statistical agency looking to efficiently apply these methods in practice. The best method to use will depend on the variables that are available on the non-probability and reference samples, the nature of the non-response on the non-probability dataset, and the quality of the data on it. When the variable of interest y is available in both the non-probability and reference samples, and there is no measurement error in the non-probability sample, it is not imperative to have the reference sample overlap with the non-probability data. When the non-probability dataset is large and can be linked to the population frame, the calibrated split population estimator can be very efficient. However, when y is subject to measurement error in the non-probability dataset, the best solution tends to be an estimator that relies more heavily on the reference sample as the basis for estimation, such as the GREG or RDI estimators.

When y is not available in the reference sample, the methods proposed by Kim and Wang (2019) and Wang et al. (2021a) were both found in the simulation study to effectively eliminate selection bias under SAR missingness. Including a subsequent calibration weighting step further improves the efficiency of the resulting estimates. Under this data scenario, it is not possible to fully address measurement error in the y variable on the non-probability sample, since there is no corresponding "true" value from the reference sample to facil-

itate the error correction. If the y variable on the non-probability dataset is likely to suffer from measurement error, it would be desirable for the statistical agency to employ a reference sample that also contains the y variable and has some overlapping sample with the non-probability data.

Chapter 3 proposed a new pseudo-likelihood approach to estimating participation probabilities when the overlap between the non-probability sample and a reference sample is known. The new approach was demonstrated to be asymptotically more efficient than other pseudo-likelihood approaches proposed to date when the non-probability sample is large, and always at least as efficient compared to the methods of Kim and Wang (2019) and Chen et al. (2020) under Poisson sampling for A. Results from a comparative simulation study corroborated this theoretical finding. A real-world data application using business taxation data provided insight into how competing methods behaved under skewed populations and different reference sample sampling schemes, and provided further evidence of the benefits from using the proposed approach. This development makes it more feasible to use administrative and big data as the primary source, alongside a small reference sample which corrects for the selection bias.

For the case when it is not possible to identify the overlapping units, I proposed a variation of the original approach which utilised the EM algorithm to simultaneously estimate the overlapping sample and produce estimated participation probabilities. Estimating the overlap between the non-probability sample and the reference sample units incurs a cost in terms of a slightly higher mean squared error, however the empirical studies I conducted showed that it would still be preferred compared with other alternatives which do not require knowledge of the overlap.

In the face of increasingly tight budgets and survey non-response, coupled with increased demand from the public for more timely and detailed information, national statistical organisations need to reconsider how to produce their statistics. The results from Chapter 3 hinted at a way forward where probability samples could cease being the primary source of data, and instead be repurposed and viewed as an audit or reference sample to address quality issues in administrative or big data. This paradigm shift would necessitate a

change in the way the probability sample is designed.

Chapter 4 considered the question of optimal sample design for the reference sample in combination with the pseudo-likelihood approach that I proposed in Chapter 3. Details of the optimal selection probabilities for the reference sample under Poisson sampling, fixed-size probability proportional to size sampling, and stratified simple random sampling are derived. The optimal sample designs require knowledge about the auxiliary variables for the population, and this is typically not available. I proposed a multiple imputation-based approach to impute values for the auxiliary variables where required and then estimate the optimal selection probabilities. Further, a split-population design approach was proposed for the case when it is possible to identify which units on the population frame are not in the non-probability sample. These design ideas were examined empirically through a simulation study and the results demonstrated their usefulness.

In practice, the probability sample and non-probability dataset will often not measure identical concepts, or will measure the same concept with varying levels of measurement error. Chapter 5 developed methods allowing the efficient combination of the probability and non-probability samples even under this challenging scenario. In particular, it considered the question of measurement error in the auxiliary variables on the reference probability sample. An approach was proposed to simultaneously correct for measurement error in the auxiliary information contained on the reference sample dataset, while also addressing selection bias in the non-probability sample. Of note, the approach is also applicable for addressing the case when the selection bias in the non-probability sample also depends on the variable of interest.

The findings from this chapter show that it is important for the statistician to conduct an assessment of the quality of the data that is on the probability and non-probability samples, and to correct any potential measurement error issues in the data being used. Ignoring these issues can lead to biased estimates.

At the same time, the new methods developed throughout this dissertation suggest that the era of the probability sample being the gold standard is changing. National statistical organisations are looking at a future where administrative and big data will become the

main source of data. While there remains an important need for high-quality probability samples, they will reduce in size and their design objectives will change to provide efficient estimates of participation probabilities. The probability samples may not even need to collect the variables of interest; instead, the focus of these samples will be on collecting accurate auxiliary characteristic information to assist in estimating the participation probabilities of units in the population. The future data collection model for the national statistical organisation may ultimately only consist of two high-quality probability samples - one for businesses and one for persons. Of course, this hinges on appropriate administrative or big data being available for the national statistical organisation to use - this may not always be the case.

6.2 Future Directions

As noted in Section 1.1, this dissertation assumes that there is full response in the reference sample. When response rates are relatively high, then non-response may reasonably be assumed to be ignorable conditional on population auxiliary variables. A calibration weighting process would account for the non-response in this case. On the other hand, if the response rate of the reference sample is low, then the reference sample itself becomes more like a non-probability sample. Effectively addressing low response in the reference sample as well as correcting for selection bias in the non-probability sample is an interesting question.

The new estimation methods proposed in this dissertation have often assumed that we are able to perform an accurate linkage of the units in the non-probability sample to their corresponding records on a reference probability sample. How to account for possible linkage errors as part of the estimation process would be of interest. Further, the proposed approach for addressing measurement error requires one to be able to link the data accurately. Whether the approach can be extended to address this error when linkage of the datasets is not possible is an open question.

The methods in this dissertation make certain assumptions regarding the selection mech-

anism into the non-probability dataset. For example, it is often assumed that selection is ignorable and a logistic regression model is applicable for the participation probabilities. Relaxing these model assumptions as much as possible would help the practitioner to avoid model misspecification issues.

The area of optimal design for the reference samples is a topic that could be further explored. While the split-population approach holds promise, there still remain unanswered questions about how to best design a split-population reference sample. For example, how are optimal selection probabilities obtained and how best should one determine the overlapping sample between the reference and non-probability samples to help address covariate measurement error, or nonignorable missingness in the non-probability sample? One idea raised in this thesis but not carried forward is whether it is possible to adopt audit or validation sampling ideas such as those described in Zhang (2021) and Zhang (2023) in designing the reference sample.

In this dissertation I have focussed on the case where there is a single non-probability sample combined with a single probability sample. In reality, there may be several non-probability samples to be combined with a single probability sample, with varying selection bias characteristics in each of the non-probability samples. How best to optimally design the probability sample in this multiple-objective case - where the participation probabilities and/or the set of auxiliary variables may differ between non-probability samples - is a challenging question for future research.

In the non-probability inference literature to date, there is generally a clearly defined data structure setup in terms of which dataset(s) the variables of interest and auxiliary variables are available in. However, there may be situations which are in essence a mixture of data structures where some of the variables of interest are available in the non-probability sample, while others may only be available through the probability sample. How should estimation best proceed for the set of variables of interest, and how should the probability sample be optimally designed to cater for this mix?

Statistical agencies are often interested in ensuring accurate estimates of movement. To facilitate this in regular probability surveys, some overlapping sample between time periods

is often required, to capture the movement in the common units. Translating this to the non-probability sample case is one question for further investigation. When the aim is to capture accurate movement estimates, what are the implications for how best to utilise the non-probability sample and the accompanying reference sample? If the non-probability sample is updated very frequently, how often should the reference probability sample be updated?

There are a growing number and variety of methods to address selection bias in non-probability samples. For the practitioner, how to choose the best method for a particular scenario can be difficult. It would be beneficial to have somes guideline for choosing the best method for a given situation. The material from Chapter 2 in this thesis goes some way towards facilitating an informed choice for a number of alternatives. At the same time, how to determine whether or not a given administrative or big dataset suffers from selection bias, and what is the best subset of auxiliary variables for treating the selection bias, remain open questions. The quality indicators proposed by Shlomo and Kim (2025) contribute towards the first question, while the variable selection procedure proposed by Beaumont et al. (2024b) provides assistance in informing the second.

Appendix A

ABS DataLab Disclaimer

The parameters used to create the artificial population for the simulation study in Chapter 2, and the data used to create the derived population for the business turnover applications in Chapters 3 and 5, are based, in part, on data supplied to the ABS under the Taxation Administration Act 1953, A New Tax System (Australian Business Number) Act 1999, Australian Border Force Act 2015, Social Security (Administration) Act 1999, A New Tax System (Family Assistance) (Administration) Act 1999, Paid Parental Leave Act 2010 and/or the Student Assistance Act 1973. Such data may only be used for the purpose of administering the Census and Statistics Act 1905 or performance of functions of the ABS as set out in section 6 of the Australian Bureau of Statistics Act 1975. No individual information collected under the Census and Statistics Act 1905 is provided back to custodians for administrative or regulatory purposes. Any discussion of data limitations or weaknesses is in the context of using the data for statistical purposes and is not related to the ability of the data to support the Australian Taxation Office, Australian Business Register, Department of Social Services and/or Department of Home Affairs' core operational requirements.

Legislative requirements to ensure privacy and secrecy of these data have been followed. For access to PLIDA and/or BLADE data under Section 16A of the ABS Act 1975 or enabled by section 15 of the Census and Statistics (Information Release and Access) De-

termination 2018, source data are de-identified and so data about specific individuals has not been viewed in conducting this analysis. In accordance with the Census and Statistics Act 1905, results have been treated where necessary to ensure that they are not likely to enable identification of a particular person or organisation.

Appendix B

Proofs and regularity conditions

B.1 Regularity Conditions for OE Participation Probabilities

It is assumed that the following regularity conditions hold:

C1 The finite population size N and the sample sizes n_A and n_B satisfy $\lim_{N\to\infty} n_A/N = f_A \in (0,1)$ and $\lim_{N\to\infty} n_B/N = f_B \in (0,1)$

C2 There exist c_1 and c_2 such that $0 < c_1 \leq N\pi_i^A/n_A \leq c_2$ and $0 < c_1 \leq N\pi_i^B/n_B \leq c_2$ for all units $i \in \mathcal{U}$.

C3 Under the finite population and the sampling design for A, we have $N^{-1}\sum_{i\in A} d_i^A \boldsymbol{u}_i - N^{-1}\sum_{i\in\mathcal{U}} \boldsymbol{u}_i = O_p(n_A^{-1/2})$ for $\boldsymbol{u}_i = \boldsymbol{x}_i, y_i$, where $d_i^A = 1/\pi_i^A$ and the order in probability is with respect to the sampling mechanism used to select A.

C4 The finite population and the propensity scores satisfy $N^{-1}\sum_{i\in\mathcal{U}} y_i^2 = O(1)$, $N^{-1}\sum_{i\in\mathcal{U}} \|\boldsymbol{x}_i\|^3 = O(1)$, and $N^{-1}\sum_{i\in\mathcal{U}} \pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T$ is a positive definite matrix for all finite population sizes N.

Condition **C1** means that the sample fractions of the probability and non-probability samples are bounded, while condition **C2** means that the inclusion probabilities for the

probability and non-probability samples do not differ in terms of order of magnitude from simple random sampling. Condition **C3** ensures consistency of the Horvitz-Thompson estimators obtained from the probability sample. Condition **C4** are finite moment conditions to validate Taylor series expansions.

B.2 Consistency and Variance Derivations for OE IPW estimator

Following Chen et al. (2020), we set up a joint estimation problem. Firstly, define $\boldsymbol{\eta}^T = (\mu, \boldsymbol{\theta}^T)$. The true (target) parameter values are $\boldsymbol{\eta}_0^T = (\mu_y, \boldsymbol{\theta}_0^T)$, while the OE IPW estimate and propensity score model parameters are combined as $\hat{\boldsymbol{\eta}}^T = (\hat{\mu}, \hat{\boldsymbol{\theta}}^T)$. $\hat{\boldsymbol{\eta}}^T$ is the solution to the combined estimating equation system

$$\begin{aligned}\boldsymbol{\Phi}(\boldsymbol{\eta}) &= \begin{pmatrix} U(\mu, \boldsymbol{\theta}) \\ S(\boldsymbol{\theta}) \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{N}\sum_{i\in\mathcal{U}} \frac{\delta_i(y_i-\mu)}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} \\ \frac{1}{N}\sum_{i\in\mathcal{U}}\Big\{\delta_i\boldsymbol{x}_i - \delta_i\pi(\boldsymbol{x}_i;\boldsymbol{\theta})\boldsymbol{x}_i\Big\} - \frac{1}{N}\sum_{i\in\mathcal{U}} I_i(1-\delta_i)d_i^A\pi(\boldsymbol{x}_i;\boldsymbol{\theta})\boldsymbol{x}_i \end{pmatrix} = \mathbf{0}.\end{aligned}$$

We consider the joint randomisation of the propensity score model and the sample design for A. Recall that $\pi_i^B = \pi(\boldsymbol{x}_i;\boldsymbol{\theta}_0)$. When Assumptions **A1**, **A2**, and the propensity score model all hold, and further assuming that the probability sample is drawn independent of the non-probability sample, then $E\{\boldsymbol{\Phi}(\boldsymbol{\eta}_0)\} = \mathbf{0}$, since

$$\frac{1}{N}\sum_{i\in\mathcal{U}}\frac{E(\delta_i)(y_i-\mu_y)}{\pi_i^B} = \frac{1}{N}\sum_{i\in\mathcal{U}}(y_i-\mu_y) = 0$$

and

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in\mathcal{U}}\Big\{E(\delta_i)\boldsymbol{x}_i - E(\delta_i)\pi_i^B\boldsymbol{x}_i\Big\} - \frac{1}{N}\sum_{i\in\mathcal{U}}(E(I_i) - E(I_i\delta_i))d_i^A\pi_i^B\boldsymbol{x}_i \\
&= \frac{1}{N}\sum_{i\in\mathcal{U}}\Big\{\pi_i^B\boldsymbol{x}_i - (\pi_i^B)^2\boldsymbol{x}_i\Big\} - \frac{1}{N}\sum_{i\in\mathcal{U}}(\pi_i^A - \pi_i^A\pi_i^B)\frac{1}{\pi_i^A}\pi_i^B\boldsymbol{x}_i \\
&= \mathbf{0}
\end{aligned}$$

Note that $\boldsymbol{\Phi}(\hat{\boldsymbol{\eta}}) = \mathbf{0}$. Under the regularity conditions in Section B.1, we apply a first-order Taylor expansion to obtain

$$\mathbf{0} = \boldsymbol{\Phi}(\hat{\boldsymbol{\eta}}) \approx \boldsymbol{\Phi}(\boldsymbol{\eta}_0) + \boldsymbol{\phi}(\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \tag{B.1}$$

where $\boldsymbol{\phi}(\boldsymbol{\eta})$ is defined as $\partial\boldsymbol{\Phi}(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$, and is given by

$$\boldsymbol{\phi}(\boldsymbol{\eta}) = \begin{pmatrix} \frac{\partial U}{\partial \mu} & \frac{\partial U}{\partial \boldsymbol{\theta}} \\ \frac{\partial S}{\partial \mu} & \frac{\partial S}{\partial \boldsymbol{\theta}} \end{pmatrix}$$

Under suitable regularity conditions as $n \to \infty$ (see Stefanski and Boos 2002), we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{\Phi}(\boldsymbol{x}_i, y_i, \boldsymbol{\eta})}{\partial\boldsymbol{\eta}} \xrightarrow{p} E\Big\{\frac{\partial\boldsymbol{\Phi}(\boldsymbol{x}, y, \boldsymbol{\eta})}{\partial\boldsymbol{\eta}}\Big\}$$

We can then re-arrange (B.1) to obtain

$$\begin{aligned}
\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 &\approx -[\boldsymbol{\phi}(\boldsymbol{\eta}_0)]^{-1}\boldsymbol{\Phi}(\boldsymbol{\eta}_0) \\
&\approx -[E(\boldsymbol{\phi}(\boldsymbol{\eta}_0))]^{-1}\boldsymbol{\Phi}(\boldsymbol{\eta}_0)
\end{aligned} \tag{B.2}$$

Now $E\{\boldsymbol{\Phi}(\boldsymbol{\eta}_0)\} = \mathbf{0}$, and drawing on the standard M-estimation argument of Stefanski and Boos (2002), we can conclude that $E(\hat{\boldsymbol{\eta}})$ converges to $\boldsymbol{\eta}_0$ asymptotically, that is, $\hat{\boldsymbol{\eta}} \xrightarrow{p} \boldsymbol{\eta}_0$ as the population and sample sizes tend to infinity.

To derive the variance for the OE IPW estimator, we apply the sandwich formula (Stefanski and Boos 2002):

$$Var(\hat{\boldsymbol{\eta}}) = E\{\boldsymbol{\phi}(\boldsymbol{\eta}_0)\}^{-1} Var\{\boldsymbol{\Phi}(\boldsymbol{\eta}_0)\}[E\{\boldsymbol{\phi}(\boldsymbol{\eta}_0)\}^{-1}]^T \tag{B.3}$$

The components of $\boldsymbol{\phi}(\boldsymbol{\eta})$ can be derived giving

$$\boldsymbol{\phi}(\boldsymbol{\eta}) = \begin{pmatrix} \frac{-1}{N}\sum_{i\in\mathcal{U}} \frac{\delta_i}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} & \frac{-1}{N}\sum_{i\in\mathcal{U}} \frac{1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta})}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})}\delta_i(y_i-\mu)\boldsymbol{x}_i^T \\ \mathbf{0} & \frac{-1}{N}\sum_{i\in\mathcal{U}}\{\delta_i + I_i(1-\delta_i)d_i^A\}\pi(\boldsymbol{x}_i;\boldsymbol{\theta})(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))\boldsymbol{x}_i\boldsymbol{x}_i^T \end{pmatrix}.$$

Then

$$E\{\boldsymbol{\phi}(\boldsymbol{\eta}_0)\} = \begin{pmatrix} -1 & \frac{-1}{N}\sum_{i\in\mathcal{U}}(1-\pi_i^B)(y_i-\mu_y)\boldsymbol{x}_i^T \\ \mathbf{0} & \frac{-1}{N}\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T \end{pmatrix}$$

and the inverse is given as

$$E\{\boldsymbol{\phi}(\boldsymbol{\eta}_0)\}^{-1} = \begin{pmatrix} -1 & \sum_{i\in\mathcal{U}}(1-\pi_i^B)(y_i-\mu_y)\boldsymbol{x}_i^T\Big[\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1} \\ \mathbf{0} & -\Big[\frac{1}{N}\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1} \end{pmatrix}. \tag{B.4}$$

To determine $Var\{\boldsymbol{\Phi}(\boldsymbol{\eta}_0)\}$, we calculate $Var(U(\mu_y,\boldsymbol{\theta}_0))$, $Var(S(\boldsymbol{\theta}_0))$, and $Cov(U(\mu_y,\boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0))$.

Firstly, $Var(U(\mu_y,\boldsymbol{\theta}_0))$ is derived as

$$\begin{aligned} Var(U(\mu_y,\boldsymbol{\theta}_0)) &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}\frac{Var(\delta_i)(y_i-\mu_y)^2}{(\pi_i^B)^2} \\ &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}\frac{(1-\pi_i^B)(y_i-\mu_y)^2}{\pi_i^B} \end{aligned} \tag{B.5}$$

To find $Var(S(\boldsymbol{\theta}_0))$, we make use of the decomposition

$$Var(S(\boldsymbol{\theta}_0)) = E_d\Big\{Var(S(\boldsymbol{\theta}_0)|A)\Big\} + Var_d\Big\{E(S(\boldsymbol{\theta}_0)|A)\Big\}$$

where E_d and Var_d refer to expectations and variances with respect to the probability sample design respectively.

Now

$$\begin{aligned} Var(S(\boldsymbol{\theta}_0)|A) &= Var\Big[\frac{1}{N}\sum_{i\in\mathcal{U}}\Big\{\delta_i\boldsymbol{x}_i - \delta_i\pi_i^B\boldsymbol{x}_i\Big\} - \frac{1}{N}\sum_{i\in\mathcal{U}} I_i(1-\delta_i)d_i^A\pi_i^B\boldsymbol{x}_i|A\Big] \\ &= Var\Big[\frac{1}{N}\sum_{i\in\mathcal{U}}\delta_i(1-\pi_i^B + d_i^A I_i\pi_i^B)\boldsymbol{x}_i - \frac{1}{N}\sum_{i\in\mathcal{U}} I_i d_i^A\pi_i^B\boldsymbol{x}_i|A\Big] \\ &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-\pi_i^B + d_i^A I_i\pi_i^B)^2\boldsymbol{x}_i\boldsymbol{x}_i^T Var(\delta_i) \end{aligned}$$

Expanding out and taking the expectation, we obtain

$$E_d\Big\{Var(S(\boldsymbol{\theta}_0)|A)\Big\} = \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big[\pi_i^B(1-\pi_i^B)(1+(d_i^A-1)(\pi_i^B)^2)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]. \tag{B.6}$$

Then

$$\begin{aligned} E(S(\boldsymbol{\theta}_0)|A) &= \frac{1}{N}\sum_{i\in\mathcal{U}}\Big[\pi_i^B\boldsymbol{x}_i - (\pi_i^B)^2\boldsymbol{x}_i\Big] + \frac{1}{N}\sum_{i\in A}(\pi_i^B)^2 d_i^A\boldsymbol{x}_i - \frac{1}{N}\sum_{i\in A} d_i^A\pi_i^B\boldsymbol{x}_i \\ \implies Var_d\Big\{E(S(\boldsymbol{\theta}_0)|A)\Big\} &= \frac{1}{N^2}Var_d\Big[\sum_{i\in A}\{(\pi_i^B)^2 d_i^A\boldsymbol{x}_i - d_i^A\pi_i^B\boldsymbol{x}_i\}\Big] \\ &= \frac{1}{N^2}Var_d\Big(\sum_{i\in A} d_i^A\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\Big) \end{aligned} \tag{B.7}$$

where $Var_d(\boldsymbol{x})$ is the design-based variance-covariance matrix for variables $\boldsymbol{x}$ under the probability sample design.

Combining (B.6) and (B.7), we obtain

$$\begin{aligned} Var(S(\boldsymbol{\theta}_0)) = \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big[\pi_i^B(1-\pi_i^B)(1+(d_i^A-1)(\pi_i^B)^2)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big] + \\ \frac{1}{N^2}Var_d\Big(\sum_{i\in A} d_i^A\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\Big) \end{aligned} \tag{B.8}$$

To find $Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0))$, we will use

$$Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0)) = E_d\Big[Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0)|A)\Big] + Cov_d\Big[E(U(\mu_y, \boldsymbol{\theta}_0)|A), E(S(\boldsymbol{\theta}_0)|A)\Big]$$

where E_d is defined as previously and Cov_d is the covariance under the probability sample design.

Now,

$$\begin{aligned} E(U(\mu_y, \boldsymbol{\theta}_0)|A) &= \frac{1}{N}\sum_{i\in\mathcal{U}}(y_i - \mu_y) = 0 \\ &\implies Cov_d\Big[E(U(\mu_y, \boldsymbol{\theta}_0)|A), E(S(\boldsymbol{\theta}_0)|A)\Big] = 0. \end{aligned} \tag{B.9}$$

Finally,

$$\begin{aligned} &Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0)|A) \\ &= Cov\Big[\frac{1}{N}\sum_{i\in\mathcal{U}}\frac{\delta_i(y_i - \mu_y)}{\pi_i^B}, \frac{1}{N}\sum_{i\in\mathcal{U}}\{\delta_i\boldsymbol{x}_i - \delta_i\pi_i^B\boldsymbol{x}_i - I_i(1-\delta_i)d_i^A\pi_i^B\boldsymbol{x}_i\}|A\Big] \\ &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}Cov\Big[\frac{\delta_i(y_i - \mu_y)}{\pi_i^B}, \delta_i\{\boldsymbol{x}_i - \pi_i^B\boldsymbol{x}_i + I_i d_i^A\pi_i^B\boldsymbol{x}_i\}|A\Big] \\ &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}(y_i - \mu_y)(1-\pi_i^B)(1-\pi_i^B + I_i d_i^A\pi_i^B)\boldsymbol{x}_i \\ &\implies E_d\Big[Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0)|A)\Big] = \frac{1}{N^2}\sum_{i\in\mathcal{U}}(y_i - \mu_y)(1-\pi_i^B)\boldsymbol{x}_i \end{aligned} \tag{B.10}$$

where we make use of the bilinear property of the covariance function and assume that $\delta_i \perp \delta_j$ for $i, j \in B$.

Combining (B.9) and (B.10), we get

$$Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0)) = \frac{1}{N^2}\sum_{i\in\mathcal{U}}(y_i - \mu_y)(1-\pi_i^B)\boldsymbol{x}_i \tag{B.11}$$

Denote the components of $E\{\boldsymbol{\phi}(\boldsymbol{\eta}_0)\}^{-1}$ and $Var\{\boldsymbol{\Phi}(\boldsymbol{\eta}_0)\}$ as

$$E\{\boldsymbol{\phi}(\boldsymbol{\eta}_0)\}^{-1} = \begin{pmatrix} -1 & \boldsymbol{b}^T \\ \mathbf{0} & \boldsymbol{d} \end{pmatrix}, Var\{\boldsymbol{\Phi}(\boldsymbol{\eta}_0)\} = \begin{pmatrix} e & \boldsymbol{f} \\ \boldsymbol{f} & \boldsymbol{h} \end{pmatrix}$$

where

$$\begin{aligned}
\boldsymbol{b}^T &= \sum_{i\in\mathcal{U}}(1-\pi_i^B)(y_i-\mu_y)\boldsymbol{x}_i^T\Big[\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1} \text{ from } (B.4)\\
\boldsymbol{d} &= -\Big[\frac{1}{N}\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1} \text{ from } (B.4)\\
e &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}\frac{(1-\pi_i^B)(y_i-\mu_y)^2}{\pi_i^B} \text{ from } (B.5)\\
\boldsymbol{f} &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}(y_i-\mu_y)(1-\pi_i^B)\boldsymbol{x}_i \text{ from } (B.11)\\
\boldsymbol{h} &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big[\pi_i^B(1-\pi_i^B)(1+(d_i^A-1)(\pi_i^B)^2)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big] + \frac{1}{N^2}Var_d\Big(\sum_{i\in A}d_i^A\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\Big)\\
&\quad \text{from } (B.8)
\end{aligned}$$

Then, using the sandwich formula, $Var(\hat{\boldsymbol{\eta}})$ can be derived as

$$\begin{aligned}
Var(\hat{\boldsymbol{\eta}}) &= E\{\boldsymbol{\phi}(\boldsymbol{\eta}_0)\}^{-1}Var\{\boldsymbol{\Phi}(\boldsymbol{\eta}_0)\}[E\{\boldsymbol{\phi}(\boldsymbol{\eta}_0)\}^{-1}]^T\\
&= \begin{pmatrix} -1 & \boldsymbol{b}^T \\ \mathbf{0} & \boldsymbol{d} \end{pmatrix}\begin{pmatrix} e & \boldsymbol{f}^T \\ \boldsymbol{f} & \boldsymbol{h} \end{pmatrix}\begin{pmatrix} -1 & \mathbf{0}^T \\ \boldsymbol{b} & \boldsymbol{d} \end{pmatrix}\\
&= \begin{pmatrix} e-\boldsymbol{b}^T\boldsymbol{f}-\boldsymbol{b}^T\boldsymbol{f}+\boldsymbol{b}^T\boldsymbol{h}\boldsymbol{b} & -\boldsymbol{f}^T\boldsymbol{d}+\boldsymbol{b}^T\boldsymbol{h}\boldsymbol{d} \\ -\boldsymbol{d}\boldsymbol{f}+\boldsymbol{d}\boldsymbol{h}\boldsymbol{b} & \boldsymbol{d}\boldsymbol{h}\boldsymbol{d} \end{pmatrix}
\end{aligned} \tag{B.12}$$

In particular, the variance of the OE IPW estimator is the first diagonal element of the

matrix (B.12):

$$
\begin{aligned}
Var(\hat{\mu}) &= e - \boldsymbol{b}^T\boldsymbol{f} - \boldsymbol{b}^T\boldsymbol{f} + \boldsymbol{b}^T\boldsymbol{h}\boldsymbol{b} \\
&= \frac{1}{N^2}\sum_{i\in\mathcal{U}}\frac{1-\pi_i^B}{\pi_i^B}(y_i-\mu_y)^2 - \frac{\boldsymbol{b}^T}{N^2}(y_i-\mu_y)(1-\pi_i^B)\boldsymbol{x}_i - \\
&\quad \frac{\boldsymbol{b}^T}{N^2}\sum_{i\in\mathcal{U}}(y_i-\mu_y)(1-\pi_i^B)\boldsymbol{x}_i + \boldsymbol{b}^T\boldsymbol{h}\boldsymbol{b} \\
&= \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-\pi_i^B)\pi_i^B\Big[\frac{(y_i-\mu_y)^2}{(\pi_i^B)^2} - 2\frac{(y_i-\mu_y)\boldsymbol{b}^T\boldsymbol{x}_i}{\pi_i^B}\Big] + \boldsymbol{b}^T\boldsymbol{h}\boldsymbol{b} \\
&= \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-\pi_i^B)\pi_i^B\Big[\frac{(y_i-\mu_y)^2}{(\pi_i^B)^2} - 2\frac{(y_i-\mu_y)\boldsymbol{b}^T\boldsymbol{x}_i}{\pi_i^B}\Big] + \\
&\quad \boldsymbol{b}^T\frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big[\pi_i^B(1-\pi_i^B)\Big\{1+(d_i^A-1)(\pi_i^B)^2\Big\}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{V}\boldsymbol{b} \\
&= \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-\pi_i^B)\pi_i^B\Big[\frac{(y_i-\mu_y)^2}{(\pi_i^B)^2} - 2\frac{(y_i-\mu_y)}{\pi_i^B}\boldsymbol{b}^T\boldsymbol{x}_i + \boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big] + \\
&\quad \boldsymbol{b}^T\frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big[\pi_i^B(1-\pi_i^B)\Big\{(d_i^A-1)(\pi_i^B)^2\Big\}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{V}\boldsymbol{b} \\
&= \frac{1}{N^2}\sum_{i\in\mathcal{U}}(1-\pi_i^B)\pi_i^B\Big[\frac{(y_i-\mu_y)}{(\pi_i^B)} - \boldsymbol{b}^T\boldsymbol{x}_i\Big]^2 + \qquad \text{(B.13)} \\
&\quad \boldsymbol{b}^T\boldsymbol{V}\boldsymbol{b} + \qquad \text{(B.14)} \\
&\quad \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big[(\pi_i^B)^3(1-\pi_i^B)(d_i^A-1)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b}\Big] \qquad \text{(B.15)}
\end{aligned}
$$

where $\boldsymbol{V} = \frac{1}{N^2}Var_d\Big(\sum_{i\in A} d_i^A\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\Big)$.

B.3 Variance of the Kim and Wang (2019) estimator

We derive the variance for the data integration estimator proposed by Kim and Wang (2019) (which we denote by KW), assuming a logistic regression model for the π_i^B.

First, noting the KW pseudo log-likelihood is

$$\begin{aligned} l_{KW}(\boldsymbol{\theta}) &= \sum_{i\in A} d_i^A \delta_i \log\pi(\boldsymbol{x}_i;\boldsymbol{\theta}) + \sum_{i\in A} d_i^A (1-\delta_i)\log(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta})) \\ &= \sum_{i\in A} d_i^A \delta_i \Big(\boldsymbol{x}_i^T\boldsymbol{\theta} - \log(1+\exp(\boldsymbol{x}_i^T\boldsymbol{\theta}))\Big) - \sum_{i\in A} d_i^A (1-\delta_i)\log(1+\exp(\boldsymbol{x}_i^T\boldsymbol{\theta})) \end{aligned} \tag{B.16}$$

we can produce the score function as

$$\begin{aligned} \frac{\partial l_{KW}}{\partial \boldsymbol{\theta}} &= \sum_{i\in A} d_i^A \delta_i (\boldsymbol{x}_i - \pi(\boldsymbol{x}_i;\boldsymbol{\theta})\boldsymbol{x}_i) - \sum_{i\in A} d_i^A (1-\delta_i)\pi(\boldsymbol{x}_i;\boldsymbol{\theta})\boldsymbol{x}_i \\ &= \sum_{i\in\mathcal{U}} I_i d_i^A (\delta_i - \pi(\boldsymbol{x}_i;\boldsymbol{\theta}))\boldsymbol{x}_i \end{aligned}$$

The combined estimating equation system $\boldsymbol{\Phi}_{KW}(\boldsymbol{\eta})$ for the KW method is thus

$$\boldsymbol{\Phi}_{KW}(\boldsymbol{\eta}) = \begin{pmatrix} U(\mu,\boldsymbol{\theta}) \\ S_{KW}(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \frac{1}{N}\sum_{i\in\mathcal{U}} \frac{\delta_i(y_i-\mu)}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} \\ \frac{1}{N}\sum_{i\in\mathcal{U}} I_i d_i^A(\delta_i - \pi(\boldsymbol{x}_i;\boldsymbol{\theta}))\boldsymbol{x}_i \end{pmatrix} = \mathbf{0} \tag{B.17}$$

For the rest of this sub-section we will omit the KW subscript for convenience, and all terms in the variance derivation will refer to the KW method unless otherwise stated.

We wish to calculate the components of the sandwich estimator (B.3). We can derive $\boldsymbol{\phi}(\boldsymbol{\eta}) = \partial\boldsymbol{\Phi}(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$ as

$$\boldsymbol{\phi}(\boldsymbol{\eta}) = \begin{pmatrix} \frac{-1}{N}\sum_{i\in\mathcal{U}} \frac{\delta_i}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} & \frac{-1}{N}\sum_{i\in\mathcal{U}} \frac{1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta})}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})}\delta_i(y_i-\mu)\boldsymbol{x}_i^T \\ \mathbf{0} & \frac{-1}{N}\sum_{i\in\mathcal{U}} I_i d_i^A \pi(\boldsymbol{x}_i;\boldsymbol{\theta})(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))\boldsymbol{x}_i\boldsymbol{x}_i^T \end{pmatrix}.$$

Then

$$E\{\boldsymbol{\phi}(\boldsymbol{\eta}_0)\} = \begin{pmatrix} -1 & \frac{-1}{N}\sum_{i\in\mathcal{U}}(1-\pi_i^B)(y_i-\mu_y)\boldsymbol{x}_i^T \\ \mathbf{0} & \frac{-1}{N}\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T \end{pmatrix}$$

and

$$E\{\boldsymbol{\phi}(\boldsymbol{\eta}_0)\}^{-1} = \begin{pmatrix} -1 & \sum_{i\in\mathcal{U}}(1-\pi_i^B)(y_i-\mu_y)\boldsymbol{x}_i^T\Big[\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1} \\ \mathbf{0} & -\Big[\frac{1}{N}\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1} \end{pmatrix} \tag{B.18}$$

which is unchanged from the OE IPW estimator.

$Var(U(\mu_y, \boldsymbol{\theta}_0))$ remains the same as before, that is,

$$Var(U(\mu_y, \boldsymbol{\theta}_0)) = \frac{1}{N^2}\sum_{i\in\mathcal{U}}\frac{(1-\pi_i^B)(y_i-\mu_y)^2}{\pi_i^B} \tag{B.19}$$

To find $Var(S(\boldsymbol{\theta}_0))$, we will again use

$$Var(S(\boldsymbol{\theta}_0)) = E_d\Big\{Var(S(\boldsymbol{\theta}_0)|A)\Big\} + Var_d\Big\{E(S(\boldsymbol{\theta}_0)|A)\Big\}.$$

First,

$$\begin{aligned} Var(S(\boldsymbol{\theta}_0)|A) &= Var(\frac{1}{N}\sum_{i\in\mathcal{U}} I_i d_i^A(\delta_i - \pi_i^B)\boldsymbol{x}_i|A) \\ &= \frac{1}{N^2}\sum_{i\in\mathcal{U}} I_i (d_i^A)^2\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T \\ \implies E_d(Var(S(\boldsymbol{\theta}_0)|A)) &= \frac{1}{N^2}\sum_{i\in\mathcal{U}} d_i^A\pi^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T. \end{aligned} \tag{B.20}$$

Then

$$\begin{aligned} E(S(\boldsymbol{\theta}_0)|A) &= E\Big(\frac{1}{N}\sum_{i\in\mathcal{U}} I_i d_i^A(\delta_i - \pi_i^B)\boldsymbol{x}_i|A\Big) = 0 \\ \implies Var_d(E(S(\boldsymbol{\theta}_0)|A)) &= 0 \end{aligned} \tag{B.21}$$

To find $Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0))$, we will again use

$$Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0)) = E_d\Big[Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0)|A)\Big] + Cov_d\Big[E(U(\mu_y, \boldsymbol{\theta}_0)|A), E(S(\boldsymbol{\theta}_0)|A)\Big].$$

$Cov_d\Big[E(U(\mu_y, \boldsymbol{\theta}_0)|A), E(S(\boldsymbol{\theta}_0)|A)\Big] = 0$ as $E(U(\mu_y, \boldsymbol{\theta}_0)|A) = 1/N \sum_{i=1}^{N}(y_i - \mu_y) = 0$. From Appendix A in Kim and Wang (2019), we have

$$\begin{aligned} Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0)|A) &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}\Big((y_i - \mu_y)I_i d_i^A(1-\pi_i^B)\boldsymbol{x}_i \\ \implies E_d(Cov(U(\mu_y, \boldsymbol{\theta}_0), S(\boldsymbol{\theta}_0)|A)) &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}(y_i - \mu_y)(1-\pi_i^B)\boldsymbol{x}_i. \end{aligned} \tag{B.22}$$

So

$$Var\{\boldsymbol{\Phi}(\boldsymbol{\eta}_0)\} = \begin{pmatrix} \frac{1}{N^2}\sum_{i\in\mathcal{U}}\frac{(1-\pi_i^B)(y_i-\mu_y)^2}{\pi_i^B} & \frac{1}{N^2}\sum_{i\in\mathcal{U}}(y_i-\mu_y)(1-\pi_i^B)\boldsymbol{x}_i \\ \frac{1}{N^2}\sum_{i\in\mathcal{U}}(y_i-\mu_y)(1-\pi_i^B)\boldsymbol{x}_i & \frac{1}{N^2}\sum_{i\in\mathcal{U}} d_i^A\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T \end{pmatrix}. \tag{B.23}$$

Feeding (B.18) and (B.23) into the sandwich formula and using (B.12), the variance of the estimator from Kim and Wang (2019) is given by

$$\begin{aligned} Var(\hat{\mu}) = &\frac{1}{N^2}\sum_{i\in\mathcal{U}}\frac{1-\pi_i^B}{\pi_i^B}(y_i-\mu_y)^2 \\ &-2\boldsymbol{b}^T\frac{1}{N^2}\sum_{i\in\mathcal{U}}(y_i-\mu_y)(1-\pi_i^B)\boldsymbol{x}_i \\ &+\boldsymbol{b}^T\frac{1}{N^2}\sum_{i\in\mathcal{U}} d_i^A\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \\ = &\frac{1}{N^2}\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\Big(\frac{y_i-\mu_y}{\pi_i^B} - \boldsymbol{b}^T\boldsymbol{x}_i\Big)^2 + \frac{1}{N^2}\sum_{i\in\mathcal{U}}(d_i^A-1)\pi_i^B(1-\pi_i^B)\boldsymbol{b}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{b} \end{aligned} \tag{B.24}$$

where $\boldsymbol{b}^T$ is defined as

$$\boldsymbol{b}^T = \sum_{i\in\mathcal{U}}(1-\pi_i^B)(y_i-\mu_y)\boldsymbol{x}_i^T\Big[\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1}$$

B.4 Common variance expression for Hájek-like IPW estimators under Poisson sampling for A

We derive a generic variance expression for the Hájek-like IPW estimator (3.1) involving the score function terms under the assumption of Poisson sampling for A. The variance expression can be applied to the various estimators considered in Chapter 3.

Define $\hat{\mu}(\boldsymbol{\theta}) = \hat{Y}(\boldsymbol{\theta})/\hat{N}_B(\boldsymbol{\theta})$, where $\hat{Y}(\boldsymbol{\theta}) = \sum_{\mathcal{U}} \delta_i y_i/\pi(\boldsymbol{x};\boldsymbol{\theta})$ and $\hat{N}_B(\boldsymbol{\theta}) = \sum_{\mathcal{U}} \delta_i/\pi(\boldsymbol{x};\boldsymbol{\theta})$. Then $\hat{\mu}_{IPW} = \hat{\mu}(\hat{\boldsymbol{\theta}})$. Recall that $\pi_i^B = \pi(\boldsymbol{x};\boldsymbol{\theta}_0)$, and additionally let $\pi_i'^B$ be the partial derivative $\partial\pi(\boldsymbol{x};\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ of π_i^B evaluated at $\boldsymbol{\theta}_0$. Under the regularity conditions in Section B.1, a first-order Taylor expansion around $\boldsymbol{\theta}_0$ gives

$$\hat{\mu}(\hat{\boldsymbol{\theta}}) - \mu_y \approx \hat{\mu}(\boldsymbol{\theta}_0) - \mu_y + \hat{\mu}'(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{B.25}$$

where $\hat{\mu}'(\boldsymbol{\theta}_0)$ is the partial derivative $\partial\hat{\mu}/\partial\boldsymbol{\theta}$ of $\hat{\mu}$ evaluated at $\boldsymbol{\theta}_0$. The terms $\hat{\mu}(\boldsymbol{\theta}_0) - \mu_y$ on the right hand side of B.25 can be expressed as

$$\begin{aligned}
\hat{\mu}(\boldsymbol{\theta}_0) - \mu_y &= \frac{\hat{Y}(\boldsymbol{\theta}_0) - \mu_y \hat{N}_B(\boldsymbol{\theta}_0)}{\hat{N}_B(\boldsymbol{\theta}_0)} \\
&= \sum_{i\in\mathcal{U}} \frac{\delta_i(y_i - \mu_y)}{\pi_i^B \hat{N}_B(\boldsymbol{\theta}_0)} \\
&= \frac{N}{\hat{N}_B(\boldsymbol{\theta}_0)} \frac{1}{N} \sum_{i\in\mathcal{U}} \frac{\delta_i(y_i - \mu_y)}{\pi_i^B} \\
&\approx \sum_{i\in\mathcal{U}} \frac{\delta_i(y_i - \mu_y)}{\pi_i^B N}
\end{aligned} \tag{B.26}$$

given that $E(\hat{N}_B(\boldsymbol{\theta}_0)) = N$. Using the quotient rule, $\partial\hat{\mu}/\partial\boldsymbol{\theta}$ can be expressed as

$$\begin{aligned}
\frac{\partial\hat{\mu}}{\partial\boldsymbol{\theta}} &= \sum_{i\in\mathcal{U}} \Big(\hat{Y}(\boldsymbol{\theta}) \frac{\delta_i \pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})^2} - \hat{N}(\boldsymbol{\theta}) \frac{\delta_i y_i \pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})^2} \Big) / \hat{N}(\boldsymbol{\theta})^2 \\
&= \frac{1}{\hat{N}(\boldsymbol{\theta})} \sum_{i\in\mathcal{U}} \frac{\delta_i \pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})^2} \Big(\hat{\mu}(\boldsymbol{\theta}) - y_i \Big) \\
&= \frac{-1}{\hat{N}(\boldsymbol{\theta})} \sum_{i\in\mathcal{U}} \frac{\pi'(\boldsymbol{x}_i;\boldsymbol{\theta})}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})^2} \delta_i \Big(y_i - \hat{\mu}(\boldsymbol{\theta}) \Big).
\end{aligned}$$

Evaluating $\partial\hat{\mu}/\partial\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$ gives

$$\begin{aligned}\hat{\mu}'(\boldsymbol{\theta}_0) &= \frac{-1}{\hat{N}(\boldsymbol{\theta}_0)} \sum_{i\in\mathcal{U}} \frac{\pi_i'^B}{(\pi_i^B)^2} \delta_i \Big(y_i - \hat{\mu}(\boldsymbol{\theta}_0)\Big) \\ &\xrightarrow{p} \frac{-1}{N} \sum_{i\in\mathcal{U}} \frac{\pi_i'^B}{\pi_i^B} \Big(y_i - \mu_y\Big). \end{aligned} \tag{B.27}$$

Expanding $S(\hat{\boldsymbol{\theta}})$ using a first-order Taylor expansion and applying similar reasoning to how (B.2) is obtained, we can express $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ approximately as $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \approx G^{-1}S(\boldsymbol{\theta}_0)$, with $G = -E(\partial S/\partial\boldsymbol{\theta})$. Combining this, along with (B.26) and (B.27), into (B.25), we obtain

$$\begin{aligned}\hat{\mu}(\hat{\boldsymbol{\theta}}) - \mu_y &\approx \sum_{i\in\mathcal{U}} \frac{\delta_i}{\pi_i^B}\Big(\frac{y_i - \mu_y}{N}\Big) - \frac{1}{N}\sum_{i\in\mathcal{U}} \frac{\pi_i'^B}{\pi_i^B}\Big(y_i - \mu_y\Big) G^{-1} S(\boldsymbol{\theta}_0) \\ &= \sum_{i\in\mathcal{U}} \Big[\frac{\delta_i}{\pi_i^B}\Big(\frac{y_i - \mu_y}{N}\Big) - \beta^T s_i\Big] \end{aligned} \tag{B.28}$$

where $\beta^T = 1/N \sum_{i\in\mathcal{U}} \pi_i'^B/\pi_i^B \Big(y_i - \mu_y\Big) G^{-1}$ and s_i are the individual unit i terms in $S(\boldsymbol{\theta}_0)$. Under the assumptions **A1** and **A3** from Section 1.1, and further that A and B are independent samples drawn with Poisson sampling,

$$Var(\hat{\mu}(\hat{\boldsymbol{\theta}})) \approx \sum_{i\in\mathcal{U}} Var\Big[\frac{\delta_i}{\pi_i^B}\Big(\frac{y_i - \mu_y}{N}\Big) - \beta^T s_i\Big] \tag{B.29}$$

B.5 Regularity conditions for split-population OE participation probabilities

The asymptotic framework described in Section 3.3 is here adopted.

It is assumed that the following regularity conditions hold:

C5 The finite population size N, the population size of $C = \mathcal{U} \setminus B$, and the sample sizes n_A and n_B satisfy $\lim_{N_C\to\infty} n_A/N_C = f_A \in (0,1)$ and $\lim_{N\to\infty} n_B/N = f_B \in (0,1)$

C6 There exist c_1 and c_2 such that $0 < c_1 \leq N_C\pi_i^A/n_A \leq c_2$ and $0 < c_1 \leq$

$N\pi_i^B/n_B \leq c_2$ for all units $i \in \mathcal{U}$

C7 Under the finite population and the sampling design for A, we have $N_C^{-1}\sum_{i\in A} d_i^A \boldsymbol{u}_i - N_C^{-1}\sum_{i\in C}\boldsymbol{u}_i = O_p(n_A^{-1/2})$ for $\boldsymbol{u}_i = \boldsymbol{x}_i, y_i$, where $d_i^A = 1/\pi_i^A$ and the order in probability is with respect to the sampling mechanism used to select A

C8 The finite population and the propensity scores satisfy $N^{-1}\sum_{i\in\mathcal{U}} y_i^2 = O(1)$, $N^{-1}\sum_{i\in\mathcal{U}}\|\boldsymbol{x}_i\|^3 = O(1)$, and $N^{-1}\sum_{i\in\mathcal{U}}\pi_i^B(1-\pi_i^B)\boldsymbol{x}_i\boldsymbol{x}_i^T$ is a positive definite matrix.

Condition **C5** means that the sample fractions of the probability and non-probability samples are bounded, while condition **C6** means that the inclusion probabilities for the probability and non-probability samples do not differ in terms of order of magnitude from simple random sampling. Condition **C7** ensures consistency of the Horvitz-Thompson estimators obtained from the probability sample. Condition **C8** are finite moment conditions to validate Taylor series expansions.

B.6 Consistency of split-population OE IPW estimator

Under the split-population approach, the probabilities of selection into the reference sample, π_i^A, depend on whether or not unit i is in B. In other words, $\pi_i^A = P(I_i = 1|\delta_i)$. In particular, $P(I_i = 1|\delta_i = 0) = \pi_i^A$, and $P(I_i = 1|\delta_i = 1) = 0$, so that

$$\begin{aligned} P(I_i = 1|\delta_i = 0) &= \pi_i^A \\ &= \frac{P(I_i = 1, \delta_i = 0)}{P(\delta_i = 0)} \\ &= \frac{P(\delta_i = 0|I_i = 1)P(I_i = 1)}{P(\delta_i = 0)} \\ &= \frac{P(I_i = 1)}{P(\delta_i = 0)} \\ \implies P(I_i = 1) = E(I_i) &= \pi_i^A P(\delta_i = 0) = \pi_i^A(1-\pi_i^B) \end{aligned} \tag{B.30}$$

where $P(\delta_i = 0|I_i = 1) = 1$ because if a unit is selected in A it must not have been in B originally.

To consider consistency of the split-population IPW estimator, the joint estimation problem is set up. Firstly, define $\boldsymbol{\eta}^T = (\mu, \boldsymbol{\theta}^T)$. The true (target) parameter values are $\boldsymbol{\eta}_0^T = (\mu_y, \boldsymbol{\theta}_0^T)$, while the split-population OE IPW estimate and propensity score model parameters are combined as $\hat{\boldsymbol{\eta}}_{SPOE}^T = (\hat{\mu}_{SPOE}, \hat{\boldsymbol{\theta}}_{SPOE}^T)$. $\hat{\boldsymbol{\eta}}_{SPOE}^T$ is the solution to the combined estimating equation system

$$\begin{aligned}\boldsymbol{\Phi}_{SPOE}(\boldsymbol{\eta}) &= \begin{pmatrix} U(\mu, \boldsymbol{\theta}) \\ S_{SPOE}(\boldsymbol{\theta}) \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{N}\sum_{i\in\mathcal{U}} \frac{\delta_i(y_i-\mu)}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta})} \\ \frac{1}{N}\sum_{i\in\mathcal{U}}\Big\{\delta_i\boldsymbol{x}_i - \delta_i\pi(\boldsymbol{x}_i;\boldsymbol{\theta})\boldsymbol{x}_i\Big\} - \frac{1}{N}\sum_{i\in\mathcal{U}} I_i(1-\delta_i)d_i^A\pi(\boldsymbol{x}_i;\boldsymbol{\theta})\boldsymbol{x}_i \end{pmatrix} = \mathbf{0}\end{aligned}$$

We consider the joint randomisation of the propensity score model and the sample design for A. From (B.30) it is given that under the split-population approach $E(I_i) = \pi_i^A(1-\pi_i^B)$. Note that since the sample A is drawn from $C = \mathcal{U} \setminus B$, $E(I_i\delta_i) = 0$. As such, when Assumptions **A1**, **A2**, and the propensity score model all hold, then

$$E(U(\mu_y, \boldsymbol{\theta}_0)) = \frac{1}{N}\sum_{i\in\mathcal{U}} \frac{E(\delta_i)(y_i-\mu_y)}{\pi_i^B} = \frac{1}{N}\sum_{i\in\mathcal{U}}(y_i-\mu_y) = 0$$

and

$$\begin{aligned}E(S_{SPOE}(\boldsymbol{\theta}_0)) &= \frac{1}{N}\sum_{i\in\mathcal{U}}\Big\{E(\delta_i)\boldsymbol{x}_i - E(\delta_i)\pi_i^B\boldsymbol{x}_i\Big\} - \frac{1}{N}\sum_{i\in\mathcal{U}}\Big(E(I_i) - E(I_i\delta_i)\Big)d_i^A\pi_i^B\boldsymbol{x}_i \\ &= \frac{1}{N}\sum_{i\in\mathcal{U}}\Big\{\pi_i^B\boldsymbol{x}_i - (\pi_i^B)^2\boldsymbol{x}_i\Big\} - \frac{1}{N}\sum_{i\in\mathcal{U}}\pi_i^A(1-\pi_i^B)\frac{1}{\pi_i^A}\pi_i^B\boldsymbol{x}_i \\ &= \mathbf{0}\end{aligned}$$

and so $E\{\boldsymbol{\Phi}_{SPOE}(\boldsymbol{\eta}_0)\} = \mathbf{0}$.

Note that $\boldsymbol{\Phi}_{SPOE}(\hat{\boldsymbol{\eta}}) = \mathbf{0}$. Under the regularity conditions in Section B.5, we apply a first-order Taylor expansion to obtain

$$\mathbf{0} = \boldsymbol{\Phi}_{SPOE}(\hat{\boldsymbol{\eta}}) \approx \boldsymbol{\Phi}_{SPOE}(\boldsymbol{\eta}_0) + \boldsymbol{\phi}_{SPOE}(\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \tag{B.31}$$

where $\boldsymbol{\phi}_{SPOE}(\boldsymbol{\eta})$ is defined as $\partial \boldsymbol{\Phi}_{SPOE}(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$, and is given by

$$\boldsymbol{\phi}(\boldsymbol{\eta}) = \begin{pmatrix} \frac{\partial U}{\partial \mu} & \frac{\partial U}{\partial \boldsymbol{\theta}} \\ \frac{\partial S_{SPOE}}{\partial \mu} & \frac{\partial S_{SPOE}}{\partial \boldsymbol{\theta}} \end{pmatrix}$$

Under suitable regularity conditions as $n \to \infty$ (see Stefanski and Boos 2002), we have

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \boldsymbol{\Phi}_{SPOE}(\boldsymbol{x}_i, y_i, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \xrightarrow{p} E\Big\{\frac{\partial \boldsymbol{\Phi}_{SPOE}(\boldsymbol{x}, y, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\Big\}$$

We can then re-arrange (B.31) to obtain

$$\begin{aligned} \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 &\approx -[\boldsymbol{\phi}_{SPOE}(\boldsymbol{\eta}_0)]^{-1} \boldsymbol{\Phi}_{SPOE}(\boldsymbol{\eta}_0) \\ &\approx -[E(\boldsymbol{\phi}_{SPOE}(\boldsymbol{\eta}_0))]^{-1} \boldsymbol{\Phi}_{SPOE}(\boldsymbol{\eta}_0) \end{aligned} \tag{B.32}$$

Now $E\{\boldsymbol{\Phi}_{SPOE}(\boldsymbol{\eta}_0)\} = \mathbf{0}$, and drawing on the standard M-estimation argument of Stefanski and Boos (2002), we conclude that $E(\hat{\boldsymbol{\eta}})$ converges to $\boldsymbol{\eta}_0$ asymptotically, that is, $\hat{\boldsymbol{\eta}} \xrightarrow{p} \boldsymbol{\eta}_0$ as the population and sample sizes tend to infinity.

B.7 Derivation of kernel score function for the SNAR case

We derive the kernel score function for the case when selection into B is considered SNAR. We begin by considering the case where $\boldsymbol{x}$ are available for the full population, for example through a population census or administrative data. This scenario is akin to the nonignorable non-response problem in survey inference, of which there is a sizeable body of previous work; see for example, Kim and Yu (2011), Riddles et al. (2016), Shao and Wang (2016), and Morikawa and Kim (2021). In the notation in this section, y is the variable of interest, and $\boldsymbol{x}$ are the auxiliary variables. We can segregate $\boldsymbol{x}$ into two parts, $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, with $P(\delta = 1|\boldsymbol{x}, y) = P(\delta = 1|\boldsymbol{x}_1, y) = \pi(\boldsymbol{x}_1, y, \boldsymbol{\theta})$. In other words, $\boldsymbol{x}_2$ does not affect the participation probability model.

The population observed likelihood function for π_i^B is

$$L(\boldsymbol{\theta}) = \prod_{i \in \mathcal{U}} \pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})^{\delta_i} \Big\{ \int (1 - \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})) f(y|\boldsymbol{x}_i) \mathrm{d}y \Big\}^{1-\delta_i} \tag{B.33}$$

The log-likelihood function is given by

$$l(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \delta_i \log \pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta}) + \sum_{i \in \mathcal{U}} (1 - \delta_i) \log \Big\{ \int (1 - \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})) f(y|\boldsymbol{x}_i) \mathrm{d}y \Big\} \tag{B.34}$$

and the corresponding score function is

$$s(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \delta_i \frac{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})'}{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})} - \sum_{i \in \mathcal{U}} (1 - \delta_i) \frac{\int \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})' f(y|\boldsymbol{x}_i) \mathrm{d}y}{\int (1 - \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})) f(y|\boldsymbol{x}_i) \mathrm{d}y} \tag{B.35}$$

where $\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})' = \partial \pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$. We do not know $f(y|\boldsymbol{x})$, the distribution of y given $\boldsymbol{x}$. However, we do know $f(y|\boldsymbol{x}, \delta = 1) = f_1(y|\boldsymbol{x})$. By using Bayes' Theorem and manipulating, we can formulate the relation

$$\begin{aligned} f_1(y|\boldsymbol{x}) &= \frac{f(y, \delta = 1|\boldsymbol{x})}{P(\delta = 1|\boldsymbol{x})} \\ &= \frac{P(\delta = 1|y, \boldsymbol{x}) f(y|\boldsymbol{x})}{P(\delta = 1|\boldsymbol{x})} \\ \implies f(y|\boldsymbol{x}) &= \frac{f_1(y|\boldsymbol{x}) P(\delta = 1|\boldsymbol{x})}{P(\delta = 1|y, \boldsymbol{x})} \end{aligned} \tag{B.36}$$

and substituting this into (B.35), we obtain

$$\begin{aligned} s(\boldsymbol{\theta}) = & \sum_{i \in \mathcal{U}} \delta_i \frac{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})'}{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})} - \\ & \sum_{i \in \mathcal{U}} (1 - \delta_i) \frac{\int \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})' f_1(y|\boldsymbol{x}_i) P(\delta_i = 1|\boldsymbol{x}_i) P(\delta_i = 1|y, \boldsymbol{x}_i)^{-1} \mathrm{d}y}{\int (1 - \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})) f_1(y|\boldsymbol{x}_i) P(\delta_i = 1|\boldsymbol{x}_i) P(\delta_i = 1|y, \boldsymbol{x}_i)^{-1} \mathrm{d}y} \\ = & \sum_{i \in \mathcal{U}} \delta_i \frac{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})'}{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})} - \sum_{i \in \mathcal{U}} (1 - \delta_i) \frac{\int \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})' (1 - \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta}))^{-1} O(\boldsymbol{x}_i, y) f_1(y|\boldsymbol{x}_i) \mathrm{d}y}{\int O(\boldsymbol{x}_i, y) f_1(y|\boldsymbol{x}_i) \mathrm{d}y} \end{aligned} \tag{B.37}$$

where $O(\boldsymbol{x}_i, y) = P(\delta_i = 0|\boldsymbol{x}_i, y) / P(\delta_i = 1|\boldsymbol{x}_i, y)$ and the $P(\delta_i = 1|\boldsymbol{x}_i)$ terms can be taken

out of the integrals and cancel out. Note that $O(\boldsymbol{x}_i, y) = O(\boldsymbol{x}_{1i}, y)$.

In order to find a solution to $s(\boldsymbol{\theta}) = 0$, we require a way to determine the integrals in the second term of (B.37). Both the numerator and denominator can be expressed as expectations, that is,

$$\int \frac{\pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})'}{1 - \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})} O(\boldsymbol{x}_{1i}, y) f_1(y|\boldsymbol{x}_i) \mathrm{d}y = E_1\Big(\frac{\pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})'}{1 - \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})} O(\boldsymbol{x}_{1i}, y)\Big|\boldsymbol{x}_i\Big) \tag{B.38}$$

$$\int O(\boldsymbol{x}_{1i}, y) f_1(y|\boldsymbol{x}_i) \mathrm{d}y = E_1(O(\boldsymbol{x}_{1i}, y)|\boldsymbol{x}_i) \tag{B.39}$$

To estimate these expectations, we employ the same approach outlined in Section 5.3. The right hand side of (B.38) and (B.39) can be replaced using

$$E_1\Big(\frac{\pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})'}{1 - \pi(\boldsymbol{x}_{1i}, y, \boldsymbol{\theta})} O(\boldsymbol{x}_{1i}, y)\Big|\boldsymbol{x}_i\Big) = \frac{\sum_{j \in B} K_h(\boldsymbol{x}_i, \boldsymbol{x}_j) \pi(\boldsymbol{x}_{1i}, y_j, \boldsymbol{\theta})' (1 - \pi(\boldsymbol{x}_{1i}, y_j, \boldsymbol{\theta}))^{-1} O(\boldsymbol{x}_{1i}, y_j)}{\sum_{j \in B} K_h(\boldsymbol{x}_i, \boldsymbol{x}_j)} \tag{B.40}$$

$$E_1(O(\boldsymbol{x}_{1i}, y)|\boldsymbol{x}_i) = \frac{\sum_{j \in B} K_h(\boldsymbol{x}_i, \boldsymbol{x}_j) O(\boldsymbol{x}_{1i}, y_j)}{\sum_{j \in B} K_h(\boldsymbol{x}_i, \boldsymbol{x}_j)} \tag{B.41}$$

The score function (B.37) thus becomes

$$\begin{aligned} s(\boldsymbol{\theta}) &= \sum_{i \in \mathcal{U}} \delta_i \frac{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})'}{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})} - \\ &\quad \sum_{i \in \mathcal{U}} (1 - \delta_i) \frac{\sum_{j \in B} K_h(\boldsymbol{x}_i, \boldsymbol{x}_j) \pi(\boldsymbol{x}_{1i}, y_j, \boldsymbol{\theta})' (1 - \pi(\boldsymbol{x}_{1i}, y_j, \boldsymbol{\theta}))^{-1} O(\boldsymbol{x}_{1i}, y_j)}{\sum_{j \in B} K_h(\boldsymbol{x}_i, \boldsymbol{x}_j) O(\boldsymbol{x}_{1i}, y_j)} \\ &= \sum_{i \in \mathcal{U}} \delta_i \frac{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})'}{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})} - \sum_{i \in \mathcal{U}} (1 - \delta_i) \sum_{j \in B} w_{ij} \pi(\boldsymbol{x}_{1i}, y_j, \boldsymbol{\theta})' (1 - \pi(\boldsymbol{x}_{1i}, y_j, \boldsymbol{\theta}))^{-1} \end{aligned} \tag{B.42}$$

where

$$w_{ij} = \frac{K_h(\boldsymbol{x}_i, \boldsymbol{x}_j) O(\boldsymbol{x}_{1i}, y_j)}{\sum_{l \in B} K_h(\boldsymbol{x}_i, \boldsymbol{x}_l) O(\boldsymbol{x}_{1i}, y_l)} \tag{B.43}$$

When A is a probability sample, we can estimate the sum over $\mathcal{U}$ using data from A, and

the sum over B using data from $A \cap B$. The resulting score function is

$$\hat{s}_n(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \delta_i \frac{\pi(\boldsymbol{x}_{1i}, y_i; \boldsymbol{\theta})'}{\pi(\boldsymbol{x}_{1i}, y_i; \boldsymbol{\theta})} - \sum_{i \in A} d_i^A (1 - \delta_i) \sum_{j \in A \cap B} \pi(\boldsymbol{x}_{1i}, y_j; \boldsymbol{\theta})' (1 - \pi(\boldsymbol{x}_{1i}, y_j; \boldsymbol{\theta}))^{-1} w_{ij}$$
$$\text{where } w_{ij} = \frac{d_j^A O(\boldsymbol{x}_{1i}, y_j) K_h(\boldsymbol{x}_j, \boldsymbol{x}_i)}{\sum_{k \in A \cap B} d_k^A O(\boldsymbol{x}_{1i}, y_k) K_h(\boldsymbol{x}_k, \boldsymbol{x}_i)}$$

where $d_i^A = 1$ and $A \cap B = B$ if A comprises the full population.

Remark B.1. Expression (B.37) is equivalent to the mean score equation used in, for example Riddles et al. (2016), Morikawa et al. (2017), and Yang and Tang (2020) to address MNAR issues. To see this, we first note that

$$\begin{aligned} f_0(y|\boldsymbol{x}) &= \frac{P(\delta = 0|y, \boldsymbol{x}) f(y|\boldsymbol{x})}{P(\delta = 0|\boldsymbol{x})} \\ &= \frac{(1 - Pr(\delta = 1|y, \boldsymbol{x})) f(y|\boldsymbol{x})}{\int Pr(\delta = 0|y, \boldsymbol{x}) f(y|\boldsymbol{x}) \mathrm{d}y} \\ &= \frac{(1 - \pi(\boldsymbol{x}_1, y, \boldsymbol{\theta})) f(y|\boldsymbol{x})}{\int (1 - \pi(\boldsymbol{x}_1, y, \boldsymbol{\theta})) f(y|\boldsymbol{x}) \mathrm{d}y} \end{aligned}$$

and then substituting in (B.36), we have

$$\begin{aligned} f_0(y|\boldsymbol{x}) &= \frac{(1 - \pi(\boldsymbol{x}_1, y, \boldsymbol{\theta})) f_1(y|\boldsymbol{x}) P(\delta = 1|\boldsymbol{x}) / P(\delta = 1|y, \boldsymbol{x})}{\int (1 - \pi(\boldsymbol{x}_1, y, \boldsymbol{\theta})) f_1(y|\boldsymbol{x}) P(\delta = 1|\boldsymbol{x}) / P(\delta = 1|y, \boldsymbol{x}) \mathrm{d}y} \\ &= \frac{f_1(y|\boldsymbol{x}) O(\boldsymbol{x}, y)}{\int O(\boldsymbol{x}, y) f_1(y|\boldsymbol{x}) \mathrm{d}y} \\ \implies f_1(y|\boldsymbol{x}) &= \frac{f_0(y|\boldsymbol{x}) \int O(\boldsymbol{x}, y) f_1(y|\boldsymbol{x}) \mathrm{d}y}{O(\boldsymbol{x}, y)} \end{aligned} \tag{B.44}$$

Let $\tilde{s}(\boldsymbol{\theta}; \delta_i, \boldsymbol{x}_{1i}, y_i) = \{\delta_i - \pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})\} \pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})' \pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})^{-1} (1 - \pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta}))^{-1}$ be the score expression from Riddles et al. (2016). Note that $\delta_i = 1$ when $i \in B$ and $\delta_i = 0$ when $i \notin B$. Using (B.44), the score function (B.37) can be expressed as

$$\begin{aligned} s(\boldsymbol{\theta}) &= \sum_{i \in \mathcal{U}} \delta_i \frac{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})'}{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})} - \sum_{i \in \mathcal{U}} (1 - \delta_i) \int \frac{\pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta})' f_0(y|\boldsymbol{x}_i)}{(1 - \pi(\boldsymbol{x}_{1i}, y_i, \boldsymbol{\theta}))} \mathrm{d}y \\ &= \sum_{i \in \mathcal{U}} \delta_i \tilde{s}(\boldsymbol{\theta}; \delta_i, \boldsymbol{x}_{1i}, y_i) + \sum_{i \in \mathcal{U}} (1 - \delta_i) \int \tilde{s}(\boldsymbol{\theta}; \delta_i, \boldsymbol{x}_{1i}, y_i) f_0(y|\boldsymbol{x}_i) \mathrm{d}y \\ &= \sum_{i \in \mathcal{U}} \delta_i \tilde{s}(\boldsymbol{\theta}; \delta_i, \boldsymbol{x}_{1i}, y_i) + \sum_{i \in \mathcal{U}} (1 - \delta_i) E_0(\tilde{s}(\boldsymbol{\theta}; \delta_i, \boldsymbol{x}_{1i}, y_i) | \boldsymbol{x}_i) \end{aligned} \tag{B.45}$$

which is equivalent to the mean score function described in Equation (5) of Riddles et al. (2016).

B.8 Conditions for consistency of the kernel score function

C9 The kernel $K(u)$ is differentiable to order d and those derivatives are bounded, $\int K(u)\mathrm{d}u = 1$, and there is a positive integer m such that for $j < m$, $K(u)$ has zero moments and a non-zero m'th order moment.

C10 For all $\boldsymbol{x}^*$, the functions $\pi(\cdot, \boldsymbol{x}^*; \boldsymbol{\theta}_0)$ and $\pi'(\cdot, \boldsymbol{x}^*; \boldsymbol{\theta}_0)$ are continuously differentiable to order d with bounded derivatives on an open set containing $\tilde{\mathcal{X}}$, where $\tilde{\mathcal{X}}$ is a compact set contained in the support of $\tilde{\boldsymbol{x}}$.

C11 There is $p \geq 4$ such that $E_1\Big[\|O(\boldsymbol{X}, \boldsymbol{z}; \boldsymbol{\theta}_0)\|^p\Big]$, $E_1\Big[\|O(\boldsymbol{X}, \boldsymbol{z}; \boldsymbol{\theta}_0|\tilde{\boldsymbol{x}})\|^p\Big] f_1(\tilde{\boldsymbol{x}})$, $E_1\Big[\|\pi'(\boldsymbol{X}, \boldsymbol{z}; \boldsymbol{\theta}_0)(1 - \pi(\boldsymbol{X}, \boldsymbol{z}; \boldsymbol{\theta}_0))^{-1} O(\boldsymbol{X}, \boldsymbol{z}; \boldsymbol{\theta}_0)\|^p\Big]$ and $E_1\Big[\|\pi'(\boldsymbol{X}, \boldsymbol{z}; \boldsymbol{\theta}_0)(1 - \pi(\boldsymbol{X}, \boldsymbol{z}; \boldsymbol{\theta}_0))^{-1} O(\boldsymbol{X}, \boldsymbol{z}; \boldsymbol{\theta}_0|\tilde{\boldsymbol{x}})\|^p\Big] f_1(\tilde{\boldsymbol{x}})$ are bounded

C12 $n^{1-(2/p)} h^r / \ln n \to \infty$, $\sqrt{n} \ln n / (n h^{r+2d}) \to 0$, and $\sqrt{n} h^{2m} \to 0$, where n is the sample size feeding into the kernel estimation, r is the dimension of $\boldsymbol{x}$, and d, m and p are defined in Conditions **C9** and **C11**.

C13 The finite population size N and the sample sizes n_A satisfy $\lim_{N\to\infty} n_A/N = f_A \in (0, 1)$

C14 There exist c_1 and c_2 such that $0 < c_1 \leq N\pi_i^A / n_A \leq c_2$ for all units $i \in \mathcal{U}$.

C15 Under the finite population and the sampling design for A, we have $N^{-1} \sum_{i\in A} d_i^A \boldsymbol{u}_i - N^{-1} \sum_{i\in\mathcal{U}} \boldsymbol{u}_i = O_p(n_A^{-1/2})$, where $d_i^A = 1/\pi_i^A$ and the order in probability is with respect to the sampling mechanism used to select A.

C16 For the sample design of A, $\max_{i\neq j} |\pi_{ij} - \pi_i \pi_j| \leq c n_A^{-1}$, and $\pi_{ij} \geq \lambda > 0$.

C17 For every integer $t \geq 1$, let $D_{t,N} = \{(i_1, i_2, \ldots, i_t) \ : \ i_1, i_2, \ldots, i_t$ are all different and each $i_j \in \{1, 2, \ldots, N\}\}$. Then for the sample design of A with I_i being

the indicator variable of selection in A,

$$\limsup_{N\to\infty} \frac{N^4}{n_A^2} \max_{(i,j,k,l)\in D_{4,N}} |E(I_i - \pi_i)(I_j - \pi_j)(I_k - \pi_k)(I_l - \pi_l)| < \infty,$$
$$\limsup_{N\to\infty} \frac{N^3}{n_A^2} \max_{(i,j,k)\in D_{3,N}} |E(I_i - \pi_i)^2(I_j - \pi_j)(I_k - \pi_k)| < \infty.$$

Conditions **C9** to **C11** relate to Assumptions 8.1 to 8.3 of Newey and McFadden (1994); see also Morikawa et al. (2017). Condition **C9** is important for controlling the bias of the kernel estimators, and ensures they are well defined. Condition **C10** imposes a level of smoothness in the function being estimated, while Condition **C11** is imposed to enable a uniform convergence rate for the kernel estimator. Condition **C12** facilitates consistency of the kernel estimator. Condition **C13** means that the sample fractions of the probability sample are bounded, while condition **C14** means that the inclusion probabilities for the probability sample do not differ in terms of order of magnitude from simple random sampling. Condition **C15** ensures consistency of the Horvitz-Thompson estimators obtained from the probability sample. Condition **C16** means that asymptotically there is near independence between the selection probabilities of units in the sample. This condition will hold for many regular sample designs like SRSWOR, stratified SRS, or high-entropy unequal probability designs, and will be met exactly for Poisson sampling. Condition **C17**, originally from Breidt and Opsomer (2000) and in the form given in Boistard et al. (2012), bounds the third and fourth-order correlations of the sampling indicators.

B.9 Proof of consistency of the survey-weighted kernel score function

We want to examine the rate of convergence of $1/N\|\hat{s}_p(\boldsymbol{\theta}) - \hat{s}(\boldsymbol{\theta})\|$. Assume that the conditions in Appendix B.8 hold. Let

$$f_{ij} = \pi'(\boldsymbol{x}_j, \boldsymbol{z}_i; \boldsymbol{\theta})(1 - \pi(\boldsymbol{x}_j, \boldsymbol{z}_i; \boldsymbol{\theta}))^{-1} O(\boldsymbol{x}_j, \boldsymbol{z}_i) K_h(\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{x}}_i),$$
$$g_{ik} = O(\boldsymbol{x}_k, \boldsymbol{z}_i) K_h(\tilde{\boldsymbol{x}}_k, \tilde{\boldsymbol{x}}_i).$$

Then from (5.8) and (5.9), $\hat{s}_p(\boldsymbol{\theta}) - \hat{s}(\boldsymbol{\theta})$ can be expressed as

$$\hat{s}_p(\boldsymbol{\theta}) - \hat{s}(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} (1 - \delta_i) \frac{\sum_{j \in \mathcal{U}} \delta_j f_{ij}}{\sum_{k \in \mathcal{U}} \delta_k g_{ik}} - \sum_{i \in A} d_i^A (1 - \delta_i) \frac{\sum_{j \in A} d_j^A \delta_j f_{ij}}{\sum_{k \in A} d_k^A \delta_k g_{ik}}.$$

Setting

$$r_i = \frac{1/N \sum_{j \in \mathcal{U}} \delta_j f_{ij}}{1/N \sum_{k \in \mathcal{U}} \delta_k g_{ik}} \text{ and } \hat{r}_i = \frac{1/N \sum_{j \in A} d_j^A \delta_j f_{ij}}{1/N \sum_{k \in A} d_k^A \delta_k g_{ik}}, \tag{B.46}$$

and further letting $a_i = I_i / \pi_i^A - 1 = d_i^A I_i - 1$, we get

$$\begin{aligned}
\frac{\hat{s}_p(\boldsymbol{\theta}) - \hat{s}(\boldsymbol{\theta})}{N} &= \frac{1}{N} \sum_{i \in \mathcal{U}} (1 - \delta_i) r_i - \frac{1}{N} \sum_{i \in A} d_i^A (1 - \delta_i) \hat{r}_i \\
&= \frac{1}{N} \sum_{i \in \mathcal{U}} (1 - \delta_i) r_i - \frac{1}{N} \sum_{i \in \mathcal{U}} (a_i + 1)(1 - \delta_i) \hat{r}_i \\
&= \frac{1}{N} \sum_{i \in \mathcal{U}} (1 - \delta_i) [(a_i + 1)(r_i - \hat{r}_i) - a_i r_i] \\
&= \frac{1}{N} \sum_{i \in \mathcal{U}} (1 - \delta_i) a_i (r_i - \hat{r}_i) - \frac{1}{N} \sum_{i \in \mathcal{U}} (1 - \delta_i)(\hat{r}_i - r_i) - \frac{1}{N} \sum_{i \in \mathcal{U}} (1 - \delta_i) a_i r_i.
\end{aligned} \tag{B.47}$$

The third term in (B.47) can be seen to be the error from a Horvitz-Thompson estimator, and by **C15** is $O_p(n_A^{-1/2})$. Noting that the second term involves the difference between a Horvitz-Thompson estimate of a ratio and its population value, using standard results (see for example Fuller 2009) this term is also $O_p(n_A^{-1/2})$.

To examine the first term, first expand $(r_i - \hat{r}_i)$ using a first-order Taylor series expansion of the form

$$\begin{aligned}
r_i - \hat{r}_i &= \frac{X_i}{Y_i} - \frac{\hat{X}_i}{\hat{Y}_i} \\
&= \frac{X_i}{Y_i^2} (\hat{Y}_i - Y_i) - \frac{1}{Y_i} (\hat{X}_i - X_i) \\
&= \hat{Y}_i \frac{X_i}{Y_i^2} - \frac{\hat{X}_i}{Y_i}
\end{aligned}$$

where X_i, Y_i, $\hat{X}_i$ and $\hat{Y}_i$ are the numerator and denominator terms in r_i and $\hat{r}_i$ respectively.

This gives

$$\frac{\hat{s}_p(\boldsymbol{\theta}) - \hat{s}(\boldsymbol{\theta})}{N} = \frac{1}{N}\sum_{i\in\mathcal{U}} a_i(1-\delta_i)\Big(\hat{Y}_i\frac{X_i}{Y_i^2} - \frac{\hat{X}_i}{Y_i}\Big) + O_p(n_A^{-1/2}). \tag{B.48}$$

The main terms on the right hand side of (B.48) are of the form

$$\begin{aligned}\frac{1}{N}k\sum_{i\in\mathcal{U}} a_i\frac{1}{N}\sum_{j\in\mathcal{U}}(1+a_j)h_{ij} &= \frac{k}{N^2}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}}(a_i + a_ia_j)h_{ij} \\ &= \frac{k}{N^2}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}} a_ih_{ij} + \frac{k}{N^2}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}} a_ia_jh_{ij}\end{aligned} \tag{B.49}$$

where h_{ij} may be $(1-\delta_i)f_{ij}$ or $(1-\delta_i)g_{ij}$ and k is a quantity that does not depend on i. The first term in (B.49) is in the form of the error from a Horvitz-Thompson estimator, and by **C15** the term is $O_p(n_A^{-1/2})$.

By **C14** and **C16** the expectation of the second term in (B.49) is $O_p(n_A^{-1})$. By **C17** the variance of the second term will be $O_p(n_A^{-2})$, and so the expectation of the square of the second term will also be $O_p(n_A^{-2})$. Applying Markov's inequality, the second term is thus $O_p(n_A^{-1})$, so (B.49) is $O_p(n_A^{-1})$. Feeding this back into (B.48) we get

$$\frac{1}{N}\|\hat{s}_p(\boldsymbol{\theta}) - \hat{s}(\boldsymbol{\theta})\| = O_p(n_A^{-1/2}).$$

References

Andridge, R. R., B. T. West, R. J. A. Little, P. S. Boonstra, and F. Alvarado-Leiton. 2019. "Indices of Non-Ignorable Selection Bias for Proportions Estimated from Non-Probability Samples." *Journal of the Royal Statistical Society Series C: Applied Statistics* 68 (5): 1465–1483. https://doi.org/10.1111/rssc.12371.

Antal, E., and Y. Tillé. 2011. "A Direct Bootstrap Method for Complex Sampling Designs From a Finite Population." *Journal of the American Statistical Association* 106, no. 494 (June): 534–543. https://doi.org/10.1198/jasa.2011.tm09767.

Australian Bureau of Statistics. 2021. *Microdata: Economic Activity Survey 2020-2021.* ABS DataLab.

———. 2022a. "Average Weekly Earnings, Australia, November 2021." (Accessed: February 9, 2023). Available at: https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/average-weekly-earnings-australia/nov-2021.

———. 2022b. "Monthly Business Turnover Indicator Methodology, October 2022 — Australian Bureau of Statistics." (Accessed: December 22, 2022). Available at: https://www.abs.gov.au/methodologies/monthly-business-turnover-indicator-methodology/oct-2022.

Beaumont, J.-F. 2020. "Are Probability Surveys Bound to Disappear for the Production of Official Statistics?" *Survey Methodology* 46 (1): 1–28.

Beaumont, J.-F., K. Bosa, A. Brennan, J. Charlebois, and K. Chu. 2024a. "Authors' Response to Comments on "Handling Non-Probability Samples through Inverse Probability Weighting with an Application to Statistics Canada's Crowdsourcing Data": Some New Developments on Likelihood Approaches to Estimation of Participation Probabilities for Non-Probability Samples." *Survey Methodology* 50 (1): 123–141.

———. 2024b. "Handling Non-Probability Samples through Inverse Probability Weighting with an Application to Statistics Canada's Crowdsourcing Data." *Survey Methodology* 50 (1): 77–106.

Berg, E., S. Chen, and C. Yu. 2025. "Combining Probability and Non-probability Samples Using Semi-parametric Quantile Regression and a Nonparametric Estimator of the Participation Probability." *Scandinavian Journal of Statistics,* 2128–2151. https://doi.org/10.1111/sjos.70020.

Bethel, J. 1989. "Sample Allocation in Multivariate Surveys." *Survey Methodology* 15 (1): 47–57.

Bethlehem, J. 2010. "Selection Bias in Web Surveys." *International Statistical Review* 78 (2): 161–188. https://doi.org/10.1111/j.1751-5823.2010.00112.x.

Boistard, H., H. P. Lopuhaä, and A. Ruiz-Gazen. 2012. "Approximation of Rejective Sampling Inclusion Probabilities and Application to High Order Correlations." *Electronic Journal of Statistics* 6:1967–1983. https://doi.org/10.1214/12-EJS736.

Boonstra, P. S., R. J. Little, B. T. West, R. R. Andridge, and F. Alvarado-Leiton. 2021. "A Simulation Study of Diagnostics for Selection Bias." *Journal of Official Statistics* 37 (3): 751–769. https://doi.org/10.2478/jos-2021-0033.

Booth, J. G., R. W. Butler, and P. Hall. 1994. "Bootstrap Methods for Finite Populations." *Journal of the American Statistical Association* 89 (428): 1282–1289. https://doi.org/10.2307/2290991.

Breidt, F. J., and J. D. Opsomer. 2000. "Local Polynomial Regression Estimators in Survey Sampling." *The Annals of Statistics* 28 (4): 1026–1053. https://doi.org/10.1214/aos/1015956706.

Breidt, F. J., and J. D. Opsomer. 2017. "Model-Assisted Survey Estimation with Modern Prediction Techniques." *Statistical Science* 32 (2): 190–205. https://doi.org/10.1214/16-STS589.

Burakauskaitė, I., and A. Čiginas. 2023. "An Approach to Integrating a Non-Probability Sample in the Population Census." *Mathematics* 11 (8): 1782. https://doi.org/10.3390/math11081782.

Business Longitudinal Analysis Data Environment (BLADE). 2020. *Microdata: Pay-As-You-Go.* Dataset. ABS DataLab.

———. 2022. *Microdata: Pay-As-You-Go.* Dataset. ABS DataLab.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective.* 2nd ed. Monographs on Statistics and Applied Probability 105. Boca Raton, FL: Chapman & Hall/CRC.

Castro-Martín, L., M. d. M. Rueda, and R. Ferri-García. 2022. "Combining Statistical Matching and Propensity Score Adjustment for Inference from Non-Probability Surveys." *Journal of Computational and Applied Mathematics* 404. https://doi.org/10.1016/j.cam.2021.113414.

Chen, S., and D. Haziza. 2023. "General Purpose Multiply Robust Data Integration Procedures for Handling Nonprobability Samples." *Scandinavian Journal of Statistics,* 697–724. https://doi.org/10.1111/sjos.12605.

Chen, S., S. Yang, and J. K. Kim. 2022. "Nonparametric Mass Imputation for Data Integration." *Journal of Survey Statistics and Methodology* 10 (1): 1–24. https://doi.org/10.1093/jssam/smaa036.

Chen, Y., P. Li, and C. Wu. 2020. "Doubly Robust Inference With Nonprobability Survey Samples." *Journal of the American Statistical Association* 115 (532): 2011–2021. https://doi.org/10.1080/01621459.2019.1677241.

———. 2023. "Dealing with Undercoverage for Non-Probability Survey Samples." *Survey Methodology* 49 (2): 497–515.

Chipperfield, J., J. Chessman, and R. Lim. 2012. "Combining Household Surveys Using Mass Imputation to Estimate Population Totals." *Australian & New Zealand Journal of Statistics* 54 (2): 223–238. https://doi.org/10.1111/j.1467-842X.2012.00666.x.

Chromy, J. 1987. "Design Optimization With Multiple Objectives." In *Proceedings of the Section on Survey Research Methods,* 194–199. American Statistical Association.

Clark, R. G., and D. G. Steel. 2022. "Sample Design for Analysis Using High-Influence Probability Sampling." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 185 (4): 1733–1756. https://doi.org/10.1111/rssa.12916.

Cochran, W. G. 1977. *Sampling Techniques.* 3rd Edition. John Wiley & Sons, Inc.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1): 1–38. JSTOR: 2984875.

Deville, J.-C., and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87 (418): 376–382. https://doi.org/10.2307/2290268.

Dharma, C., P. Smith, T. Salway, D. Gesink, M. Escobar, and V. Landsman. 2025. "A Two-Step Approach to Simultaneously Correct for Selection and Misclassification Bias in Nonprobability Samples from Hard-to-Reach Populations." *American Journal of Epidemiology* 194 (11): 3267–3272. https://doi.org/10.1093/aje/kwaf132.

Elisson, H., and E. Elvers. 2001. "Cut-off Sampling and Estimation." *Proceedings of Statistics Canada Symposium 2001,* 8.

Elliot, M. R. 2009. "Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights." *Survey Practice* 2 (6): 1–7. https://doi.org/10.29115/SP-2009-0025.

Elliott, M. R., and R. Valliant. 2017. "Inference for Nonprobability Samples." *Statistical Science* 32 (2): 249–264. https://doi.org/10.1214/16-STS598.

Ferri-García, R., and M. d. M. Rueda. 2020. "Propensity Score Adjustment Using Machine Learning Classification Algorithms to Control Selection Bias in Online Surveys." Edited by S. Marchetti. *PLOS ONE* 15 (4): e0231500. https://doi.org/10.1371/journal.pone.0231500.

Fuller, W. A. 2009. *Sampling Statistics.* Hoboken, New Jersey: John Wiley & Sons, Inc.

———. 1987. *Measurement Error Models.* Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Gershunskaya, J., and V. Beresovsky. 2024. "Comments on "Handling Non-Probability Samples through Inverse Probability Weighting with an Application to Statistics Canada's Crowdsourcing Data"." *Survey Methodology* 50 (1): 107–115.

Golini, N., and P. Righi. 2024. "Integrating Probability and Big Non-Probability Samples Data to Produce Official Statistics." *Statistical Methods & Applications* 33:555–580. https://doi.org/10.1007/s10260-023-00740-y.

Hammon, A., and S. Zinn. 2024. "Validating an Index of Selection Bias for Proportions in Non-Probability Samples." *International Statistical Review,* https://doi.org/10.1111/insr.12590.

Hartley, H. O. 1962. "Multiple Frame Surveys." In *Proceedings of the Social Statistics Section,* 203–206. Alexandria, VA: American Statistical Association.

———. 1974. "Multiple Frame Methodology and Selected Applications." *Sankhya C* 36:99–118.

Haziza, D., G. Chauvet, and J.-C. Deville. 2010. "Sampling and Estimation in the Presence of Cut-Off Sampling." *Australian & New Zealand Journal of Statistics* 52 (3): 303–319. https://doi.org/10.1111/j.1467-842X.2010.00584.x.

Hidiroglou, M. A., and P. Lavallée. 2009. "Sampling and Estimation in Business Surveys." In *Handbook of Statistics: Design, Methods and Applications,* edited by D. Pfeffermann and C. R. Rao, vol. 29A, 441–470. Elsevier. https://doi.org/10.1016/S0169-7161(08)00017-5.

Kim, J. K., and Y. Kwon. 2024. "Comments on "Exchangeability Assumption in Propensity-Score Based Adjustment Methods for Population Mean Estimation Using Non-Probability Samples"." *Survey Methodology* 50 (1): 57–63.

Kim, J. K., and K. Morikawa. 2023. "An Empirical Likelihood Approach to Reduce Selection Bias in Voluntary Samples." *Calcutta Statistical Association Bulletin* 75 (1): 8–27. https://doi.org/10.1177/00080683231186488.

Kim, J. K., and Z. Wang. 2019. "Sampling Techniques for Big Data Analysis." *International Statistical Review* 87 (1): 177–191. https://doi.org/10.1111/insr.12290.

Kim, J. K., and C. L. Yu. 2011. "A Semiparametric Estimation of Mean Functional With Nonignorable Missing Data." *Journal of the American Statistical Association* 106 (493): 157–165. JSTOR: 41415541.

Kim, J.-K., and S.-M. Tam. 2021. "Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference." *International Statistical Review* 89 (2): 382–401. https://doi.org/10.1111/insr.12434.

Köhler, M., A. Schindler, and S. Sperlich. 2014. "A Review and Comparison of Bandwidth Selection Methods for Kernel Regression." *International Statistical Review* 82 (2): 243–274. https://doi.org/10.1111/insr.12039.

Little, R. J. A., B. T. West, P. S. Boonstra, and J. Hu. 2020. "Measures of the Degree of Departure from Ignorable Sample Selection." *Journal of Survey Statistics and Methodology* 8 (5): 932–964. https://doi.org/10.1093/jssam/smz023.

Liu, A.-C., S. Scholtus, and T. De Waal. 2022. "Correcting Selection Bias in Big Data by Pseudo-Weighting." *Journal of Survey Statistics and Methodology,* 1–23. https://doi.org/10.1093/jssam/smac029.

Liu, Z., X. Chen, R. Li, and L. Hou. 2024. "Bayesian Inference for Nonprobability Samples with Nonignorable Missingness." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 17 (1): e11667. https://doi.org/10.1002/sam.11667.

Liu, Z., D. Wang, and Y. Pan. 2025. "Superpopulation Model Inference for Non-Probability Samples under Informative Sampling." *Communications in Statistics - Simulation and Computation* 54 (10): 4213–4234. https://doi.org/10.1080/03610918.2024.2374056.

Lohr, S., and J. N. K. Rao. 2006. "Estimation in Multiple-Frame Surveys." *Journal of the American Statistical Association* 101 (475): 1019–1030. https://doi.org/10.1198/016214506000000195.

Lohr, S. L. 2011. "Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames." *Survey Methodology* 37 (2): 197–213.

———. 2021. "Multiple-Frame Surveys for a Multiple-Data-Source World." *Survey Methodology* 47 (2): 229–263.

———. 2022. "Comments on "Statistical Inference with Non-Probability Survey Samples"." *Survey Methodology* 48 (2): 331–338.

Lohr, S. L., and T. E. Raghunathan. 2017. "Combining Survey Data with Other Data Sources." *Statistical Science* 32 (2): 293–312. https://doi.org/10.1214/16-STS584.

Marella, D. 2023. "Adjusting for Selection Bias in Nonprobability Samples by Empirical Likelihood Approach." *Journal of Official Statistics* 39 (2): 151–172. https://doi.org/10.2478/jos-2023-0008.

Mashreghi, Z., D. Haziza, and C. Léger. 2016. "A Survey of Bootstrap Methods in Finite Population Sampling." *Statistics Surveys* 10. https://doi.org/10.1214/16-SS113.

Matei, A., and Y. Tillé. 2005. "Evaluation of Variance Approximations and Estimators in Maximum Entropy Sampling with Unequal Probability and Fixed Sample Size." *Journal of Official Statistics* 21 (4): 543–570.

Medous, E., C. Goga, A. Ruiz-Gazen, J.-F. Beaumont, A. Dessertaine, and P. Puech. 2023. "QR Prediction for Statistical Data Integration." *Survey Methodology* 49 (2): 385–410.

Meng, X.-L. 2018. "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election." *The Annals of Applied Statistics* 12 (2): 685–726. https://doi.org/10.1214/18-AOAS1161SF.

———. 2022. "Comments on "Statistical Inference with Non-Probability Survey Samples" – Miniaturizing Data Defect Correlation: A Versatile Strategy for Handling Non-Probability Samples." *Survey Methodology* 48 (2): 339–360.

Morikawa, K., and J. K. Kim. 2021. "Semiparametric Optimal Estimation with Nonignorable Nonresponse Data." *The Annals of Statistics* 49 (5). https://doi.org/10.1214/21-AOS2070.

Morikawa, K., J. K. Kim, and Y. Kano. 2017. "Semiparametric Maximum Likelihood Estimation with Data Missing Not at Random." *Canadian Journal of Statistics* 45 (4): 393–409. https://doi.org/10.1002/cjs.11340.

Nadaraya, E. A. 1964. "On Estimating Regression." *Theory of Probability & Its Applications* 9 (1): 141–142. https://doi.org/10.1137/1109020.

Newey, W. K., and D. McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics,* 4:2111–2245. Elsevier.

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625. https://doi.org/10.2307/2342192. JSTOR: 2342192.

Rao, J. N. K. 2021. "On Making Valid Inferences by Integrating Data from Surveys and Other Sources." *Sankhya B* 83 (1): 242–272. https://doi.org/10.1007/s13571-020-00227-w.

Rao, J. N. K., and W. Fuller. 2017. "Sample Survey Theory and Methods: Past, Present, and Future Directions." *Survey Methodology* 43 (2): 145–160.

Riddles, M. K., J. K. Kim, and J. Im. 2016. "A Propensity-Score-Adjustment Method for Nonignorable Nonresponse." *Journal of Survey Statistics and Methodology* 4 (2): 215–245. https://doi.org/10.1093/jssam/smv047.

Righi, P., G. Bianchi, A. Nurra, and M. Rinaldi. 2019. "Integration of Survey Data and Big Data for Finite Population Inference in Official Statistics: Statistical Challenges and Practical Applications." *Statistica & Applicazioni* 17 (2): 135–158. https://doi.org/10.26350/999999_000025.

Rivers, D. 2007. "Sampling for Web Surveys." In *Joint Statistical Meetings 2007, Section on Survey Research Methods.* Joint Statistical Meetings 2007.

Rivière, P. 2002. "What Makes Business Statistics Special?" *International Statistical Review* 70 (1): 145–159. https://doi.org/10.1111/j.1751-5823.2002.tb00353.x.

Rubin, D. B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–592. https://doi.org/10.2307/2335739.

———. 1987. *Multiple Imputation for Nonresponse in Surveys.* 1st ed. Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/9780470316696.

Rueda, M. d. M., S. Pasadas-del-Amo, B. C. Rodríguez, L. Castro-Martín, and R. Ferri-García. 2023. "Enhancing Estimation Methods for Integrating Probability and Nonprobability Survey Samples with Machine-Learning Techniques. An Application to a Survey on the Impact of the COVID-19 Pandemic in Spain." *Biometrical Journal* 65 (2): 1–19. https://doi.org/10.1002/bimj.202200035.

Rueda-Sánchez, J. L., R. Ferri-García, M. D. M. Rueda, and B. Cobo. 2025. "Evaluation of Available Techniques and Their Combinations to Address Selection Bias in Nonprobability Surveys." *AStA Advances in Statistical Analysis,* https://doi.org/10.1007/s10182-025-00530-9.

Salvatore, C. 2023. "Inference with Non-Probability Samples and Survey Data Integration: A Science Mapping Study." *METRON,* https://doi.org/10.1007/s40300-023-00243-6.

Särndal, C.-E., and S. Lundström. 2006. *Estimation in Surveys with Nonresponse.* Wiley Series in Survey Methodology. Chichester: Wiley.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling.* New York, NY, US: Springer-Verlag Publishing. https://doi.org/10.1007/978-1-4612-4378-6.

Savitsky, T. D., M. R. Williams, J. Gershunskaya, and V. Beresovsky. 2023. "Methods for Combining Probability and Nonprobability Samples under Unknown Overlaps." *Statistics in Transition new series* 24 (5): 1–34. https://doi.org/10.59170/stattrans-2023-061.

Schenkel, J. F., and L.-C. Zhang. 2022. "Adjusting Misclassification Using a Second Classifier with an External Validation Sample." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 185 (4): 1882–1902. https://doi.org/10.1111/rssa.12845.

Scholtus, S., A.-C. Liu, and T. De Waal. 2025. "Notes on the Variance of a Pseudo-Weighted Estimator for Selection Bias Correction." *METRON* 83:31–42. https://doi.org/10.1007/s40300-024-00284-5.

Scott, D. W. 2014. *Multivariate Density Estimation: Theory, Practice, and Visualization.* 2nd Edition. Hoboken, New Jersey: Wiley.

Shao, J., and L. Wang. 2016. "Semiparametric Inverse Propensity Weighting for Nonignorable Missing Data." *Biometrika* 103 (1): 175–187. https://doi.org/10.1093/biomet/asv071.

Shlomo, N., and M. S. Kim. 2025. "Quality Indicators for Representation in Administrative Data." *Survey Methodology* 51 (2): 509–532.

Statistics Canada. 2018. "Modernizing the National Statistical System - Stakeholder Consultations," July 26, 2018. (Accessed: November 21, 2022). Available at: https://www150.statcan.gc.ca/n1/pub/89-20-0003/892000032019001-eng.htm.

Stefanski, L. A. 2000. "Measurement Error Models." *Journal of the American Statistical Association* 95 (452): 1353–1358. https://doi.org/10.2307/2669787.

Stefanski, L. A., and D. D. Boos. 2002. "The Calculus of M-Estimation." *The American Statistician* 56 (1): 29–38. https://doi.org/10.1198/000313002753631330.

Tam, S.-M., and F. Clarke. 2015. "Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics." *International Statistical Review* 83 (3): 436–448. https://doi.org/10.1111/insr.12105.

Tam, S.-M., and A. Holmberg. 2020. "New Data Sources for Official Statistics – A Game Changer for Survey Statisticians?" *The Survey Statistician* 81:21–35.

Tam, S.-M., and G. Van Halderen. 2020. "The Five V's, Seven Virtues and Ten Rules of Big Data Engagement for Official Statistics." *Statistical Journal of the IAOS* 36 (2): 423–433. https://doi.org/10.3233/SJI-190595.

Tillé, Y., M. Debusschere, H. Luomaranta, M. Axelson, E. Elvers, A. Holmberg, and R. Valliant. 2022. "Some Thoughts on Official Statistics and Its Future (with Discussion)." *Journal of Official Statistics* 38 (2): 557–598. https://doi.org/10.2478/jos-2022-0026.

Vale, C. D., and V. A. Maurelli. 1983. "Simulating Multivariate Nonnormal Distributions." *Psychometrika* 48 (3): 465–471. https://doi.org/10.1007/BF02293687.

Valliant, R., and J. A. Dever. 2011. "Estimating Propensity Adjustments for Volunteer Web Surveys." *Sociological Methods & Research* 40 (1): 105–137. https://doi.org/10.1177/0049124110392533.

Wand, M. P., and M. C. Jones. 1995. *Kernel Smoothing.* 1st Edition. Monographs on Statistics and Applied Probability 60. London ; New York: Chapman & Hall.

Wang, L., R. Valliant, and Y. Li. 2021a. "Adjusted Logistic Propensity Weighting Methods for Population Inference Using Nonprobability Volunteer-based Epidemiologic Cohorts." *Statistics in Medicine* 40 (24): 5237–5250. https://doi.org/10.1002/sim.9122.

———. 2021b. "Supporting Document for Adjusted Logistic Propensity Weighting Methods for Population Inference Using Nonprobability Volunteer-Based Epidemiologic Cohorts." *Statistics in Medicine* 40:1–9.

Wang, S., J. Shao, and J. K. Kim. 2014. "An Instrumental Variable Approach for Identification and Estimation with Nonignorable Nonresponse." *Statistica Sinica* 24 (3). https://doi.org/10.5705/ss.2012.074.

Watson, G. S. 1964. "Smooth Regression Analysis." *Sankhya A* 26 (4): 359–372. JSTOR: 25049340.

West, B. T., and R. R. Andridge. 2023. "Evaluating Pre-Election Polling Estimates Using a New Measure of Non-Ignorable Selection Bias." *Public Opinion Quarterly* 87 (S1): 575–601. https://doi.org/10.1093/poq/nfad018.

Wright, R. L. 1983. "Finite Population Sampling With Multivariate Auxiliary Information." *Journal of the American Statistical Association* 78 (384): 879–884.

Wu, C. 2022. "Statistical Inference with Non-Probability Survey Samples." *Survey Methodology* 48 (2): 283–311.

———. 2023. "Challenges and Strategies in Dealing with Non-Probability Survey Samples." In *Proceedings of the 2nd Workshop on Methodologies for Official Statistics,* 13–23. 2nd Workshop on Methodologies for Official Statistics. Istituto Nazionale Di Statistica.

Wu, C., and M. E. Thompson. 2020. *Sampling Theory and Practice.* ICSA Book Series in Statistics. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-44246-0.

Xu, Z., and B. Nandram. 2024. "Supplementing a Non-Probability Sample With a Probability Sample to Predict the Finite Population Mean." *International Journal of Statistics and Probability* 13 (2): 16. https://doi.org/10.5539/ijsp.v13n2p16.

Yang, S., and J.-K. Kim. 2020. "Statistical Data Integration in Survey Sampling: A Review." *Japanese Journal of Statistics and Data Science* 3, no. 2 (December): 625–650. https://doi.org/10.1007/s42081-020-00093-w.

Yang, S., J.-K. Kim, and Y. Hwang. 2021. "Integration of Data from Probability Surveys and Big Found Data for Finite Population Inference Using Mass Imputation." *Survey Methodology* 47 (1): 29–58.

Yang, Z., and N. Tang. 2020. "Empirical Likelihood for Nonlinear Regression Models with Nonignorable Missing Responses." *Canadian Journal of Statistics* 48 (3): 386–416. https://doi.org/10.1002/cjs.11540.

Yi, G. Y. 2017. *Statistical Analysis with Measurement Error or Misclassification.* Springer Series in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-6640-0.

Yorgason, D., B. Bridgman, Y. Cheng, A. H. Dorfman, Y. K. Liu, J. Miranda, and S. Rumburg. 2011. "Cutoff Sampling in Federal Surveys: An Inter-Agency Review." Available at: https://www.bls.gov/osmr/research-papers/2011/pdf/st110050.pdf.

Zhang, L.-C. 2019. "On Valid Descriptive Inference from Non-Probability Sample." *Statistical Theory and Related Fields* 3 (2): 103–113. https://doi.org/10.1080/24754269.2019.1666241.

Zhang, L.-C. 2021. "Proxy Expenditure Weights for Consumer Price Index: Audit Sampling Inference for Big-Data Statistics." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184 (2): 571–588. https://doi.org/10.1111/rssa.12632.

———. 2023. "Audit Sampling as a Quality Standard for Multisource Official Statistics." *Spanish Journal of Statistics* 5 (1): 67–83. https://doi.org/10.37830/SJS.2023.1.05.

Zhu, T., L. J. Gamble, M. Klapman, L. Xue, and V. M. Lesser. 2023. "Using Auxiliary Information in Probability Survey Data to Improve Pseudo-Weighting in Nonprobability Samples: A Copula Model Approach." *Journal of Survey Statistics and Methodology,* 1–27. https://doi.org/10.1093/jssam/smad032.