The London School of Economics and Political Science

A Structured Approach to Web Panel Surveys

The Use of a Sequential Framework for non-Random Survey Sampling Inference

Yehuda Dayan

A thesis submitted to the Department of Statistics of the London School of Economics for the degree of Doctor of Philosophy, London, April 2014

To my father who showed me the direction
To my mother who kept me on the path

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 51,924 words.

Abstract

Web access panels are self selected panels constructed with the aim of drawing inference for general populations, including large segments of the population who rarely or never access the Internet. A common approach for modeling survey data collected over access panels is combing it with data collected by a randomly selected reference survey sample from the target population of Interest. The act of joining the panel is then treated as a random process where each member of the population has a positive probability of participating in the survey. The combined reference and panel survey sample can then be used for different estimation approaches which model either the selection process or the measurement of interest, or some case the two together. Most practitioners and academics who have considered this combined sample approach, model the selection process by a single phase process from the target population directly to the observed sample set.

In the following work, I assume selection into the panel is a sequential rather than a single phase process and offer several estimators that are underlined by appropriate sequential models. After a careful investigation of a variety of single phase methods applied in practice, I demonstrate the benefits a sequential framework has to the panel problem. One notable strength of this approach is that by assuming a sequential framework the modeler can include important variables associated with Internet and Web usage. Under a single phase model inclusion of such information would invalidate basic assumptions such as independence between selection and model covariates.

In this work I also suggest a carefully structured panel estimation strategy, combining a sample selection design with chosen estimator. Under the sequential framework I demonstrate the potential of combining a within-panel random sampling procedure, that is balanced on a sequence of target statistics, with estimators that are modeled over both the selection process and the variable of interest. I show that this strategy has several robustness properties over and beyond currently applied estimators. I conclude by describing an estimation algorithm which applies this estimation strategy to the combined panel and reference survey sample case.

Acknowledgements

I would like to thank my supervisor, Professor Jouni Kuha, for his consistent support from the early stages of this research to the last days of writing this document. His sharp advice and confidence in my ability to complete this journey has been invaluable. I also would like to thank wholeheartedly the LSE Department of Statistics who overlooked my part-time student status and have supported me with real generosity.

This research could not have happened without Karsten Shaw, Dr. David Herron and Dr. Patten Smith of Ipsos-MORI who provided financial and professional backing, as well as a significant level of flexibility at work to allow me to pursue this project.

Above all, I want to thank my parents, Yossi, Ayelet and Meyrav for their unconditional love and support.

And finally, to Julia, Ester and Nessy II. For sticking through, and making this worth it.

Contents

Chapter 1

Web Access Panels

1.1 Introduction and Opening Words

The expansion of Online research and Online survey research is a major trend in the past fifteen years. From a very small base Inside research (2009) put the US total Online research in 2009 at about \$2 billion, while the UK market research society estimates that 26% of all research in the UK in 2010 was done Online. The Lion's share of this research is done over the platform of Web panels. This is largely a replacement technology with 85% of the research conducted over these panels have been migrated from traditional research modes such as face to face or telephone. The main engine for this growth has been market research studies such as brand evaluation, product concept trails, advertising testing, customer satisfaction surveys. Political polling, always having an outsized influence[1] on the research market and the public perception of research, has also shifted increasingly to Online platforms and has had a major part in a wider public and commercial acceptance of estimates based on Web panel data.

Panels can be recruited by means of traditional probability designs (such as the Knowledge Networks panel), however, the vast majority of panels are assembled through a non-probability and loosely controlled process. As means of recruitment panels employ a large range of procedures to solicit population members to join the panel. These offers normally suggest the prospect of receiving monetary remuneration for their survey response, but also stress social aspects such as the participation in public debate or the possibility of influencing the development of new products or improving relationships with companies. After joining, the panel collects profiling information such as social and demographic indicators as well as some attitudinal and behaviour information. The panel company then communicates with their members by email or when panelists actively log on to the panel website.

[1] With only 2% of total research spend, inside research (2009)

The non controlled and non-probability method of panel recruitment are in stark contrast to the conventional framework of survey sampling theory underlying much of the survey based research for over 50 years (Baker et al., 2010), and so it is not surprising the resistance Web panel faces from much of the key stake holders of the industry including the academic, governmental as well as the commercial communities. But given the deep societal changes of the last decade reflected in dramatic new methods of communication and social interactions, changes in consumer habits, as well as commercial pressures the march of surveys moving Online is inevitable.

However, this relatively fast shift in the adaptation of Web panels in the industry has exceeded the pace of methodological advancement. And while still fiercely debated in academia, Web panels are now an integral part of survey based research. Thus, it is necessary to acknowledge the importance of Web surveys, instead of neglecting their potentials by regarding them as a cheap and dirty method. It becomes now the methodologists responsibility to devise ways to improve and devise frameworks that can underline a statistical approach to Web panel based survey inference.

Luckily, there have been a number of substantial attempts by social scientists in the design aspect of Web surveys particularly in questionnaire design and usability issues. However, findings in these studies do not cover the full picture of Web survey methodology, as they are limited to improving the quality of data collected from persons who already participate in the surveys. At the same time, academic researchers have in recent years started investigating and publishing work on the topic and along with valuable work done in the commercial sector there exists a good initial body of empirical and theoretical work to build on.

Still, a great deal more is left to develop in the field and my hope is that this work may contribute to a more structured and detailed framework for the statistical inference of general population parameters based on Web panel survey sampling data.

1.2 Online Panels

Traditionally, the principal assignment for any survey based research is to establish a sample frame of the target population of interest. When the sampling mode is Online, the researcher encounters the limitation that not all members of the target population can access the mode on which the survey collection platform is built on - the Internet[2]. When the target population of interest is certain classes of the population such as the student population or a commercial organization, this issue is of minor concern given their high Internet penetration levels. However for general (e.g. national or international) populations a non negligible and distinct segment of the population is not Online. Despite the digital revolution of the past decade and the seemingly omnipresence of the Internet in our daily lives in 2012 still only 76% of the EU population had used the internet at least once, while only 67.5% of the EU population are regular users - that is use the Internet at least one time a week (European Commission, 2013). In the US the Pew research center estimates that only 72% of Internet users go Online at least once a week (Rainie, 2010). Furthermore, the use of Internet is strongly associated with education level, income, profession, health disability and age: for example in the UK, while almost all people aged 16-44 have used the Internet, only 3 out of 10 people aged 75+ have (ONS 2014). In the US, demographic groups such as Hispanics and blacks are underrepresented in the Online population (Baker et al., 2010).

It is true that effectively all Internet users have an email account, but still there is no complete list of these addresses. Furthermore, while several Internet users may share a single email address, more and more commonly users acquire multiple email accounts from different providers which are used for different activities. Many of these addresses are forgotten, left unused without being disabled. In general , the non standardised format of emails blocks the possibility of establishing an Online equivalent of the telephone sampling random digit dialing (RDD) methodology. It is worth noting that even if a useful list of email account could be compiled, legal and commercial body regulations would block the use of such a list for privacy and consumer rights reasons (Baker et al., 2010).

The Online panel platform is a widely used solution to this lack of sampling frame or sampling methodology for general population research. The ISO 26362: Access Panels in Market, Opinion, and Social Research offers the following definition of Online panels: “A sample database of potential respondents who declare that they will cooperate with future [online] data collection if selected” (International Organization for Standardization 2009). This is a broad definition which can include (as noted in section 1.1) panel databases collected by an Offline random probability design as well as one that is collected through a non probability selection procedure - namely through a wide placement of Online ads and offers to join the panel and partake in future surveys.

A further distinction can be made by the population targeted for the panel based research. *General Population panels* are collected to correspond to the general population Online, or when census balanced, to represent the general popula-

[2] Or more specifically in this work, the World Wide Web

tion. *Consumer Panels* are similar but may be limited to certain ages (25-55) and demographics more relevant to market research exercises such as a study of the consumer packaged goods market. *Sub-panels* aim to represent certain sub-populations for which unambiguous features or attributes are available linking them to groups of panelist for specific areas of studies. One example is a Handicap or disability panel, where members of the general population panel can be identified as having a disability and can be gauged on unique health issues, social views, economic status and even shopping patterns. This is a cost effective way of researching such low incidence population segments. Similarly *Specialty Panels* and *Business to Business (B2B)* panels which aim to survey specific members of the population such as business executives legal professionals, musicians, hunters and other members of the panel representing small, low incidence, groups of the population. *Proprietary Panels* are panels usually built and maintained by a research company but exclusively used by a certain company for their own research needs. Banks, Content providers or large retailers with large customer databases are companies which benefit from owning a panel representing their client population.

As noted my research focuses on the problem of inference of general population statistics based on data collected from panels assembled Online through a non probability process, however, the ideas are immediately transferable to any of the above panel types.

1.3 Data-Collection Methods in the Face of Social, Economic and Technological changes

The field of survey methodology and specifically its data collection and data measurement tools evolve dynamically along with the cultural and technological changes. The survey methodology field advances by expanding the variety of measurements of day to day activities and the views of the society it studies. Over the previous century, among the evolutions most notable in the field was the introduction of telephone interview (Groves and Kahn, 1979; Dillman, 1998; and Dillman, 2002). It is noteworthy then that when the idea of conducting surveys over telephone was first introduced, researchers were skeptical and not fully convinced of the effectiveness of the method. One possible explanation was the famous failed Literary Digest poll which was based as well on a list of telephone owners. It is also true that the prevailing belief at the time was that surveys should involve face-to-face interactions.

However, this changed and since the Health Survey Methods Conference in 1972, where telephone interviewing first received attention as a serious data collection mode (Dillman, 1998), there was (and still is) a great effort to build and improve telephone survey methodology (e.g., Groves and Kahn, 1979). Meanwhile, an innovative concept of balancing survey costs and errors to the maximum degree has influenced researchers to design surveys within some fixed amount of budget - The total survey error paradigm (e.g., Groves, 1989). A well-defined probability sampling procedure by random digit dialing was developed for telephone surveys (e.g., Mitofsky, 1970; Waksberg, 1978; Lepkowski, 1988; Casady and Lepkowski, 1993). Inevitably, practical considerations and societal changes boosted the legitimacy of telephone interviews. For example, increased telephone usage and a lowered household contactability for face-to-face interviews due to an increase in female workforce and a decrease in household size have made surveys by telephone more feasible and cost-effective. Now, telephone surveys are a standard data collection method in developed countries.

It is clear now that the digital revolution and the internet have created another significant societal change and with it another leap in the sources of measurement available. The internet is profoundly changing the way we communicate with one another and so, the survey research field is experiencing another challenging transition into implementing Internet based surveys.

As noted already Web surveys[3] have been both hyped for their capabilities and criticized for their limitations. However, to put this in historical perspective, it is instructive to return to what was written about telephone and mail surveys when they were still regarded as unproven survey methodologies. In 1978, Don Dillman, a noted authority on surveying, said the following about mail and telephone survey questionnaires: 'Neither mail nor telephone has been considered anything more than a poor substitute...' for the highly regarded face-to-face interview. At the time this view was probably justified, because the two methods had many deficiencies and problems. Surveys by mail typically elicited extremely low response rates, even with short questionnaires. Further, it was not possible to reach many people by mail questionnaires, and among those to whom questionnaires could be delivered, the best educated were far more likely to respond. Even completed questionnaires left much to be desired in terms on non answered questions. It is not surprising, then, that users of the mail questionnaire treated response rates well below 50 percent as acceptable and explained away problems of data quality with disclaimers such as, this is the best we can expect from a mail questionnaire (Dillman, 1978, pp. 12).

Not unlike the situation with mail surveys in the 1970s, many questions and concerns exist about how to best conduct Web surveys and whether they are, in fact, scientifically valid. However, reflecting on Dillman quotation and making the necessary substitutions between web, mail and Web, the statement will ac-

[3] In Web surveys the respondent visits the survey Web site by either clicking a hyperlink in an e-mail or in another Web site, or by typing the Web address directly into the address box in the browser window.

curately reflect much of the criticism directed at Internet-based surveys today. Therefore, it is inevitable to consider Internet surveys as an alternative or rather a new survey tool alongside traditional mail and phone survey methods.

More positively, Web-surveys do have advantages over more-traditional methods in certain applications, and the use of this medium will continue to expand. Compared to traditional survey platforms proponents argue (Kellner, 2008; Rivers, 2010) that: (1) they are less time consuming; (2) they are just as good or better than more-traditional surveys; (3) they are much cheaper to conduct; and (4) they are easier to execute. However, these assumptions may or may not be true depending on the individual circumstances of the survey. Therefore, researchers need to understand the current limitations of Web surveys.

According to Dillman (2002) our survey methods are more a dependent variable of society rather than an independent variable constructed by the survey community in isolation. In fact the ideal survey methodology is likely to reflect the society and its culture. As Taylor and Terhanian (2003) argue, just as telephone surveys began to be adopted extensively a few decades ago mirroring the societal and technological trends, the survey methodology field is currently witnessing a widespread growth in the use of Web surveys. These changes are simply manifestation of societal trends.

1.4 The Web access Panel- Mechanics and a Total Error Perspective

Web access panels have their roots in the prior post-mail panels started by US market research companies, including Synovate, NOP and others (Baker et al., 2010). These mail panels were assembled by a non random processes, similar to the Web panels, but through non Internet sources. These panels were marketed by commercial research companies based on their superior speed, cost and ability to capture low incidence population - again similar to the strengths of today's Web panels.

Each Web panel is built and maintained in a very idiosyncratic process. This is a reflection of the commercial nature of the panels, the relative novelty of the idea as well as a lack of an agreed theoretical framework to model the process and the analysis of the survey data recorded. In the following I touch on the main elements considered by a panel management team from an operational and methodological perspective.

Population Coverage

Our interest is in estimating certain statistics of a target population. In survey sampling theory a target population is a finite population of units, that is a population which at least in theory can be counted over a certain period of time and are thus observable. A classic example is a survey targeting the US population of adult household members in a certain month. Usually a target population is defined already while considering a sampling frame from which to collect population members. That is a researcher may want to survey the entire adult population, however technically and financially limiting the aim to the household member population is a more feasible proposition. Any sampling frame (e.g. the Postcode Address File in Great Britain) will have limitations that may cause errors in the eventual estimation stage. In traditional sampling frames these include undercoverage, such as non registered households but also overcoverage- caused by multiple mapping and duplications which usually are a function of several household members clustered under the same address. To deal with these issues the survey community established recognized and tested procedures to identify and minimize such problems

In the Web access panel case the idea of a population sampling frame is substituted by the panel itself and evaluation of frame errors are replaced by a focus of collecting a large and diverse enough set of the target population members, allowing the researcher to sample a representative sample. The problem is that the term diverse cannot be defined properly in practice. For different research topics, a different diversity is required - for some instances social demographic diversity is enough , in other areas it is a range of political views while in another consumer habits. Randomization theory cuts through this problem by assuring that, over repeated applications, for large enough sample the population and sample distributions are similar on all observable and non observable variables. When randomization is not possible, we are left to empirically compare to available information on the target population from sources such as national statistics or reputable survey based statistics, but even then the researcher can test balance only on available data and can only hope that this indicates balance on non observable distributions. Finally, it is worth noting that the problem of overcoverage by multiple mapping also effects the Web panel as the panel frame, the list of Internet emails, is likely to have many cases where several population members are represented by a single address. Another problem, is the phenomena of single population members participating in several panels simultaneously. For example Walker et al. (2009), investigating 17 different Web panels, found a duplication rate, a population unit which is a member of multiple panels, of between 16-40%.

Recruitment, Registration and Profiling of Panelists

Non response or non cooperation is an error affecting any survey mode as it severs the link between the planned selection design from the actual sampling distribution. From the nature of Web access panels' assembly procedure, which relies on a multitude of contact points with the target population (many of them loosely controlled by the panel management team), the magnitude and characteristics of such non cooperation with the collection process is significant.

The recruitment to the panel is mainly done Online, but can have an Offline element as well. Panel companies attempt to advertise and solicit volunteering to as many members of the population as possible. The extent of the recruitment drive is a function of the cost constraints and the desired social and demographic profile of the panelists. Motivation to join and participate in surveys can be driven by aspects such as monetary incentive - either fixed or contingent, the desire of the member for self-expression of certain views and influence public opinion, the opportunity to view survey results and as a means of entertainment. It may be somewhat surprising that when studying panelists' motivation Poynter and Comley (2003) found a fairly even mix of the above list with monetary reward (42%) the highest stated, followed by curiosity (40%), entertainment (20%) and a desire for self expression of views (28%).

As noted, panel development is done in several fronts. Probably the most common way of recruitment is through the purchasing of email lists from specialised vendors. These lists are compiled through co-registration agreements between the email list vendor and web site owners, who actively maintain email lists of visitors who sign up voluntarily and agree to be contacted by third party 'partners' (Baker et al., 2010). This is the approach taken in the US originally by Harris Interactive in the late 90's and YouGov in the UK. Another common recruitment platform is Web banners or display ads. A panel owner has the option to contact websites directly, through affiliate networks to which those websites belong or through a media agency (Michael 2010). Yet anther recruitment source is through search engines such as Google, Yahoo or Bing where the panel company or search engine consultants attempt to manipulate the rank of their panel website on the search engine list when relevant search terms are typed in. The same search engines are popular web sites for placement of display ads, where again, the panel companies compete to have their web site link displayed by search terms deemed relevant to population members who are likely to self select into the access panel. As noted above. additional recruitment initiatives can be done Offline, these include offering survey respondents in face to face, telephone or post mail surveys to join the Online access panel. Another method is referral, snow balling or viral recruitment where panelists are encouraged to suggest family, friends or other members of their social circle to join the panel.

Given the uncontrolled nature of the process, it is almost impossible to estimate or even define the response rate of the overall selection process. In one such attempt Alvarez et. al, (2003) studied the case of Web banner recruitment and found that the recruitment rate per Web banner impression was 0.002% while the recruitment rate per actual Web banner ad click-through was just above 6%. Another study by the same authors found that a vendor email list compiled through a co registration agreement yielded a 32% recruitment rate. This is one indication of the reason for the prevalence of vendor list based recruitment.

Baker et al. (2010) summarize several empirical evidence of the recruitment or self selection error for US and EU panels. The evaluations are conclusive in that population member who join Web access panels differ significantly from the general Online population. For example a study of a Dutch panel (Vonk et al. 2006) found it underrepresented ethnic minorities and immigrants while over represented frequent Internet users, voters for certain political parties and religious groups. Similar evidence has been documented by US studies (Krosnick and Chang, 2004; Chang and Krosnick, 2009; Dever et al., 2008; Couper, 2000, and others).

The majority of large commercial panels follow a *double opt-in* procedure of joining the panel, which means that after the collection of a population member's (email) address, the panel sends him or her an email offering to join the panel by following an attached link leading to the panel web site. On arrival most panel companies screen potential panelists through a recruitment questionnaire, recording a long list of profiling data which can range from basic social demographic characteristics to attitudinal, behaviour and psychogrpahic[4] information which in turn will be used to identify the overall profile of the panel as well as sampling covariates for specific survey studies. The profiling phase is used as well to maintain the integrity of the panel by means of different validation procedures. The aim of these procedures is to prevent fraudsters from joining the panel and can include comparing profiling data to third party databases, checks on the stated home address against postal records and ISP address as well as data integrity evaluation (to test if the profession, age and other pieces of information given are 'reasonable') and digital fingerprinting to check for multiple registration (duplication tests) to the panel.

Incentive and Panel Maintenance

As is evident from the discussion above, acquiring a panelist is an expensive undertaking and so an important part of the management's team resources is to decrease the level of panel attrition. This is a natural process which occurs in any panel, with the largest group of panel drop-outs in fact are the group of new

[4] the study of personality, values, opinions, attitudes, interests, and lifestyles

members. Longer term, attrition can be influenced by the length of the surveys taken, the topic of studies when they do not match the respondents interests, or inversely by a panel not being engaged enough in the sense that the member is not sent survey requests over a long period of time.

The vast majority of panels offer panelists incentives. A common incentive is to use redeemable points that are collected for each survey completed. The amount of points gained can be a function of survey characteristics such as the length of the survey or the perceived interest (or how tedious, boring) members may find answering the questions. Another factor is the panelist characteristics, where low incidence panel members (such as top business executives, members with a specific disability) will likely be offered a higher number of points for completion of the survey. Points can be redeemed in numerous ways: gift certificates or other forms of vouchers, cash by check or bank transfer, purchase services or products through participating websites, multiple participation in sweepstakes.

An important guiding principle the panel management must always follow is that regardless of the incentive method, the level of remuneration should be high enough to incentive panelists to participate, but must not be large enough to encourage population members for which panel membership is a major source of income, that is professional panelists, to be encouraged to join.

Within Panel Survey Sampling

Given the unrepresentativeness of the panels, it is uncommon for a researchers to design a simple random sample from the panel. More likely, panel management will sample by non random purposive methods. In the panel management terminology 'sampling is the process of drawing a sample from the panel that fulfills certain criteria, often referred to as quotas' (Michael 2010). These criteria can be basic social demographic quotas taken from national statistic organizations, but depending on the survey topic can include constraints on the sample based on lifestyle and attitudes, brand and category usage, ownership health conditions. To estimate the quotas necessary, several panel companies run in parallel a smaller ad hoc random reference survey to estimate the population distribution of these specific variables. Such reference surveys, for cost reasons, can also be conducted on an ongoing basis and used to design several surveys simultaneously. This approach was pioneered by Harris Interactive. Clearly, the reference survey must match in wording the panel survey questions and should be as much as possible mode and time insensitive.

While purposive sampling, and especially quota sampling, is viewed dimly in the traditional survey sampling world[5] a strength of Online panels is their complete knowledge of the distribution of the panel members across the many profiling variables and the control on the distribution of the achieved sample. An interesting example of the use of this knowledge are political polls, where many panel companies at the time of a general (or in some cases local) elections, take a 'snap shot' of the electorate map by surveying all panel members over the days before and after the election. Companies use this profiling data to anchor the panel to the correct political landscape which they can then use on an ongoing basis for political, social and even economic research.

1.5 Structure of Thesis

In chapter 1 I have introduced the idea of Web panels as a substitute to the classical survey sampling population frame. I also have described there the main methodological challenges practitioners face when using this new survey platform. In chapter 2 I review common approaches to the problem, beginning with the important fixed population framework. In chapter 2 I also explain the statistical logic of common panel practices such as purposive and quota sampling. In chapter 3 I propose a sequential framework to the Web panel process and adapt the most common estimators reviewed in chapter 2 to the sequential case. I then discuss the strengths these estimators have compared to their single phase framework counterparts, and demonstrate these properties through several simulation studies. In chapter 4 I introduce the idea of using a random reference random survey as a surrogate to the unknown general population distribution and describe how the sequential estimators of chapter 3 can be computed over a combined sample set collected over both Web panel and reference surveys. In the final sections of chapter 4 I move away from post survey adjustment methods and propose a within-panel sampling design, which aims to balance the achieved sample to a set of population statistics. I show that combining this balanced sampling design with estimators built over a combination of selection and measurement of interest models achieve an additional level of robustness to misspecification.

[5] Although there is a significant divergence across the Atlantic where in Europe quota sampling is far more accepted than in the US.

Chapter 2

Modelling Panel Data - Introduction to the Problem and a Review of Relevant Approaches

2.1 Overview of Chapter

In chapter 1, I described the complex and to an extent intractable survey process of Web access panels. The objective of this expensive undertaking is to draw inference about the population from which the panelists originate. This fits the survey sampling problem of estimating population quantities from a subset of the population dataset, However, the non random and highly complex nature of the underlying selection and survey process, and the lack of (even a proxy) sampling frame stands out from the main stream of survey research datasets. It may be only a small exaggeration to state that since the inception of scientific survey sampling in the 1920s Web panel surveys are the first survey platform possessing such characteristics that have gained traction and have not been dismissed outright by the research community.

In this first chapter I review common approaches relevant to the question of inference based on Web panel samples. Broadly, these approaches can be categorized into (i) those which rely on a probability model of the measurement of interest, (ii) those which rely on modelling the selection process or (iii) a combination of the two. For brevity and when unambiguous I will refer to these three approaches as m, π and πm-estimation receptively. Within these categories, for a given estimand the specific estimation procedures will differ on the inferential framework, the estimation procedure and the specific estimator. Regardless of these specifics, the different approaches have several common weaknesses and strengths in the context of the Web panel survey sampling setting which I will try and highlight.

Even before, I can state that one common weakness of these approaches is that they usually reduce the complex survey selection process into a single model. More clearly, any inference over data collected from a sample of a population must address explicitly or implicitly the selection (or missingness) mechanism. It will be evident from the discussion that as a general rule, commonly used approaches model the panel selection mechanism as a *one phase process*, which from the short introduction it is clearly not to be the case. This reduction of what is a sequence of conditional processes into a one phase process, I will argue, weakens both the ability of the analyst to offer a satisfactory model for the selection mechanism and may increase the likelihood of model misspecification of the measurement of interest - because it reduces artificially the complexity of the selection process, and prevents the use of valuable information which could otherwise be used directly.

A modelling framework which takes into account the sequential nature of the selection process will be given in the following chapter 3, while here I consider approaches which could be used *if* selection could be treated as being generated by a single phase mechanism.

In the following I start by setting up the notation and formalizing the basic problem, then section 2.3 discusses the applicability of the fixed population approach of survey sampling inference to the Web panel problem. This is a natural starting point given the popularity of the framework in the survey community. Section 2.5 covers the 'translation' of observational study methods, which are highly relevant to the Web problem, into the context and terminology of survey methodology. After this, over sections 2.7-2.9 I review separately common π, m and πm-estimation procedures, their strengths and weaknesses - all within the context of the Web panel problem.

2.2 Notation, Terminology and Basic Set Up of the Problem

A finite population $s^0 = \{1, ..., k, ..., N\}$ consists of $N < \infty$ units where k represents the k^{th} unit of the population, a physically existing element on which we can make measurements or observations. For convenience, when describing basic results I assume that the size N, at least conceptually, is known.

Let y_k quantify without error the value of y, our variable of interest associated with the k^{th} unit of the population. When y_k is in bold font then $\mathbf{y}_k = (y_{1k}, ..., y_{Qk})'$, that is for each unit k sampled, a $Q-$vector of variables is recorded - usually Q is of high dimension. The unknown set of population values is denoted in general by the $N \times Q$ matrix $\mathbf{y}_{s^0} = [\mathbf{y}_1, ...\mathbf{y}_k, ..., \mathbf{y}_N]^{'}$ and in the case where only one variable is recorded by the $N-$vector $\mathbf{y}_{s^0} = (y_1, ..., y_k, ..., y_N)^{'}$. When unambiguous, I may drop the s^0 which here indicates the relevant set of population units. These study variables can be continuous, as in situations where y_{qk}, $q = 1, .., Q$ stands for the 'income' or 'hours listening to the radio' of unit k of the population. In many other cases, however, y_{qk} is categorical , for example a dichotomous variable such that $y_{qk} = 1$ if k has the attribute '*The Times* reader' and $y_{qk} = 0$ '*Not The Times* reader'. But for specific cases, we consider our objective is in estimating linear functions of $\mathbf{y}$ such as the population mean $\overline{\mathbf{y}}_{s^0} = N^{-1} \sum_{s^0} y_k$.

In the most general setting we assume the value y_k is a single draw from the distribution of the possible outcomes and is therefore itself stochastic. The unobserved measurements $\mathbf{y} = [\mathbf{y}_1, ..., \mathbf{y}_N]^{'}$, the fixed finite population, are assumed to be the realized outcome of random variables $\mathbf{Y} = [\mathbf{Y}_1, ..., \mathbf{Y}_N]^{'}$, the *Superpopulation,* with joint distribution about which certain features are assumed known. This idea of a Superpopulation from which the finite population is a sample was first proposed by Deming and Stephan (1941).

Formally, let $f(\mathbf{y})$ be the probability density function (for continuous random variables) or the probability function (for discrete random variables) of the Superpopulation distribution of the random variables $\mathbf{Y}$. Normally when the values $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_N]^{'}$ of $\mathbf{X}$ an $N \times P$ matrix with typical $\mathbf{x}_k$ a $P \times 1$ vector are (or assumed) known for all members of the population s^0 then I shall use the conditional density function $f(\mathbf{y}|\mathbf{x})$. In some cases I assume that the family of Superpopulation distributions is indexed by an unknown parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_T)$ of finite dimension T so that the density $f(\mathbf{y})$ is denoted $f(\mathbf{y};\boldsymbol{\theta})$ and the conditional density function is denoted $f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$.

The survey statistician observes only a subset of the population of interest. As noted above, selection is underlined by a sequence of panel processes, however, most published work on Online access panel data (for example see Isaksson et al., 2004; Lee, 2006; Rivers and Bailey, 2009 and Lee and Valliant, 2009) reduces this into a single selection model from finite population to the sample surveyed. This is the approach we take throughout this chapter, but will be relaxed in subsequent chapters.

Formally let $\mathbf{S} = (S_1, ..., S_N)^{'}$ be the $N \times 1$ vector of indicator variables where $S_k = 1$ if unit k belongs to the observed set and $S_k = 0$ otherwise, with distribution $Pr(\mathbf{S} = \mathbf{s}) = p(\mathbf{s})$ where $\mathbf{s} = (s_1, ..., s_k, ..., s_N)^{'}$. Define s to be the set of population units participating in the survey, that is $s = \{k : S_k = 1\}$, then for any selection mechanism with no duplication (such as discussed in this work), the event $S = s$ is equivalent to the event $\mathbf{S} = \mathbf{s}$ that is $Pr(S = s) = Pr(\mathbf{S} = \mathbf{s})$. When necessary I associate with the selection distribution an unknown parameter vector $\boldsymbol{\phi}$ so we write $p(\mathbf{s}; \boldsymbol{\phi})$ or $p(s; \boldsymbol{\phi})$ and when appropriate a conditional distribution $p(s|\mathbf{x}; \boldsymbol{\phi})$.

Given s, the population $\mathbf{y}$ is partitioned $\mathbf{y}_{s^0} = (\mathbf{y}_s, \mathbf{y}_{\overline{s}})$ where $\overline{s}$ is the compliment set of unselected population members. Throughout assume that the observed sample size $\sum_{k \in s^0} S_k = n_S$ is large enough so that estimation and inference is reasonable.

Throughout I assume exchangeability of $p(s)$ and $f(\mathbf{y})$ in the sense that all permutations, for example for $(S_{k1}, ..., S_{kN})$, have the same $N-$dimensional probability distribution. In other words, the labels k are uninformative in the distribution. See Cassel et al. (1977, p.72) for a more detailed definition.

2.3 The Coverage Constraint of the Fixed Population Framework

As a survey sampling platform, an investigation of the Web access panel inference problem should start with the fixed population approach which is practised by the overwhelming majority of survey statisticians. In the following I discuss briefly the inferential framework and evaluate the difficulty of implementing relevant estimation strategies in the panel context. I assume throughout this section a basic familiarity of the reader with the fixed population framework of survey sampling inference. Standard references for this idea is given by Cochran (1977), Cassel et al. (1992), Särndal et al. (1992) and Kish (1995).

In the fixed population approach all values of variables are considered fixed - but unknown - at the outset. Variables $\mathbf{y}$ and $\mathbf{x}$ are not random, and are variables only in the sense of taking possibly different values (Särndal et al., 1992, section 9.2). Fundamentally, randomness is due only to the subset selection process. The problem is that whatever modelling approach one takes within this framework, when tackling the panel question the researcher is constraint by the fact a subset of the population cannot be selected into the sampling set, resulting in a biased estimator.

Classically, it is understood that under a certain probability model $p(\cdot)$ we can state before selection the probability of observing units $s \subseteq s^0$, the sample. More formally, for any $s \in \mathcal{S}$ where $\mathcal{S}$ is the set of all possible subsets of s^0, selection satisfies $p(s) \geq 0 \ \forall s \in \mathcal{S}$ and that $\sum_{s \in \mathcal{S}} p(s) = 1$. An important element in fixed population statistical inference is the calculation of the inclusion probability of a random unit into the set s. For a given $p(\cdot)$,

$$\pi_k \ = \ Pr(k \in S) \quad \text{that is } Pr(S_k = 1) = \sum_{S \ni k} p(S)$$

where the notation $\sum_{S \ni k}$ is understood as the sum over all samples s which include unit k of the population s^0.

In fixed population inference, a selection model (or sampling design) is said to be *measurable* if $\pi_k > 0$ and $\pi_{kl} = Pr(k \& l \in S) > 0$ for all $k, l \in s^0$. The positive probability of selection for each unit is a necessary and sufficient condition for the existence of unbiased estimators, while the positive joint probability for each pair of population units allows calculation of valid variance estimators, and using observed survey data, confidence intervals that achieve the theoretical coverage rate (see Lemma 3.4 in Cassel et al., 1977). Estimators are evaluated by properties such as expectation, variance, mean square error (MSE) which are all defined with reference to the specific selection model $p(\cdot)$.

Traditionally, the estimator accompanying the fixed population approach is the HT-estimator (Horvitz and Thompson, 1952)

$$\hat{\bar{\mathbf{y}}}_{s^0} = N^{-1} \sum_{k \in s^0} S_k \frac{y_k}{\pi_k} = N^{-1} \sum_{k \in S} \frac{y_k}{\pi_k}. \tag{2.1}$$

A popular alternative replaces N by $\hat{N} = \sum_s \pi_k^{-1}$.

Classical theory assumes that π_k are known and specified by the researcher. In Web panels, however, respondents self select into the panel. There is a rich academic and commercial literature dealing with self selection and unit non response in the survey sampling process. A popular approach is to describe the self selection as a separate conditional phase to the sampling design. That is, we assume first that a sample s^1 is selected by known design $p_1(\cdot)$, and subset s^2 selects to respond to the survey questions with unknown response process $p_2(\cdot)$ with (conditional) unit selection probability $\pi_{2k} = \sum_{k \ni s^2} p_2(s^2|s^1)$ where $s^2 \supset s^1$.

After invoking two selection models estimator (2.1) is impractical[1], and an appropriate estimator in this case is

$$\hat{\bar{\mathbf{y}}}_{s^0} = N^{-1} \sum_{k \in S^2} y_k (\pi_{1k}\pi_{2k})^{-1}. \tag{2.2}$$

[1] Even if both phases mechanism are known as π_k is not easily defined- we need to know $p_1(s^1)$ for all s^1 which in normal setting is available, but also for each s^1 we must know π_{2k} which can't be calculated without observing s^1.

The two estimators (2.1) and (2.1) are mostly of not the same form as

$$\begin{aligned}\pi_k &= Pr(k \in s^1)Pr(k \in s^2|k \in s^1) \\ &= \pi_{1k}\,(\pi_k/\pi_{1k}) \neq \pi_{1k}\pi_{2k}\end{aligned}$$

where the difference lies in conditioning the probability in selecting k to s^2 on the observed set s^1 rather than on the event that k is in the realized set s^1. See Särndal et al. (1992, section 9.2) for further discussion.

A popular approach of estimating the self selection probabilities π_2 is assuming a response homogeneity groups (RHG) which states that $s^1 = \cup_{h=1}^{H_{s^1}} s_h^1$ so that $\pi_{2k} = \pi_{2h}$ for any $k \in s_h^1$; $h = 1, ..., H_{s^1}$. Let $n_1 = \cup_{h=1}^{H_{s^1}} n_{1h}$ and $n_2 = \cup_{h=1}^{H_{s^2}} n_{2h}$ denote the sizes of s^1 and s^2 respectively then under RHG $\pi_{2h} = \frac{n_{2h}}{n_{1h}}$ so

$$\hat{\bar{\mathbf{y}}}_{s^0} = N^{-1}\sum_h \sum_{k\in s^2\cup s_h^1} (\frac{n_{2h}}{n_{1h}})^{-1}\pi_{1k}^{-1} y_k$$

which can be shown to be unbiased by referring to the law of total expectations, $E(\hat{\bar{\mathbf{y}}}_{s^0}) = EE(\hat{\bar{\mathbf{y}}}_{s^0}|s^1, \mathbf{n}_1)$ where $\mathbf{n}_1 = (n_{11,...}, n_{1h}, ..., n_{1H_{s^1}})$.

However, in the Web panel case the process inverts. Now s^1 represents the self selected set of Web panelists with $p_1(s^1)$ being unknown while s^2 is the survey sample of panelists with conditional probability $p_2(\cdot)$ known by design. The RHG model here is defined over s^0, that is $s^0 = \cup_{h=1}^{H} s_h^0$ define homogeneous selection strata such that $\pi_{1k} = \frac{n_{1h}}{N_h}$ for all $h = 1, ..., H$ where sizes $\mathbf{N} = (N_1, ..., N_H)$ are fixed as they are population characteristics. Assuming $\mathbf{N}$ is available from external sources and for any design p_2 estimator (2.2) is now

$$\begin{aligned}\hat{\bar{\mathbf{y}}}_{s^0} &= N^{-1}\sum_h \frac{N_h}{n_{1h}}\hat{n}_{1h}\frac{\sum_{k\in s^2\cup s_h^1} y_k\pi_{2k}^{-1}}{\sum_{k\in s^2\cup s_h^1}\pi_{2k}^{-1}} \\ &= \sum W_h'\hat{\bar{\mathbf{y}}}_{s_h^1\pi_2}\end{aligned}$$

where $W_h' = \frac{N_h}{n_{1h}}\hat{n}_{1h}/N$ and $\hat{n}_{1h} = \sum_{k\in s^2\cup s_h^1}\pi_{2k}^{-1}$ estimates the size of s^1. An alternative is to replace N with $\hat{N} = \sum_h \frac{N_h}{n_{1h}}\hat{n}_{1h}$.

However, strata now are over s^0 and define panel volunteering response groups using fixed population characteristics. It is unavoidable that most homogeneous (and likely largest such) group identifies non (at least regular) Internet users. Define s_1^0 to be this strata of non connected population units of size N_1. With $\pi_{1k} = 0$ by definition for s_1^0 let $h = 2, ..., H$ and let $B(\cdot)$ denote the expected bias of an estimator then

$$B(\hat{\bar{\mathbf{y}}}_{s^0}) = \frac{N - N_1}{N}(\bar{\mathbf{y}}_{s_1^0} - \bar{\mathbf{y}}_{\bar{s}_1^0})$$

the *coverage bias,* where $\overline{\mathbf{y}}_{\bar{s}_1^0}$ represents the average over population excluding members of s_1^0. Thus it is not surprising that most survey statisticians tackling the panel question limit their inference to the 'Internet connected' population (e.g. Bethlehem, 2008, 2010, Lee, 2009 and others). Counter intuitively, the only possibility now for an unbiased estimate will be if the RHG model *does not hold* as then the coverage error may cancel out with the selection model misspecification error.

2.4 The Use of Model Assisted and Calibration Estimators in Web Panels

The HT estimator and its variants such as the RHG model are only the most basic estimator in the fixed population framework, while Calibration (Deville and Särndal, 1992) and model assisted methods such as the generalized regression (GREG) estimator (Särndal, 1982; Robinson and Särndal, 1983) build on the approach by including population level auxiliary data. However, as I shall show in this short section, these methods share the same constraint due to the fixed population assumption.

Calibration is a popular approach many practitioners take which seems to avoid the coverage bias of the fixed population perspective (Lee and Valliant 2010, Chang and Kott 2005). The approach in essence is a systematic way of introducing auxiliary information into HT type estimation, with the aim to increase efficiency or increasingly to remove self selection bias. The calibration estimator is

$$\hat{\overline{\mathbf{y}}}_{s^0} = N^{-1} \sum (g_k/\pi_k)\mathbf{y}_k \tag{2.3}$$

such that π_k is the selection probability into the survey set, and g_k are calibration weights which minimize the distance

$$E_{p(s)}\left(\sum_s G_k(g_k, \pi_k^{-1})\right) \quad \text{s.t.} \sum_s (g_k/\pi_k)\mathbf{x}_k = \sum_{s^0} \mathbf{x}_k$$

for well behaved distance metric $G_k(\cdot)$ As Särndal and Lundstrom (2008) note calibration estimation development is intimately link to practice and the 'fixation' of national statistical agencies in estimators which are representable as the sum of weighted survey values.

To understand intuitively the problem of using calibration in the panel context, consider the specific case where $G_k(\cdot) = (g_k - \pi_k^{-1})^2/\pi_k^{-1}q_k$ for a known positive weight q_k. Then calibration simply provides an alternative derivation of the generalized regression (GREG) estimator

$$\hat{\bar{\mathbf{y}}}_{s^0} \quad = \hat{\bar{\mathbf{y}}}_{\pi s^0} + (\overline{\mathbf{x}}_{s^0} - \hat{\bar{\mathbf{x}}}_{\pi s^0})\widehat{\boldsymbol{B}} \tag{2.4}$$

underlined by the 'assisting' linear model

$$\begin{array}{ll} E_M(y_k) = & \sum_{j=1}^{p} \beta_j x_{jk} = \boldsymbol{x'\beta} \quad \text{and} \\ V_M(y_k) = & \sigma_k^2 \qquad\qquad\qquad\quad ; k = 1, ..., N \end{array} \tag{2.5}$$

where $\widehat{\mathbf{B}} = (\widehat{B}_1, ..., \widehat{B}_p)' = (\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k^{'} / \sigma_k^2 \pi_k)^{-1} \sum_s \mathbf{x}_k y_k / \sigma_k^2 \pi_k$. When $q_k = \sigma_k^2$ the estimator (2.3) is equivalent to (2.4) with $g_{\pi k} = 1 + (\overline{\mathbf{x}}_{s^0} - \hat{\bar{\mathbf{x}}}_{\pi s^0})\widehat{\boldsymbol{T}}^{\mathbf{-1}} \mathbf{x}_k / \sigma_k^2$ where $\widehat{\mathbf{T}}^{-1} = \sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k^{'}}{\sigma_k^2 \pi_k}$.

Now linking to the panel problem, a useful exemplar is the *post stratification* estimator

$$\hat{\bar{\mathbf{y}}}_{ps} = \sum_h W_h \overline{\mathbf{y}}_{s_h^2}$$

where $W_h = N_h/N$. However, to justify this estimator is to pick one of several invalid assumptions. From a systematic survey sampling perspective the $\hat{\bar{\mathbf{y}}}_{ps}$ is assisted by model (2.5) such that $\boldsymbol{\beta} = (\beta_1, ..., \beta_H)'$ and $\sigma_k^2 = \sigma_h^2 x_k \quad \forall k : k \in s_h^0$ for $h = 1, ..., H_{s^0}$ implying a stratification of the population $s^0 = \cup_{h=1}^{H} s_h^0$ defined by homogeneity in y. In addition we must choose to either (1) ignore the unknown selection probabilities π_k, (2) assume an equal probability selection process or (3) that strata $h = 1, ..., H$ overlap into an RHG type of selection mechanism. The first two are clearly false while the last brings back the coverage bias we wish to ignore.

However, what seems to be a dead end, does show us the appeal in the Web panel context of introducing models which explain y. Taking either of (invalid) assumptions listed above we can find that

$$B(\hat{\bar{\mathbf{y}}}_{ps}) \approx \sum_h W_h (\tilde{\mathbf{y}}_{s_h^0} - \overline{\mathbf{y}}_{s_h^0})$$

where $\tilde{\mathbf{y}}_{s_h^0} = N_h^{-1} \sum_{s_h^0} y_k \pi_k / \overline{\pi}_{1h}$ and the approximation is due to Taylor linearisation, with $\overline{\pi}_{1h}$ denoting the unknown average selection probability into panel segment s_h^1. The selection bias is a weighted sum of the strata biases, where each strata bias can be rewritten as

$$\tilde{\mathbf{y}}_{s_h^0} - \overline{\mathbf{y}}_{s_h^0} = \sum_{s_h^0} (\pi_{1k} - \overline{\pi}_{1h})(y_k - \overline{\mathbf{y}}_{s_h^0}) / \overline{\pi}_{1h} N_h.$$

While the post stratification and RHG estimator share the same structure, differing only in the weights $W_h \neq W_h^{'}$, the post stratification estimator (and more broadly- calibration or model assisted estimators) seem to avoid coverage bias by defining the strata on the population distribution of y. If its model holds the bias will tend to zero. However, the invalidity of the selection model assumption makes this approach questionable.

2.5 The Interpretation of Observational Studies into the Panel Case

The question of inference from self selected samples such as the Web panel samples, and more broadly to any non-probability sampling vis a vis the standard probability survey sampling theory is somewhat analogous to the comparison of observational studies against the scientific gold standard of randomized experiments for causal inference. By observational studies we mean here a study of a possible effect of a treatment on subjects, where the assignment of subjects into a treated group versus a control group is outside the control of the investigator. From analysis of observational studies we can draw ideas for analysis of non-probability samples. As a background for that, I review here key ideas of observational studies.

The web access panel is a survey sampling platform in purpose, but is assembled as an observational study, and it is not surprising that early attempts to address the panel question viewed Web panel surveys as a new type of observational studies (Boruch and Terhanian, 1996) and used popular observational study estimation tools. However, for the survey statistician the translation to our finite population estimand context and the specific data structure available in practice requires clarifications.

As noted above, observational studies refer to studies of causality based on convenience samples, that is non randomized treatment assignment. The problem is that substantial differences in the variable of interest across factors - tacitly implying causality - after closer examination however, may be partly or entirely be linked to variation of confounding covariates, rather than the specific characteristic studied. The fundamental problem is that unless a study sufficiently controls for conditions other than the experimental factor under study, distributional differences and subsequent causality statements may not be any more than an artefact (Lee, 2004).

In the basic setting of a causality problem for population unit k, the indicator S_k takes the value 1 if unit k takes the action and 0 otherwise[2]. If the unit takes the action we say that it is in the treatment group, otherwise it is in the control group. In the tradition of Rubin (78) and Holland (86), we define causal effects in terms of counterfactuals or potential outcomes. Specifically, $Y_k^{(1)}$ is what unit $k's$ outcome would have been if she had taken the action while the corresponding potential outcome for not taking the action is $Y_k^{(0)}$. The individual causal effect

[2] The unit may or may not be the decision-maker with regard to the action. Sometimes the action is taken by the unit, other times, others perform the action/treatment on the unit. In the panel survey context, action is a combination of active and passive decision making.

for unit k, then, would simply be $Y_k^{(1)} - Y_k^{(0)}$, the difference between what would happen under action and what would happen under inaction.

Note that the causal effects are defined in terms of potential outcomes, not observed quantities[3]. To connect the observed data and the potential outcomes, we make use of a *consistency* assumption: when we observe a unit taking an action, we observe their potential outcome for that action. Formally

$$Y_k = Y_k^{(s)} \quad \text{if} \quad S_k = s \tag{2.6}$$

where s is either 0 or 1. Without the consistency assumption the two distinct variables Y_k and $Y_k^{(s)}$ need not be the same if the potential outcomes depend on the actions of other units. The consistency assumption directly connects the observed and potential outcomes, ignoring any of these possible spillover effects[4]. In the design based survey sampling approach we assume the variable of interest is unknown but fixed (as well as all other population characteristics which may be measured), so the notation implicitly assumes consistency and sidesteps this issue.

The quantity of interest in casual inference is typically the average treatment effect (*ATE* or *ACE*)

$$E[\mathbf{Y}^{(1)} - \mathbf{Y}^{(0)}] = \overline{\mathbf{Y}}^{(1)} - \overline{\mathbf{Y}}^{(0)},$$

Now, consistency implies

$$E[Y_k^{(1)}|S_k = 1] = E[Y_k|S_k = 1]$$

where $E[Y_k|S_k = 1]$ is the observed average outcome among those who have been treated. But this implies nothing about the counterfactual quantity $E[Y_k^{(1)}|S_k = 0]$, which is what the average outcome would be for non treated had they been treated. The 'cost' of making the consistency assumption is that we cannot simultaneously observe a unit's outcome under both action and inaction. This is commonly known as the fundamental problem of causal inference and makes individual causal effects difficult to estimate without strong assumptions.

To overcome this we need the further assumption of conditional independence of the potential outcomes. That is the potential outcomes are independent of the

[3] To emphasize this some authors denote Y_k by Y_k^{obs} (Imbens and Rubin, 2010, chapter 5) or in a dose-response setting the outcome (the curve) is sometimes denoted by a different letter, say $D_k^{(1)}$ (Wasserman, 1999), to the observed value Y_k.

[4] Rubin (78) refers to this assumption as the stable unit treatment value assumption or SUTV It may seem that consistency is a definition rather than an assumption, but a few recent studies have called that assertion into question (Cole and Frangakis, 2009; VanderWeele, 2009; Pearl, 2010)

action, conditional on a set of covariates $\mathbf{X}_k$,

$$Y^{(1)}, Y^{(0)} \perp S_k | \mathbf{X}_k \tag{2.7}$$

where $\perp$ indicates conditional independence (Dawid, 1979). Political scientists call this assumption *no omitted variables*, economists call it *no selection on unobservables*, epidemiologists call it *no unmeasured confounders* and survey statisticians *non informative design.* In words, the assumption means that the distribution of the potential outcomes is the same for those who have and have not been treated .

Augmenting consistency with ignorability allows us to fully connect the observed data and the potential outcomes. For $s = 1$ or 0

$$\begin{aligned} E(Y_k^{(s)}) &= \int y_k^{(s)} f(y^{(s)}) dy^{(s)} \\ &= \int E(Y_k^{(s)} | \mathbf{X}_k = \mathbf{x}) f(\mathbf{x}) d\mathbf{x} \\ &= \int E(Y_k^{(s)} | \mathbf{X}_k = \mathbf{x}, S_k = s) f(\mathbf{x}) d\mathbf{x} \\ &= \int E(Y_k | \mathbf{X}_k = \mathbf{x}, S_k = s) f(\mathbf{x}) d\mathbf{x} \qquad (2.8) \end{aligned}$$

which some term the '*g-computation algorithm formula*' (Robins et al., 1999) or the *g-equation* (Wasserman, 1999) and is the standard adjustment for a confounder as discussed by many statisticians, notably Rubin and Rosenbaum. The strength of *g-equation* is in the sequential framework I discuss in chapter 3. Equation (2.8) suggests an estimator which relies on a model of the distribution of the conditional mean Y over the population distribution of $\mathbf{X}$. I discuss this approach in the following section.

We are now ready to reinterpret these ideas to the Web panel context. First, 'treatment' here is participation in a Web panel and response to a specific survey. Non treatment is is simply non participation in a Web panel. We assume that conditional on observed covariates $Y_k^{(0)} = Y_k^{(1)}$ and our focus is on estimating the fixed population parameter $\overline{\mathbf{y}}_{s^0}^{(1)}$, the average outcome if everyone in the population would join our panel and be measured by a survey, or when a superpopulation model is envoked $E(\overline{\mathbf{Y}}_{s^0}^{(1)})$, but otherwise the discussion on estimation of the ATE is the same but for the objective- instead of estimating *treatment effect* we are measuring (or attempting to correct for) *estimation bias* by relying on the conditional independence assumption. Here, (2.7) would be violated, for instance, if Liberal Democrat supporters were more likely than other party supporters to join and respond to the panel survey and we failed to control for political party affiliation. In a case of a political poll, joining the panel would be correlated with potential outcomes and inflate the real support for Liberal democratic party in the population.

In some cases $s = 0$ will denotes not the counterfactual but rather $Y_k^{(0)}$ indicates the response through a traditional survey setting such as a randomly selected sample surveyed by telephone. In such instances the two states $s = 0$ or 1 do not complement each other, are not counterfactual, and may happen simultaneously. This causes a potential overlap with some implications to the estimation procedures which I discuss in detail in chapter 4 along with the estimation methods relying on a randomly selected 'reference' survey sample.

2.6 Survey sampling as a Missing Data Problem

In the tradition of Rubin (1976) it is useful to describe the survey sampling problem as a missing data problem where we want to draw inference on the full population distribution $D = (\mathbf{Y}, \mathbf{X},\mathbf{S})$ while we observe only a subset of the full population $D_s = (\mathbf{Y}_s, \mathbf{X},\mathbf{S})$. This is not necessarily with aim to model the full distribution but rather, outlying the full distribution will clarify the assumptions and mechanism for valid inference. These conditions vary depending on the estimand, the mode of inference (Survey sampling, Likelihood, Bayesian, Generalized Estimation Equations), the estimation procedure and the estimator used. Further discussion on the idea of describing survey sampling inference as a general missing data problem can be found in Little (1983); Smith (1983); Smith and Sugden (1988) and Gelman et al. (2004, chapter 7).

In general the full population $D = (\mathbf{Y}, \mathbf{X},\mathbf{S})$ with data d can be thought to have distribution

$$\begin{aligned} f(d) &= f(\mathbf{y}|\mathbf{x})p(\mathbf{s}|\mathbf{y},\mathbf{x})f(\mathbf{x}) \text{ or} \\ &= f(\mathbf{y}|\mathbf{s},\mathbf{x})p(\mathbf{s}|\mathbf{x})f(\mathbf{x}). \end{aligned} \tag{2.9}$$

Assuming throughout that values $\mathbf{x}$ are known, the analyst observes the dataset d_s with distribution

$$f(d_s) = \int f(d) d\mathbf{y}_{\overline{s}} \tag{2.10}$$

where $\mathbf{y}_{\overline{s}}$ is the set of unobserved responses from $\mathbf{Y}_{\overline{s}}$, the unsampled units of the population. In the case $\mathbf{Y}$ and $\mathbf{S}$ are conditionally independent given $\mathbf{X}$, the observed data distribution will follow

$$f(d_s) = f(\mathbf{y}_s|\mathbf{x})p(\mathbf{s}|\mathbf{x})f(\mathbf{x}). \tag{2.11}$$

As Tsiatis and Davidian (2007) discuss we can outline broadly three estimation strategies, by positing different statistical models that may have generated the observed data, by making different assumption on the components of (2.9). These broadly characterize three estimation approaches (m, π and πm mentioned earlier), that is

1. Make no assumptions on the forms of $f(\mathbf{x})$ or $p(\mathbf{s})$, leaving these entirely unspecified. Make a specific assumption on $f(\mathbf{y}|\mathbf{x})$, for example that $E(\mathbf{Y}_k|\mathbf{X}_k = \mathbf{x}) = m(\mathbf{x}_k; \boldsymbol{\beta})$ for some given function $m(\mathbf{x}; \boldsymbol{\beta})$ depending on parameters $\boldsymbol{\beta}$ $(P^{\beta} \times 1)$.
2. Make no assumption on the forms of $f(\mathbf{x})$ or $f(\mathbf{y}|\mathbf{x})$, but make a specific assumption on $p(\mathbf{s})$, for example that $Pr(S_k = 1|\mathbf{X}_k = \mathbf{x}) = E(S_k|\mathbf{x}) = \pi(\mathbf{x}_k; \boldsymbol{\alpha})$ for some given function $\pi(\mathbf{x}; \boldsymbol{\alpha})$ depending on parameters $\boldsymbol{\alpha}$ $(P^{\alpha} \times 1)$ and also that $0 > Pr(S_k = 1|\mathbf{X}_k = \mathbf{x}) > 1$ for all $\mathbf{x}$.
3. Make no assumption on the form of $f(\mathbf{x})$, but make specific assumptions on $f(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{s}|\mathbf{x})$, for example, that assumptions outlined in preceding items 1. and 2. hold.

denote $f^0(d_s)$ as the true joint density generating the observed data, and let m_0 and π_0 denote receptively the true functions $E(\mathbf{Y}|\mathbf{x})$ and $E(S|\mathbf{x})$ corresponding to $f^0(d_s)$. If model 1. is correct then $m_0 = m(\mathbf{x}_k; \boldsymbol{\beta})$ and if model 2. is correct then $\pi_0 = \pi(\mathbf{x}_k; \boldsymbol{\alpha})$. In the next three sections I discuss such estimation procedures starting with the case of modelling the unknown selection process.

2.7 Modelling the Selection Process- $\pi-$Estimation

As discussed in section 2.3 classic survey sampling theory relies explicitly on a known $p(s)$ and conditioning on the population values $\mathbf{y}_{s^0}, \mathbf{x}_{s^0}$ which lends the term 'fixed population inference'. This reliance on the single distribution $p(s)$, however complicated, makes it operationally appealing for official statistic agencies and survey based research companies where large amount of items $\mathbf{Y}$ are processed. However, as shown, the fixed population inference suffers from a coverage bias over the non 'Web-covered' subset of the population, and furthermore the panel selection process is unknown and so $p(s)$ must be modelled. In the following I discuss the idea of π-estimation methods such as the HT- estimator from a superpopulation perspective.

I start with discussing in detail the propensity score estimation strategy, which as noted by the AAPOR report on Online Panels (Baker et al., 2010) attracted arguably the greatest amount of attention in Online panel survey research and estimation. The approach was first introduced for use in online panels by Harris Interactive (Terhanian et al., 2000; Taylor et al., 2001) and further examined by Lee (2004, 2006); Lee and Valliant (2009); Schonlau et al. (2004); Loosveldt and Sonck (2008); Rivers (2007) and others.

The analyst observes d_s from $D_s = (\mathbf{Y}_s, \mathbf{X}, \mathbf{S})$ and denote the conditional selection given covariates by $\pi(\mathbf{x}) = p(s|\mathbf{x}_{s^0})$. When s depends only on fully observed covariates, we call $\pi(\mathbf{x})$ the propensity score and a conditional independence between Y and s exists, that is

$$f(d_s|\mathbf{x}_{s^0}) \quad = \quad f(\mathbf{y}_s|\mathbf{x}_{s^0})p(\mathbf{s}|\mathbf{x}_{s^0}). \tag{2.12}$$

Commonly, it is assumed that

$$p(s|\mathbf{x}_{s^0}) = \Pi_{k\in s}\pi(\mathbf{x}_k)\Pi_{k\in\bar{s}}\left[1-\pi(\mathbf{x}_k)\right] \tag{2.13}$$

with $\pi(\mathbf{x}_k)$ being the individual unit selection probability. The strict independence assumption, which implies a form of *Poisson selection process* Särndal et al. (1992) popular in survey literature, is not essential but simplifies notation and discussion.

To understand the specific mechanics when applying propensity score estimation, it is instructive to start with the original motivation for the method- finding a low dimensional substitute to the conditioning covariate set $\mathbf{x}$ while still allowing (2.12) to hold. When $\mathbf{X}$ takes a high dimension, such substitutes, termed *balancing scores*, can facilitate more practical estimation strategies.

Formally, an appropriately constructed balancing score $b(\mathbf{x})$ has the property

$$\mathbf{X}_k \perp S_k | b(\mathbf{x}_k).$$

(Rosenbaum and Rubin, 1983, sec 2.2) discuss the special role of propensity scores within the family of balancing scores and show that (i) the propensity score is a balancing score, so $\mathbf{X}\perp S|\pi(\mathbf{x})$, that (ii) it is the coarsest such balancing score in the sense that $\pi(\mathbf{x}) = f\{b(\mathbf{x})\}$, for any balancing score $b(\mathbf{x})$ where the finest score is $\mathbf{X}$ itself, thus it is the most efficient score, and (iii) that if $\mathbf{X}$ provides enough information for (2.12) to hold, then conditional independence holds given any balancing score including the propensity score. That is

$$f(d_s|\pi(\mathbf{x})) = f(\mathbf{y}_s|\pi(\mathbf{x}))p(\mathbf{s}|\pi(\mathbf{x})) \tag{2.14}$$

which can be derived by showing that $p(s|\pi(\mathbf{x}), \mathbf{y}) = E\left\{E(S|\pi(\mathbf{x}), \mathbf{x}, \mathbf{y})|\pi(\mathbf{x}), \mathbf{y}\right\}$ which is equal to $\pi(\mathbf{x})$ because of conditional independence given $\mathbf{X}$.

An immediate consequence is that

$$\begin{aligned}
E_{\pi(\mathbf{x})}\{E(\mathbf{Y}_s|s,\pi(\mathbf{x}))|\pi(\mathbf{x})\} &= \int\int yf(y|s,\pi(\mathbf{x}))f(\pi(\mathbf{x}))dyd\pi(\mathbf{x}) \\
&= \int y\int f(y|\pi(\mathbf{x}))f(\pi(\mathbf{x}))d\pi(\mathbf{x})dy = E(\mathbf{Y})
\end{aligned}$$

and under *probabilistic selection*, that is $0 < \pi(\mathbf{x}) \leq 1$ valid inference is possible. Probabilistic selection is closely related to the concept of *common support* in matching literature, *positivity* in causal inference literature and *measurability* in sampling theory and simply requires each unit to have a chance to participate in a the survey. This lets the entire support of $\mathbf{x}$ to be represented in s while if a particular subpopulation has zero probability of being in the sample, estimates for this subpopulation must by necessity rely on extrapolation through a supporting m-model.

Several competing estimation procedures have been popularized following Rosenbaum and Rubin (1983) paper, including weighting, classification, matching and covariance adjustment; all of which have been applied to the Web panel problem. In the following I touch on the first two starting with the most widely cited- the $\pi-$weighted estimator. For further reading see Rubin (2006); Rosenbaum (2002); Abadie and Gardeazabal (2008); Dehejia and Wahba (2002, 1999); Olson (2006). For specific application to the Web panel case see Isaksson and Lee (2005); Lee and Valliant (2009); Rivers (2007) and Schonlau et al. (2004).

The $\pi-$weighted estimator can be viewed as a model based version of the $HT-$ estimator introduced in section 2.3 and takes the same form

$$\hat{\bar{\mathbf{Y}}}_\pi = N^{-1}\sum_{k\in s}\mathbf{y}_k\hat{\pi}_k(\mathbf{x})^{-1} \tag{2.15}$$

where $\hat{\pi}_k(\mathbf{x}) = \pi(\mathbf{x}_k;\hat{\boldsymbol{\alpha}})$ are consistent estimators of the unit selection probabilities. As with the HT- estimator an alternative version is to replace the denominator with $\hat{N}_\pi = \sum_{k\in s}\hat{\pi}_k(\mathbf{x})^{-1}$. For some selection processes $p(\cdot)$ the two are identical, but in general the latter is superior as (i) it is more efficient with both parts of the ratio increasing in size for large sample sizes, decreasing with small sample sizes. (ii) It is a *bounded estimator* - as a convex combination of y_k's it always lies in the interval $[y_{min}, y_{max}]$ with endpoints the minimum and maximum observed Y-values and so will fall in the parameter space of $\mathbf{y}_s$ with probability 1 (see Robins et al., 2007). The latter, classic Horvitz Thompson, estimator is not bounded (a property Basu, 1971 famously exploited in attacking survey sampling theory). Lastly, (iii) estimation of the population mean is feasible even when the population size, N is unknown.

Estimator (2.15) can be motivated by the *g-equation* (2.8)

$$\begin{aligned} E(\overline{\mathbf{Y}}_{s^0}) &= N^{-1}\sum_{s^0}\int \frac{E(Y_k|\mathbf{X}_k=\mathbf{x})E(S_k|\mathbf{X}_k=\mathbf{x})}{E(S_k|\mathbf{X}_k=\mathbf{x})}f(\mathbf{x})d\mathbf{x} \qquad (2.16)\\ &= N^{-1}\sum_{s^0}\int \frac{E(Y_kS_k|\mathbf{X}_k=\mathbf{x})}{\pi(\mathbf{x}_k)}f(\mathbf{x})d\mathbf{x} = N^{-1}\sum_{s^0} E\left(\frac{Y_kS_k}{\pi_k}\right) \end{aligned}$$

The idea of inverse weighting is to recover the joint distribution of $(\mathbf{X}, Y)$ by attaching weight $\propto \pi(\mathbf{x})^{-1}$ to each point in $\{(\mathbf{X}_k, Y_k) : S_k = 1\}$. This estimation principle is visualized in figure 2.1. Early discussion on the strengths and weaknesses of direct weighting by inverse estimated propensity scores can be found in Little (1986) and Little and Rubin (1987, p.58). See Tan (2006) for a likelihood formulation.

Assuming $\hat{\pi}_k(\mathbf{x})$ are consistently estimated consider now the the potential bias of the $\pi-$weighted estimator. Let $Pr(S_k = 1|\mathbf{x}, y) = \pi_k(\mathbf{x}, y)$ and denote the joint density $(\mathbf{X}, Y)$ in the population by $f(\mathbf{x}, y)$, while in the sample by $f(\mathbf{x}, y|s) = \frac{\pi_k(\mathbf{x},y)f(\mathbf{x},y)}{\pi(\cdot,\cdot)}$ where $\pi(\cdot,\cdot) = \int\int \pi(\mathbf{x}, y)f(\mathbf{x}, y)dyd\mathbf{x}$ is the average selection probability. Also let $Cov(\cdot)$ denote covariance, $\sigma(\cdot|\mathbf{x})$ the conditional standard deviation and $\rho(\cdot|\mathbf{x})$ the conditional correlation coefficient. By noting that $\int\int \pi(\mathbf{x})^{-1}\pi(\mathbf{x}, y)f(\mathbf{x}, y)dyd\mathbf{x} = 1$ then the potential bias of $\hat{\overline{\mathbf{y}}}_\pi$ is

$$\begin{aligned} B(\hat{\overline{\mathbf{Y}}}_\pi) &= E\left(\frac{\sum_{s^0} S_kY_k/\pi_k(\mathbf{x})}{\sum_{s^0} S_k\pi_k(\mathbf{x})^{-1}}\right) - E(\overline{\mathbf{Y}})\\ &\approx \int\int y\left(\frac{\pi(y,\mathbf{x})}{\pi(\mathbf{x})} - 1\right) f(y,\mathbf{x})dydx\\ &= Cov\left(Y, \frac{\pi(Y,\mathbf{X})}{\pi(\mathbf{X})}\right)\\ &= \int\left\{\sigma(Y|\mathbf{x})\frac{\sigma\{\pi(\mathbf{x},Y)|\mathbf{x}\}}{\pi(\mathbf{x})}\rho\{Y,\pi(\mathbf{x},Y)|\mathbf{x}\}\right\} f(\mathbf{x})d\mathbf{x} \qquad (2.17) \end{aligned}$$

where the approximation is of order $o(n^{-1})$.

Under correct specification $\pi(y, \mathbf{x}) = \pi(\mathbf{x})$ and the bias is exactly zero. Otherwise, the magnitude of $B(\hat{\overline{\mathbf{Y}}}_\pi)$ will depend both on the level of departure from the $\pi-$model and distributional properties of $f(\mathbf{y}|\mathbf{x})$ and $p(s|\mathbf{x})$. Specifically, bias will be small if at least one of the factors is small

(i) the conditional standard deviation of Y $\sigma(Y|\mathbf{x})$ is small for all $\mathbf{x}$, which happens if $Y \approx E(Y|\mathbf{x})$ - that is the Y values are predicted well by the $\mathbf{X}$, or

(ii) $\pi(\mathbf{x}, y)$ varies little with y for fixed $\mathbf{x}$ which happens if at least approximately, $p(s|\mathbf{x}, y) \approx p(s|\mathbf{x})$ for all $\mathbf{x}$ and y, which happens in a good prediction model of S by $\mathbf{x}$, or

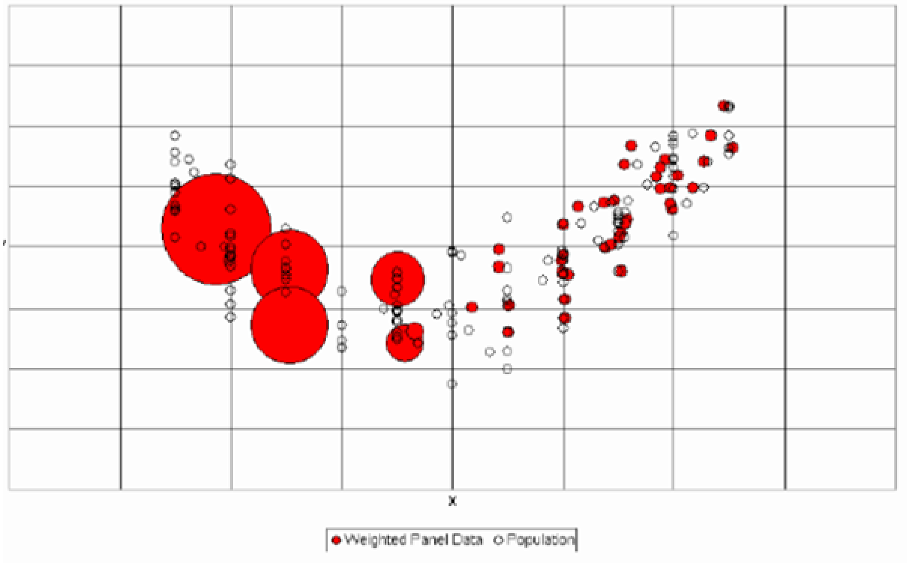


Figure 2.1: The π-estimator: $\pi-$ estimation recovers the joint distribution of $(\mathbf{X}, Y)$ by attaching weight $\propto \pi(\mathbf{x})^{-1}$ to each point in $\{(\mathbf{X}_k, Y_k) : S_k = 1\}$ but is sensitive to cases where $\pi \approx 0$.

(iii) the conditional correlation $\rho\{Y, \pi(\mathbf{x}, \mathbf{Y})|\mathbf{x}\}$ is close to zero for all $\mathbf{x}$.

The main objection to $\hat{\overline{\mathbf{Y}}}_\pi$ is its sensitivity to the common support assumption. Direct weighting can be highly unstable as respondents with very low values of π are sharply adjusted - an idea visualized in figure 2.1 as well. This sensitivity can be seen both in the potential bias (2.17) which is a function of π^{-1} and in the expected variance; for example the selection-model variance of the $\pi-$estimator with known selection probabilities is $V(\hat{\overline{\mathbf{Y}}}_\pi|\mathbf{y}) = \sum_{s\in S} p(s)(\hat{\overline{\mathbf{Y}}}_\pi - E(\hat{\overline{\mathbf{Y}}}_\pi))^2$ which under Poisson selection is

$$V(\hat{\overline{\mathbf{Y}}}_\pi) \approx N^{-2} \sum_{k\in U} \left(\pi_k^{-1} - 1\right) (y_k - \overline{\mathbf{y}})^2 \tag{2.18}$$

which also is inflated in regions where $\pi(\mathbf{x}) \approx 0$.

At the extremes when there is complete separation in the distribution $f(\mathbf{x})$ between s and $\overline{s}$ the estimator breaks down and is not defined. An interesting counterargument to this criticism (Tan, 2007) puts that this instability in fact properly reflects the separation in $f(\mathbf{x})$ and thus the lack of information on $\overline{\mathbf{y}}_{s^0}$ in the data from the non sampled units. It is a consequence of transparency rather than a disadvantage and as such, makes clear that there is no basis for inferring the expected outcome of non selected units. In contrast ML, multiple imputation and other m-estimation methods will still be produced even in complete separation because they are based implicitly on extrapolation. An additional strength of π-estimation is that it is modelled over $(\mathbf{S}, \mathbf{X})$ both assumed observed over the entire finite population. This is in contrast to $m-$estimation which is fitted over $(\mathbf{Y}_s, \mathbf{X}_s)$, a major weakness I discuss in the following section.

In practice after choosing the $\pi-$weighted estimator, we must decide (i) the covariate set $\mathbf{X}$ to include in the $\pi-$model and (ii) the relevant model fitting/goodness of fit tests.

As for covariate selection, the fundamental recommendation, supported by the potential bias (2.17) is to include all covariates associated either to Y or S . This is echoed in advice such as in Rubin and Thomas (1996) and Heckman et al. (1998) who argue that there is no distinction between highly predictive covariates and weakly predictive ones in the performance of propensity score adjustment and suggest including all observable covariates. Drake (1993) finds that misspecifications of the propensity score in terms of functional form have much smaller biases than similar misspecifications in $m-$model estimation. Millimet and Tchernis (2009) report a simulation that focuses on, essentially, the functional form of the propensity score and conclude that the penalty for overfitting is minimal. More recently Rubin (2009, p.1421) reiterates by stating that *not* controlling for an observed covariate is bad practical advice in all but the most unusual circumstances.

On the other hand, work on the effects of including covariates (into the propensity model) that have only weak or no effect on either the selection or outcome variables has been published countering the above consensus. For example in a simulation study Augurzky and Schmidt (2001) include a set of variables $\mathbf{X}^s$, which strongly influence S, but do not or only weakly determine Y and a second set $\mathbf{X}^y$, that influence Y, but are irrelevant to S. Their result indicates that including both sets of covariates results in an increase in the MSE and recommend including only highly significant variables in the propensity score equation. Similarly, Brookhart et al. (2006) find that one should include covariates that are thought to be related to $\mathbf{Y}$, whether or not they are related to the selection, while the opposite of including covariates only related to S increases the variance of and estimator without decreasing its bias. Clarke (2005, 2009); Clarke et al. (2011) persuasively demonstrates that, for a misspecified model, the inclusion of additional control variables, which influence both S and Y increases the estimation bias dramatically.

Given a set of covariates, the particular model fitting process common in propensity score literature is understood by recalling that the goal is to create balance, on observed covariates, between the selected and non selected groups. Starting with Rosenbaum and Rubin (1983, 1984) and Rubin (1997) the process is of recycling between checking for balance on the covariates (e.g. by t-tests) and reformulating the propensity score. For example, when large mean differences in an important covariate are found to exist between the two groups, even after its inclusion in the model, then the square of the variable and interactions with other variables can be tried. Ho et al. (2007) put this idea as the propensity score tautology ': *The estimated propensity score is a balancing score when we have a consistent estimate of the true propensity score; We know we have a consistent estimate of the score when matching on the propensity score balances the raw covariate.* More recent balancing tests include testing for mean differences within strata of the propensity score (Dehejia and Wahba, 1999, 2002) or testing for the joint equality of covariate means across groups using the Hotelling test or F-test (Smith and Todd, 2005). See Lee (2006) for further details.

Kosuke et al. (2006), Sekhon (2011) have criticized the use of an absolute cut off point after which balance is achieved noting that hypothesis tests are not monotonic functions of balance and are driven in part by factors other than balance such as the number of observations, the ratio of selected and non selected units, and their respective variances. Thus tests such as the t test can get better while balance gets worse. They suggest alternatives such as using empirical QQ plots to compare the full empirical distributions for the two groups (either univariate, or for $\pi(\mathbf{x})$ as it offers a good low dimensional summary) and numerically summarize these plots with mean and maximum deviation between the two distributions on the scale of the variables being measured.

Still, the model fitting discussion remains within the boundaries of balancing properties and in that light it is not surprising that in a a systematic literature review (Weitzen et al., 2004) of the approximately 50 published articles inspected only 6 considered usual modelling steps such as evaluating the goodness of fit of a logistic regression model, while the majority inspected the balance achieved after propensity score modelling.

Beyond direct weighted π-estimators, arguably the most popular application of the propensity score is *subclassification estimation* (term coined by Cochran, 1968)*,* in essence a model based version of the RHG estimator of section 2.3 . For any balancing score, the basic setting involves ordering its values and forming classes according to the values $b(\mathbf{x}_k)$, $k = 1, ..., N$. Specifically let S_h define a set of values of $b(\mathbf{x})$ so that $b(\mathbf{a}) \in S_h$ implies units with $\mathbf{x} = \mathbf{a}$ fall into subclass $h = 1, ..., H$. This allows classification of the population into strata $s^0 = \cup_{h=1}^{H} s_h^0$ and similarly classifies s the selected sample. Denote the respective population and samples sizes of these classes by N_h and n_h for all $h = 1, ..., H$.

When subclasses are perfectly homogeneous in $b(\mathbf{x})$, the model covariates are perfectly balanced in the sense that for large populations with $\frac{n_h}{N_h} \approx 0$ the distribution over s_h^0 and s_h is identical. And so a weighted average, with population total weights, is an unbiased estimator for the population average. However, as $\pi(\mathbf{x})$ is the coarsest balancing score in the sense that $\pi(\mathbf{x}) = f\{b(\mathbf{x})\}$ for some function f than we expect residual bias due to imbalances in model covariates.

Noting that assignment is ignorable given $b(\mathbf{x})$ this residual bias of the covariates after subclassification can be quantified by

$$B(\mathbf{x}) = \sum_{h=1}^{H} W_h[\int E(\mathbf{x}|b)\{p(b|s_h\} - p(b|\bar{s}_h\}db]$$

where $\bar{s}_h = s_h^0 - s_h$ within each $h = 1, ..., H$ which can be approximated to s_h^0 again assuming a small sample fraction. Thus the residual bias is a weighted sum of the difference between the distribution of the balancing score in the sample and that of the population.

A common problem which often surprises applied researchers is that such imperfect adjustment carries potentially the undesirable property of increasing the bias for some linear functions of $\mathbf{x}$ even if all univariate means are closer post subclassification. That is, there exists a vector $\mathbf{w}$ such that $\mathbf{w}B > \mathbf{w}B^*$ where B^* represents the original covariate imbalance. With this issue in mind Rosenbaum and Rubin (1983) give the theoretical conditions (the equal percentage bias reduction property) where the subclassification does not increase the bias.

To limit the residual bias one can create numerous subclasses so as to refine units in each subclass to have almost identical propensity scores and negate the residual bias. The norm, however, is to adopt five subclasses based on quintiles of the propensity scores[5]. This is based on work by Cochran (1968) and Rosenbaum (1984) investigating subclassification on $\mathbf{x}$ and $\pi(\mathbf{x})$ respectively.

The most obvious application of the approach (e.g. Little, 1986) involves the following steps: (i) Estimate $\boldsymbol{\alpha}$ and calculate $\hat{\pi}_k = \pi(\mathbf{x}_k; \hat{\boldsymbol{\alpha}})$ for all k; (ii) Order the estimated probabilities over s^0 and form H classes according to the quantiles of $\hat{\pi}_k$, where the hth quantile $\hat{q}_h$, $h = 1, ..., H$, is such that the proportion of $\hat{\pi}_k \leq \hat{q}_h$ is roughly h/H, $\hat{q}_0 = 0$, and $\hat{q}_H = 1$; (iii) Within each class calculate $\hat{\bar{\pi}}_h = (n_{sh} + n_{\bar{s}h})^{-1}(\sum_{s_h} \hat{\pi}_k + \sum_{\bar{s}_h} \hat{\pi}_k)$ the average selection probability; and (iv) Estimate $\bar{\mathbf{y}}_{s^0}$ by simple $\pi-$estimator $\hat{\bar{\mathbf{Y}}}_{s^0} = \sum_s y_k \hat{\pi}_k^{-1} / \sum_s \hat{\pi}_k^{-1}$ which here can be expressed as

$$\hat{\bar{\mathbf{Y}}}_{s^0} = \sum_h \hat{W}_h \bar{\mathbf{y}}_{sh} \;\; \text{where} \, \hat{W}_h = n_{sh} \hat{\bar{\pi}}_h^{-1} / \sum_h^H n_{sh} \hat{\bar{\pi}}_h^{-1}.$$

An alternative popular in the Web panel literature (Lee, 2004; Valliant and Dever, 2011; Isaksson et al., 2005) is underlined by the following rationale. Assume

[5] Rather than classes based on homogeneity of π. This preference can be understood as a preference of inexact classification over incomplete classification, see Rosenbaum.Rubin:85b for a relevant discussion.

H non overlapping classes, then the sample average can be written as

$$\overline{\mathbf{y}}_s = \sum_{h=1}^{H} n_{sh}\overline{\mathbf{y}}_{sh}.$$

If we adjust each sampled unit by weights $W_k = \frac{N_h/N}{n_{sh}/n}$ for $k \in s_h^0$, $h = 1, ..., H$ then we get a post stratified estimator

$$\hat{\overline{\mathbf{Y}}}_{s^0} = \sum_{h=1}^{H} \frac{N_h}{N}\overline{\mathbf{y}}_{sh}$$

which can be seen to be unbiased as $E(\mathbf{Y}) = \int E(\mathbf{Y}|\mathbf{x}, s)f(\mathbf{x})d\mathbf{x}$ where here $f(\mathbf{x}) = \frac{N_h}{N}$. Note the relationship between $\pi-$weights and these adjustment weights

$$W_k = \frac{f(\mathbf{x})}{f(\mathbf{x}|s)} = \frac{f(\mathbf{x})p(s)}{f(\mathbf{x})p(s|\mathbf{x})} \propto \pi(\mathbf{x})^{-1}$$

for each class h. Thus, when $\pi(\mathbf{x})$ are homogeneous in h the post stratification should remove effectively the selection bias.

In practice we calculate

$$\hat{\overline{\mathbf{Y}}}_{s^0} = \sum_{h=1}^{H} \hat{W}_h\overline{\mathbf{y}}_{sh} \text{ where} \hat{W}_h = \frac{\hat{N}_h}{N}$$

with $\hat{N}_h$ denoting the number of population units that fall into class h of estimated unit selection probabilities.

An important point is that numerous published work demonstrating the estimator suggests defining the classes by quantiles of $\pi(\mathbf{x})$ over the combined set of sampled and non sampled units, (see Terhanian et al., 2000; Taylor et al., 2001; Schonlau et al., 2002; Lee, 2004; Isaksson et al., 2004; as well as Valliant, 2009 among others). However, defining classes by equal unit counts by definition suggests $\hat{W}_k \approx 1/H$ depending on the number of classes predetermined and results with the simple sample average remaining unadjusted. Anecdotally Valliant and Dever (2011), when studying web panel inference using a reference survey, have published a critique of the estimator supplemented with a simulation study stating that '...although the mathematics needed to anticipate the large biases ... are not obvious, the message ... is clear—propensity post stratification should not be used.' but fail to recognize that the fault is in the specific mechanics of classification by quantiles they choose rather than the approach.

I conclude with setting the previous discussion in a wider family of π-weighted estimators, all which can be applied to the Web panel problem. The following is based on work attributed to Robins and Rotnitzky (1994) together with colleagues, who developed an elegant semiparametric theory underlying estimation equation estimators from general missing data problems. I develop this approach in more detail in the next section dealing with m-estimators where I discuss concepts such as the *influence function* and a *linear estimator* which are left here undefined.

Let $p_0(d)$ be the true joint density, assume that $E(S_k|\mathbf{X}_k) = \pi(\mathbf{x}_k; \boldsymbol{\alpha})$ for all $k \in s^0$ leaving the rest of the components of $p(d_k)$ unspecified and that conditional independence holds then the theory of Robins et al. shows that all asymptotically linear estimators of $E(Y) = \mu$ under model $\pi_0(\mathbf{x}_k)$ of the true function $E(S_k|\mathbf{x}_k)$ have influence functions of the form

$$\frac{S_k Y_k}{\pi_0(\mathbf{x}_k)} - \frac{S_k - \pi_0(\mathbf{x}_k)}{\pi_0(\mathbf{x}_k)} h(\mathbf{x}_k) - \mu$$

for arbitrary function $h(\cdot)$ of $\mathbf{x}$, which suggests that any estimator of the population average consistent and asymptotically normal under π_0 must be asymptotically equivalent to an estimator of the form

$$N^{-1} \sum_{k=1}^{N} [\frac{S_k Y_k}{\pi(\mathbf{x}_k; \boldsymbol{\alpha})} - \frac{S_k - \pi(\mathbf{x}_k; \boldsymbol{\alpha})}{\pi(\mathbf{x}_k; \boldsymbol{\alpha})} h(\mathbf{x}_k)]$$

when parameters $\boldsymbol{\alpha}$ are known. See Tsiatis and Davidian (2007, p. 570-571) for the appropriate formulation when parameters are estimated.

Setting $h = -\mu$ or $h = 0$ result in the estimators $\sum_s Y_k \pi_k^{-1} / \sum_s \pi_k^{-1}$ and $\sum_s Y_k \pi_k^{-1}/N$ respectively. Another interesting example is when we set $h = -E(Y(1 - \pi(\mathbf{x}))/E(1 - \pi(\mathbf{x}))$ giving a π-imputation type estimator

$$\hat{\bar{\mathbf{Y}}}_{s^0} = f_s \bar{\mathbf{y}}_s + (1 - f_s) \hat{\bar{\mathbf{y}}}_{\pi\bar{s}}$$

with $\hat{\bar{\mathbf{y}}}_{\pi\bar{s}} = \sum_s \mathbf{y}_k \hat{\pi}_k^{-1}(1 - \hat{\pi}_k) / \sum_s \hat{\pi}_k^{-1}(1 - \hat{\pi}_k)$ which is approximately unbiased for the average of the unselected population set.

Using the fact the expectation of the square of the influence function is the asymptotic variance, the most efficient h (limiting $h(\mathbf{x})$ to be constant) can be found resulting with the estimator

$$\hat{\bar{\mathbf{Y}}}_{s^0} = \sum_s Y_k \hat{\pi}_k^{-1}(1 - C_s \hat{\pi}_k^{-1}) / \sum_s \hat{\pi}_k^{-1}(1 - C_s \hat{\pi}_k^{-1})$$

where $C_s = \sum_{s^0}(\hat{\pi}_k^{-1} S_k - 1) / \sum_{s^0}\{(\hat{\pi}_k^{-1} S_k - 1)\}^2$, which is the normal ratio $\pi-$estimator with each weight $\hat{\pi}_k^{-1}$ adjusted by a quantity estimating the deviation of $\pi_k^{-1} S_k$ from it's expectation of 1. For large samples C_s should be close to zero while for small samples the adjustment is a way of countering this variance. See Lunceford and Davidian (2004, page 2943) and Tsiatis and Davidian (2007) for further details.

An interesting result by Lunceford and Davidian (2004) show by estimation equations theory that estimating $\pi_k(\mathbf{x}; \boldsymbol{\alpha})$, even if its true value is known, leads to smaller (large-sample) variance for these estimators than using the true value. The reason is that the fitted values balance for random sampling errors as well as the panel selection bias. Thus, even if the functional form of the propensity score is known exactly, it is beneficial from an efficiency standpoint to estimate it anyway.

The move from HT-estimation over a fixed population to $\pi-$estimator paradigm invoking a superpopulation does allow at least hypothetically a way to avoid the coverage constraint and achieve valid inference under correct specification of the selection process. Still, in chapter 3 I take one step further and model $p(s)$ as a sequence of selection phases where the first models Internet usage. Subsequently, the 'Web covered' segment of the population- and associated covariates such as Internet related characteristics- become stochastic with each population unit having a positive probability of joining the segment. This will allow a more complete solution to this fundamental problem limiting population parameter estimation from a Web based data selection platform.

2.8 Modelling the Outcome of Interest- $m-$Estimation

In $m-$estimation, the model for the survey outcomes $\mathbf{Y}$ is used to predict the non-sampled values of the population, allowing estimation of finite population quantities. The approach does not overtly consider a distribution for S and they are not the basis for the inference. $m-$estimation procedures are not widely used in Web panel published work, which can be explained by the dominance of $\pi-$estimation techniques in survey sampling community.

Classically, the key concept underpinning the reliance only on $m-$models for inference is the idea of ignorability of Missing at Random (MAR) selection processes. The basic idea can be formulated under a likelihood perspective by first specifying a joint probability model over the entire finite population $d_{s^0} = (\mathbf{y}_{s^0}, s)$ given parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and covariates $\mathbf{x}_{s^0}$

$$f(d_{s^0}|\mathbf{x}_{s^0}; \boldsymbol{\theta}, \boldsymbol{\phi}) = f(\mathbf{y}_{s^0}|\mathbf{x}_{s^0}; \boldsymbol{\theta}) p(s|\mathbf{y}_{s^0}, \mathbf{x}_{s^0}; \boldsymbol{\phi}) \tag{2.19}$$

where the parameter vector $(\boldsymbol{\theta}, \boldsymbol{\phi})$ denote the parameters of the conditional distributions of the outcome of interest matrix and the selection vector, respectively. The actual observed data likelihood on the otherhand is

$$f(d_s|\mathbf{x}_{s^0};\boldsymbol{\theta},\boldsymbol{\phi}) = \int f(d_{s^0}|\mathbf{x}_{s^0};\boldsymbol{\theta},\boldsymbol{\phi})d\mathbf{y}_{\overline{s}} \tag{2.20}$$

where $d_s = (\mathbf{y}_s, \mathbf{x}_{s^0})$. An $m-$estimation approach would ignore the selection process, and work with the likelihood

$$f(\mathbf{y}_s|\mathbf{x}_{s^0};\boldsymbol{\theta}) = \int f(\mathbf{y}_{s^0}|\mathbf{x}_{s^0};\boldsymbol{\theta})d\mathbf{y}_{\overline{s}}. \tag{2.21}$$

Rubin (1976) establishes the conditions under which inference from (2.20) and (2.21) will be identical from a Bayesian, Likelihood as well as a frequentist theory viewpoints. Principally, if

$$p(s|\mathbf{y}_{s^0},\mathbf{x}_{s^0};\boldsymbol{\phi}) = p(s|\mathbf{x}_{s^0};\boldsymbol{\phi}) \tag{2.22}$$

that is a propensity score type assumption. If this holds

$$\begin{aligned} f(d_s|\mathbf{x}_{s^0};\boldsymbol{\theta},\boldsymbol{\phi}) &= p(s|\mathbf{x}_{s^0};\boldsymbol{\phi}) \int f(\mathbf{y}_{s^0}|\mathbf{x}_{s^0};\boldsymbol{\theta})d\mathbf{y}_{\overline{s}} \\ &= p(s|\mathbf{x}_{s^0};\boldsymbol{\phi}) f(\mathbf{y}_s|\mathbf{x}_{s^0};\boldsymbol{\theta}) \end{aligned} \tag{2.23}$$

and **S** and **Y** satisfy a conditional independence type of factorization, as in Dawid (1979). From a frequentist viewpoint (2.20) and (2.21) will be identical only if s is fixed, since otherwise the labels in the integral (2.21) are not well defined, thus variables are conditionally independent in a probabilistic sense only if (2.23) holds for all possible s, Rubin (1976). For notational simplicity I will, however, avoid conditioning explicitly on s in the remaining of the section.

Provided that the parameters $\boldsymbol{\phi}$ are not functions of the parameters $\boldsymbol{\theta}$, that is the parameters are distinct, the distributions generated from (2.21) for given s will be identical to those generated by (2.23) , and so selection can be ignored. Bayesian inference requires a somewhat weaker condition on the selection mechanism than (2.22), that selection is MAR $p(s|\mathbf{y}_{s^0},\mathbf{x}_{s^0};\boldsymbol{\phi}) = p(s|\mathbf{y}_s,\mathbf{x}_{s^0};\boldsymbol{\phi})$ that is selection evaluated at the observed value of $\mathbf{y}_s, \mathbf{x}_{s^0}, s$ and $\boldsymbol{\phi}$ must be free only of $\mathbf{y}_{\overline{s}}$, while in addition parameters $(\boldsymbol{\theta},\boldsymbol{\phi})$ are *a priori* independent given $\mathbf{x}_{s^0}$.

How relevant, however, is this description for m-estimation over a Web panel survey data? In the setting of data collected through a known (random or purposive) sampling design (the backdrop of Rubin 1976 and subsequent work) the MAR assumption is reasonable and explains how and when data analysts may choose to ignore the selection mechanism. In the Web panel setting this emphasis seems misguided.

Thus here, in $m-$estimation, we make no modelling assumption on $p(s)$ and focus on the $m-$model. A description closer to our problem is this- Our survey process collects both $(\mathbf{x}_s, \mathbf{y}_s)$ from a population subset of Web panel volunteers. Let $\mathbf{W}_{s^0}$ be an indicator variable, a vector of 0's and 1's, which explains joining the panel set. Selection is then of the form $p(s|\mathbf{w};\boldsymbol{\phi})$. Ignorability then holds for Superpopulation inference about $\boldsymbol{\theta}$ based on the conditional distribution of $f(\mathbf{y}_s|\mathbf{w};\boldsymbol{\theta})$ for a given s. But viewed as an indicator variable, $\mathbf{W}$ will limit inference to the panel population subset while not necessarily to the wider population of interest. For this $f(\mathbf{y}_{\overline{s}}|\mathbf{w};\boldsymbol{\theta})$ must be known, and the question is then how to model this unobserved distribution.

A solution is to use $\mathbf{x}_s$ and examine the conditional distribution. If

$$f(\mathbf{y}_s|\mathbf{w},\mathbf{x}_s;\boldsymbol{\theta}) = f(\mathbf{y}_s|\mathbf{x}_s;\boldsymbol{\theta}), \tag{2.24}$$

holds, Superpopulation inference about $\boldsymbol{\theta}$ can be made directly from (2.24). That is panel volunteering, represented here by $\mathbf{W}$, contains no information beyond that in the auxiliary values $\mathbf{x}_s$ in explaining the distribution of $\mathbf{y}_s$. For finite population questions such as inference on $\overline{\mathbf{y}}_{s^0}$ involving $\mathbf{y}_{\overline{s}}$ requires knowledge of $\mathbf{x}_{\overline{s}}$ as well. Thus, in $m-$estimation, the focus is on correct model specification rather than considering MAR selection.

In the following subsections I build a robust m-estimation stratgey relying on balanced sampling suitable for Web panel question by the following steps. I introduce first the general idea of $m-$estimation using Valliant et al. (2000) general linear model prediction framework and comment as well on the likelihood and Bayesian approaches. I then briefly discuss Robins and Rotnitzky (1994) semi parametric theory for estimating equation type estimation which I referred to in the previous section, its relevance to the use of independent reference surveys in the panel question and its strength in finding estimators of the estimation variance. I then show that the common problem of model misspecificaion and its associated bias is inflated by the sample selection problem and propose as a layer of defense to this problem a purposive random sampling strategy- balanced sampling- as a method of enhancing the robustness to model misspecification of $m-$estimators while taking advantage of the ability of the panel administrators to control the within panel sampling process.

2.8.1 The Best Linear Unbiased Predictor

Suppose population vector $\mathbf{Y}_{s^0}$ has been generated by

$$m(\boldsymbol{\beta}) : \begin{cases} E(\mathbf{Y}_{s^0}) & = \mathbf{m}_{s^0}(\mathbf{x}_{s^0};\boldsymbol{\beta}) \\ V(\mathbf{Y}_{s^0}) & = \mathbf{V}_{s^0} \end{cases} \tag{2.25}$$

where $\mathbf{m}_{s^0}(\mathbf{x}_{s^0};\boldsymbol{\beta}) = (m(\mathbf{x}_1;\boldsymbol{\beta}),...,m(\mathbf{x}_N;\boldsymbol{\beta}))$ with $m(\cdot)$ indicating a function - linear or non linear - of the unknown parameters $\boldsymbol{\beta}$ and the observed auxiliaries. The $N \times N$ covariance matrix for $\mathbf{Y}$ is $\mathbf{V}_{s^0}$. The population units can be rearranged so that model components are expressed by

$$\begin{aligned}
\mathbf{m}_{s^0}(\mathbf{x}_{s^0};\boldsymbol{\beta}) &= \left[\mathbf{m}_s(\mathbf{x}_s;\boldsymbol{\beta})', \mathbf{m}_{\bar{s}}(\mathbf{x}_{\bar{s}};\boldsymbol{\beta})'\right] \\
\mathbf{V}_{s^0} &= \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{\bar{s}s} \\ \mathbf{V}_{s\bar{s}} & \mathbf{V}_{\bar{s}\bar{s}} \end{bmatrix}
\end{aligned}$$

representing the selected and non selected units of the population and their covariance structure.

Under model (2.25) with $\boldsymbol{\beta}$ assumed known, among the class of unbiased estimators linear in $\mathbf{Y}_s$, that is of the form $\hat{\overline{\mathbf{Y}}}_{s^0} = \mathbf{g}_s'\mathbf{Y}_s$, the error variance $E\{\hat{\overline{\mathbf{Y}}}_{s^0} - \overline{\mathbf{Y}}_{s^0}\}^2$ is minimized[6] by

$$\hat{\overline{\mathbf{Y}}}_{s^0} = N^{-1}\{\mathbf{1}_s'\mathbf{Y}_s + \mathbf{1}_{\bar{s}}'[\mathbf{m}_{\bar{s}}(\mathbf{x}_{\bar{s}};\boldsymbol{\beta}) + \mathbf{V}_{s\bar{s}}\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s - \mathbf{m}_s(\mathbf{x}_s;\boldsymbol{\beta}))]\} \quad (2.26)$$

a sum of sample units used directly and predicted values adjusted based on the model residuals. This is the best linear unbiased predictor (BLUP) estimator given by Valliant et al. (2000, theorem 11.2.1).

In practice an estimator can be suggested by computing separately the unknown $\boldsymbol{\beta}$, for example by minimizing the generalized least squares $[\mathbf{Y}_s - \mathbf{m}_s(\mathbf{x}_s;\boldsymbol{\beta})]'\mathbf{V}_{ss}^{-1}[\mathbf{Y}_s - \mathbf{m}_s(\mathbf{x}_s;\boldsymbol{\beta})]$ and solving iteratively by the Newton-raphson method, plugging the estimator into (2.26). A closed form approximation of the error variance $V(\hat{\overline{\mathbf{Y}}}_{s^0} - \overline{\mathbf{Y}}_{s^0})$ can be found by first using (2.26) to write down the variance in general and applying standard Taylor series approximations using first partial derivatives to compute variance and covariance components as necessary. Valliant (1985) covers some of the details. For the linear regression case the variance can be derived directly.

Models of particular interest which fit into this framework are Generalized Linear Models (GLM) which assumes a distribution in the exponential family underlying (2.25) and an appropriate link function. For example suppose $\mathbf{Y}_{s^0}$ are independent Bernoulli random variables

$$\begin{aligned}
E(Y_k) &= m(\mathbf{x}_k,\boldsymbol{\beta}) \qquad \text{where } m(\mathbf{x}_k,\boldsymbol{\beta}) \in \{0,1\} \\
V(Y_k) &= m(\mathbf{x}_k,\boldsymbol{\beta})[1 - m(\mathbf{x}_k,\boldsymbol{\beta})] \;;\forall k \in s^0 \qquad (2.27)
\end{aligned}$$

[6] Whereas the linear estimator minimizing the estimator variance $V(\hat{\overline{\mathbf{Y}}}_{s^0}) = \mathbf{g}_s'V(\mathbf{Y}_{ss})\mathbf{g}_s$ is simply $\hat{\overline{\mathbf{Y}}}_{s^0} = N^{-1}\mathbf{1}_{s^0}'\mathbf{m}_{s^0}(\mathbf{x}_{s^0};\boldsymbol{\beta})$, that is the value for each unit in the population is estimated as its expected value from the regression model.

assuming $m(\mathbf{x}_k, \boldsymbol{\beta}) = [1 + \exp(-\mathbf{x}_k^{'}\boldsymbol{\beta})]^{-1}$ such that $\text{logit}[m(\mathbf{x}_k, \boldsymbol{\beta})] = \mathbf{x}_k^{'}\boldsymbol{\beta}$. The estimator (2.26) under (2.27) is simply

$$\hat{\overline{\mathbf{Y}}}_{s^0} = N^{-1}[\mathbf{1}_s^{'}\mathbf{Y}_s + \mathbf{1}_{\bar{s}}^{'}\mathbf{m}_{\bar{s}}(\mathbf{x}_{\bar{s}}; \hat{\boldsymbol{\beta}})] \tag{2.28}$$

where the residual-based adjustment of (2.26) is dropped as $\mathbf{V}_{s\bar{s}} = 0$ under the assumption of independence between units. This is an identical format to the regression estimator popular in survey sampling theory. The The MLE of the model parameters can be found by solving the likelihood estimating equations using the Fisher scoring algorithm. The approximate variance is the sum of the sample variance $\mathbf{V}_{ss}$ and a function of $\mathbf{V}_{\bar{s}\bar{s}}$ and the first partial derivative $\partial\mathbf{m}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ over the sampled and non sampled units. An estimator is computed by plugging $\hat{\mathbf{V}}_{ss} = diag[\mathbf{m}_s(\mathbf{x}_s; \hat{\boldsymbol{\beta}})(1 - \mathbf{m}_s(\mathbf{x}_s; \hat{\boldsymbol{\beta}}))]$ and and similarly for $\mathbf{V}_{\bar{s}\bar{s}}$ and replacing $\hat{\boldsymbol{\beta}}$ into the first partial derivatives. Under certain conditions on the finite population and sampling set it is consistent of the approximate variance while the estimator is asymptotically normal. For the specific formulation and conditions, see Valliant et al. (2000, sec 11.2.1), and Valliant (1985).

Two other estimation strategies are likelihood and Bayesian approaches. Take for example the simple case where (y_k, x_k) $k = 1, .., N$ are assumed to be independent observations from a bivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_y, \mu_x)$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_x^2, \sigma_y^2, \sigma_{xy})$. Under correct specification selection is ignored and by simple factorization the joint distribution is

$$f(y_k, x_k|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = f(x_k|\mu_x, \sigma_x^2)f(y_k|x_k; \beta_0, \beta_1, \sigma_{y|x}^2) \tag{2.29}$$

where β_0, β_1 and $\sigma_{y|x}^2$ are the regression coefficients and its model residual variance. The parameter $(\mu_x, \sigma_x^2, \beta_0, \beta_1, \sigma_{y|x}^2)$ are one to one function of the original parameter $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. MLE can be found then by independently maximizing the two components of (2.29) leading to the common regression estimator such as (2.28) under independent unit distributions where $m(\mathbf{x}_k, \boldsymbol{\beta}) = \mathbf{x}_k^{'}\boldsymbol{\beta}$.

The more popular $m-$approach in survey sampling is Bayesian (Ericson 1969, 1988; Binder 1982; Rubin 1983, 1987; Ghosh and Meeden 1997, Little 2004, Little Rubin, Gelman et al.) which would determine the posterior distribution of

$$\overline{\mathbf{Y}}_{s^0} = f\overline{\mathbf{Y}}_s + (1 - f)\overline{\mathbf{Y}}_{\bar{s}} \;\; ; f = \frac{n}{N}$$

by using simulations of $\mathbf{Y}_{\bar{s}}$. Bayesian inference separates the analysis into two steps: (1) Superpopulation inference, that is, drawing $\boldsymbol{\theta}^l$ $l = 1, ..., L$ from the posterior distribution $f(\boldsymbol{\theta}|\mathbf{y}_s, \mathbf{x}_{s^0})$ and then (2) finite population inference by drawing a vector of missing values $\mathbf{Y}_{\bar{s}}$ from $f(\mathbf{y}_{\bar{s}}|\mathbf{x}_{s^0}, \mathbf{y}_s; \boldsymbol{\theta}^l)$ the posterior predictive distribution. As $\mathbf{Y}_s$ is observed, draws from the posterior predictive distribution of $\overline{\mathbf{Y}}_{\bar{s}}$ is equivalent to draws from the posterior predictive distribution from $\overline{\mathbf{Y}}_{s^0}$. The simulation of $\mathbf{Y}_{\bar{s}}$ from its posterior distribution are called *multiple imputations* (Gelman et al., 2002).

The fact is, however, that but for the linear regression case, there has been only a limited number of studies applying estimators such as 2.26 in the survey field. This can be explained by the need for explicit summation of the predictions for non sampled population units which is rarely available for general population cases. A notable exception is the case of small area estimation (SAE) where aggregated population level data is used to model averages of measurements of interest over units confined to a certain geographical region, mainly in the context of official national statistics. There is a large body of literature in the field (see Ghosh and Rao, 1994; Rao, 2003; Datta, 2009; Lehtonen and Veiganen, 2009; Chandra and Chambers, 2011 or Pfeffermann, 2013) which has expanded on predictors such as (2.26), however, with little relevance to our Web panel problem in terms of objectives or data structure.

A way of dealing with this constraint is by incorporating in the estimation strategy data collected from a reference survey allowing for unit level summation. A useful framework facilitating such integration is to consider finite population estimation from the perspective of estimation equation functions which I discuss in the next few paragraphs.

2.8.2 Estimating Equations in Survey Sampling

The use of estimation equations (EE) for analytic estimands such as regression coefficients over random survey samples is not new, see for example Binder (1983), Pfeffermann (1993), Kovacevic and Binder (1997). Regression techniques for complex surveys, as implemented in software packages like SUDAAN (Shah et al. 1997), are based on weighted EE functions.

To fix the idea, suppose $\{d_k,\ k = 1...N\}$ are independent observations which depend on an unknown single parameter θ. The estimator of θ is defined as the solution to

$$\sum_N u(d_k; \hat{\theta}) = 0$$

where $u(d_k; \theta)$ is the estimation function, a known function that does not depend on k or N. Godambe & Thompson (1986) show that the optimal estimation function $h(d_k; \theta)$ applied to a random sample data, unbiased of $u(d_k; \theta)$ with regards to the sample selection process, is simply a $\pi-$estimator of the 'census' estimation equations

$$\sum_s h(d_k; \theta) = \sum_s u(d_k; \theta)\pi_k^{-1}$$

where π_k , $k \in s^0$ are constants known by design. The authors restrict to the case where population level estimation function $u(d_k;\theta)$ is linear with regards to d_k, and optimality refers to minimizing the variance of the estimation function divided by the expected partial derivative squared. See Pfeffermann (1993, sec 7.2.4) for a short review. When fully specified joint distribution of the population values is assumed this result provides justification for the use of pseudo maximum likelihood estimation (PMLE) which utilizes sampling design weights to estimate the likelihood equations that would have been obtained in the case of a census. A general discussion of PMLE is entailed in Skinner (chapter 3, 1989), Binder (1983) and Chambless & Boyle (1985)).

More recently, Robins and Rotnitzky (together with colleagues) developed an elegant semiparametric theory underlying EE-type estimators for more general missing data problems attracting great interest in the causal inference domains (Gustafson 2012). Interestingly and despite the applicability, there is little evidence of a crossover to finite-population sample surveys (Lumley et al., 2011). Tsiatis (2006) gives an exhaustive introduction to this work, and in the following I give a short description of some basic ideas within our m-estimation perspective.

First, for a parameter θ in a parametric or semiparametric statistical model, the influence function $\varphi(d_k, \theta)$ of estimator $\hat{\theta}$ based on independent, identically distributed d_k, $k = 1, ..., N$ can be defined as satisfying

$$\hat{\theta} - \theta_0 = N^{-1}\Sigma_{k=1}^{N}\varphi(d_k, \theta_0) + o_p(N^{-1/2}) \tag{2.30}$$

with $E[\varphi(d_k, \theta)] = 0$ and $E[\varphi(d_k, \theta)^2] < \infty$ where expectation is with respect to the true distribution of d and θ_0 the true value of θ generating the data. An estimator satisfying (2.30) is said to be *asymptotically linear* and is consistent and asymptotically normal with asymptotic variance $E[\varphi(d, \theta)^2]$. An estimator function $u(d_k; \theta)$ is linked to the influence function by

$$u(d_k; \theta) = \varphi(d_k, \theta_0) - (\theta - \theta_0).$$

Tsiatis and Davidian demonstrate (2007) a wide variety of estimators for the population mean can be expressed as the solution to an estimating equation based on an influence function, and show that exploiting the relationship between influence functions and estimators is a fruitful approach to studying and contrasting the (large-sample) properties of estimators

For our assumed data structure, we do not observe $(\mathbf{Y}_{s^0}, \mathbf{X}_{s^0})$ nor a simple sample of it but rather $d_s = (S, \mathbf{X}_{s^0}, \mathbf{Y}_s)$. Under correct specification of the m-model, Y and $\mathbf{X}$ are conditionally independent and the observed distribution is

$$p(d_k) = p(s_k|\mathbf{x}_k)f(y_k|\mathbf{x}_k)^{I(s_k=1)}f(\mathbf{x}_k). \tag{2.31}$$

Let $p_0(d)$ be the density in the class of densities of form (2.31) generating the observed data (the true joint density). Assume only that $E(Y_k|\mathbf{X}_k) = m(\mathbf{x}_k; \boldsymbol{\beta})$ leaving the rest of the components of $p(d_k)$ unspecified and denote by $m_0(\mathbf{x}_k)$ the true function $E(Y_k|\mathbf{x}_k)$. Tsiatis (2006, Section 4.5) shows that all estimators of $E(Y) = \mu$ have influence functions of the form

$$m_0(\mathbf{x}_k) - \mu + S_k a(\mathbf{x}_k)[Y_k - m_0(\mathbf{x}_k)] \tag{2.32}$$

for arbitrary function $a(\cdot)$ of $\mathbf{x}$, which suggests that any estimator of the population average consistent and asymptotically normal under m_0 must be asymptotically equivalent to an estimator of the form

$$N^{-1}\sum_{k=1}^{N}[m(\mathbf{x}_k; \boldsymbol{\beta}) + S_k a(\mathbf{x}_k)\{Y_k - m(\mathbf{x}_k; \boldsymbol{\beta})\}] \tag{2.33}$$

for example $a = 1$ will give the regression estimator discussed earlier, while the estimator when $a = 0$ is equivalent to the BLUP estimator (2.28) when $\mathbf{V}_{s^0}\mathbf{1_N} \in M(\mathbf{X}_{s^0})$ where $M(\mathbf{X}_{s^0})$ denotes the vector space spanned by all linear combinations of the columns of $\mathbf{X}$- the linear manifold. Tsiatis and Davidian (2006, eq. 7) give the correct formulation of the influence function underlying estimator (2.33) when parameters are estimated[7].

The use of influence functions allow as well a simple method for calculating and estimating the standard error of estimators in the finite population case. In appendix 5.1 I discuss variance estimation and demonstrate for the important univariate case where $Y_k = \beta x_k + \varepsilon_k,\ \ \varepsilon_k \sim (0, \sigma_k^2)$ and $a(\mathbf{x}_k) = 1$ which would lead to the ratio estimator. The resulting estimator based on the influence is equivalent to that proposed by (Valliant et al., 2000, p.145) for the ratio estimator.

The use of estimating equation based on an influence function is not limited to only m-type estimations and as I've discussed already in the previous section covers the cases where assumptions are made on the distribution of the selection process alone or, as shall be seen in the following section, alongside the outcome model of interest leading to the important class of double robust estimators.

[7] Alternatively, from an estimation function perspective, for example for $a = 1$ assuming normal distribution we may suggest solving

$$\begin{pmatrix} \sum_{s^0}[S_k(Y_k - \overline{\mathbf{y}}_{s^0})/\sigma^2 + (1 - S_k)(m(\mathbf{x}_k; \boldsymbol{\beta}) - \overline{\mathbf{y}}_{s^0})/\sigma^2] \\ \sum_{s^0} S_k \partial m(\mathbf{x}_k; \boldsymbol{\beta})/\partial\beta'\ [Y_k - m(\mathbf{x}_k; \boldsymbol{\beta})] \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}$$

where $\mathbf{0}$ is a vector of all zeros of an appropriate dimension.

2.8.3 $m-$Model Misspecification Over non-Random Samples

Thus far in this discussion little emphasis has been on what is usually the fundamental question in survey sampling, namely 'how have the observations been sampled?' More precisely we are anchored on the assumption of correct specification of the outcome of interest model while disregarding any thought on the sampling of panelists. Given the self selection of the panel it is convenient to dismiss the within panel sampling mechanism as irrelevant. However, what if the m-model is incorrect? Is there a link between the panel unrepresentativeness to successful model fitting?

In the following I start by arguing that the common problem of m-model misspecification is more likely in a self selected sample such as the Web panel set, even when the conditional independence assumption is attainable given the available covariates. However, by taking advantage of the panel managment's control of the within panel sampling process a proper estimation strategy, of selection and estimator, becomes *robust* in the sense that it is robust to m-model misspecification.

First, I introduce the following tool- For any selection mechanism

$$f(\mathbf{y}|s,\mathbf{x}) = f(\mathbf{y}|\mathbf{x}) + [f(\mathbf{y}|s,\mathbf{x}) - f(\mathbf{y}|\overline{s},\mathbf{x})][1-\pi(\mathbf{x})]$$

with $\overline{s}$ denoting the non selected set of units. Consider, the basic case where we we assume M: $Y_k = \beta + \varepsilon_k$, which would underline estimating by the simple mean, the potential bias can then be given by

$$B(\hat{\overline{\mathbf{Y}}}_{s^0}|s) = \int \mathbf{y}[f(\mathbf{y}|s) - f(\mathbf{y}|\overline{s})]d\mathbf{y}[1-\pi]$$

where π denotes the unconditional probability of being a panel member. This is an m-estimation version of the coverage error discussed in the context of the fixed population framework where the size of the bias depends on (i) the difference in expected value of Y between panel and non panel members and (ii) the probability of (not) participating in the panel survey. Returning to (2.34) note that the bias is largest for values of $\mathbf{x}$ where the participation rate $\pi(\mathbf{x})$ is low and the gap in the distributional form of $f(y|\cdot)$ is wide[8].

Now, to assist derivation I introduce a new formulation of the BLUP estimator

[8] Clearly a researcher considering running a Web access panel survey should ask whether important segments of the population vis a vis the research topic overlap with the non participating segments of Web panels, and whether these segments should behave fundamentally different in regards to these measurements from the rest of the population studied.

(2.26) assuming independent errors and an identity link function

$$\hat{\overline{\mathbf{Y}}}_{s^0} = \sum_{k=1}^{n} g_k Y_k \quad \text{where } g_k = 1 + \mathbf{1}'_{\overline{s}}\mathbf{X}_{\overline{s}}\mathbf{A}_s^{-1}\mathbf{X}_k w_k$$

where $\mathbf{A}_s = \mathbf{X}'_s\mathbf{W}\mathbf{X}_s$. Together, under a working model M we can write

$$E(\hat{\overline{\mathbf{Y}}}_{s^0}|s) = \sum_{k=1}^{n} E(g_k Y_k) + \sum_{k=1}^{n} \int g_k \int y_k[f(y|\mathbf{x}, s) - f(y|\mathbf{x}, \overline{s})][1 - \pi(\mathbf{x})]f(\mathbf{x})dyd\mathbf{x}. \tag{2.34}$$

More generally assume that the true underlying model of the population follows a BLUP type model $\tilde{M}$ also with independent errors while our working model is as before denoted by M. If we let $\tilde{g}$ be the prediction coefficients under the correct linear model

$$B(\hat{\overline{\mathbf{Y}}}_{s^0}|s) = \sum_{k=1}^{n} E[(g_k - \tilde{g}_k)Y_k] + \sum_{k=1}^{n} \int g_k \int y_k[f(y|\mathbf{x}, s) - f(y|\mathbf{x}, \overline{s})][1 - \pi(\mathbf{x})]f(\mathbf{x})dyd\mathbf{x}$$

a combination of model misspecification error and the effect of a self selected sample set.

When the available conditioning covariate set $\mathbf{X}$ is not sufficient to attain conditional independence, that is $f(y|\mathbf{x}, s) \neq f(y|\mathbf{x}, \overline{s})$, then the model is not correctly specified by definition and the bias is the sum of the two components. If, on the other hand the model is correctly specified then $g_k = \tilde{g}_k$ and by definition $f(y|\mathbf{x}, s) = f(y|\mathbf{x}, \overline{s})$ so the expected bias is zero.

Our interest is in the third possibility where the covariate set is sufficient for $f(y|\mathbf{x}, s) = f(y|\mathbf{x}, \overline{s})$ but our specific model specification is incorrect so that the first component is not zero. The problem is that in practice for regions of $\mathbf{X}$ where $\pi(\mathbf{x}) \approx 0$ the likelihood of model misspecification increases above and beyond the normal difficulties any modeller faces in other areas of statistics.

A useful illustration of the problem is given in figure 2.2 which describes the case where the panel selection process is such that

$$p(s|\mathbf{x}_1, \mathbf{y}) = p(s|\mathbf{x}_1)$$

and that we assume

$$\begin{aligned} M: \quad Y_k = &\quad \beta + \beta x_{1k} + \varepsilon_k \quad \text{while in fact} \\ \tilde{M}: \quad Y_k = &\quad \beta + \beta x_{1k} + \beta_2 x_{2k} + \varepsilon_k \end{aligned}$$

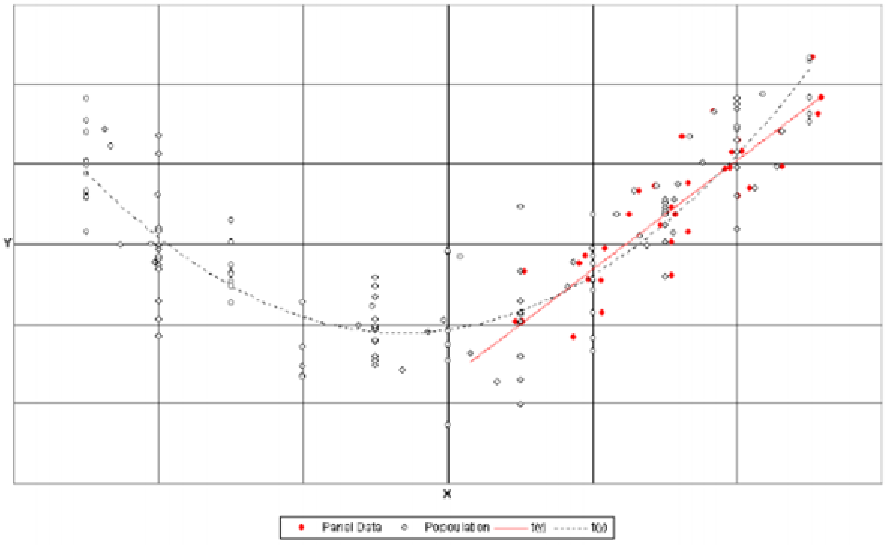


Figure 2.2: Extrapolation of the BLUP estimator

where $x_2 = x_1^2$. Modelling is on the truncated data $(\mathbf{y}_s, \mathbf{x}_s)$ and prediction, as can be seen in the graph, may rely on model extrapolation at areas (values) different from the majority of the $\mathbf{x}_s$, i.e. where $\pi(\mathbf{x}) \approx 0$.

What makes matters worse, is that this extrapolation from the panel data to the non panel population is not evident in the standard errors of BLUP estimators (or any imputation, multiple imputation, ML or mean score methods). The usual standard goodness of fit techniques called upon to assess the adequacy of the working model have limited relevance for detecting this misspecification. Note how well the prediction model fits over the panel records as shown in figure 1.2, compared to their complete inadequacy over the entire population.

As Tan (2007) argues model checking *in this region* of $\mathbf{x}$ is not capable of detecting m- misspecification, while leverage points may indicate the existence of such region but not model misspecification. This problem holds for low or high-dimensional x, and is separate from the difficulty to capture $m(x)$ over panel members when x is high-dimensional.

A final comment is to say that this problem does not exist in the $\pi-$estimation approach as we assume that we observe $(\mathbf{S}, \mathbf{X}_{s^0})$ over the entire population.

2.8.4 Robust $m-$estimation Under Balanced Random Sampling

One way of increasing the quality of m-estimators is to couple estimators with sampling designs which may provide an additional layer of robustness to model misspecification. To understand the idea, let us start with a simple example.

Consider an m-estimator which assumes a single covariate working model m : $E(Y_k|x) = x_k\beta$ and $V(Y_k|x) = x_k\sigma^2$ with independent errors. Under m the $m-$estimators is $\hat{\overline{\mathbf{Y}}}_{s^0} = \overline{\mathbf{X}}_{s^0}\overline{\mathbf{Y}}_s/\overline{\mathbf{X}}_s$ the simple ratio estimator. If however, under the true model of the population $E(Y_k|x) = \alpha + x_k\beta$ then the bias given s of the $m-$estimator defined as the expectation over the true population model is

$$B(\hat{\overline{\mathbf{Y}}}_{s^0}|s) = \alpha(\overline{\mathbf{X}}_{s^0} - \overline{\mathbf{X}}_s)/\overline{\mathbf{X}}_s$$

and so when our sample is representative in the sense that $\overline{\mathbf{X}}_s = \overline{\mathbf{X}}_{s^0}$ the bias is removed. This example suggests that in the estimation of finite population quantities such as totals and averages, we can protect against model misspecification by ensuring that the sample is representative of the population.

We can generalize this idea by first exposing an interesting link between the optimal linear m-type estimation and the simple sample average under a representative sample. Consider as before $m-$estimators under a general linear model

$$m(\mathbf{x}) : \begin{cases} E(Y_k) & = \mathbf{X}_k'\boldsymbol{\beta} \\ V(Y_k) & = \sigma_k^2 \end{cases} \tag{2.35}$$

where $\mathbf{x}_k$ and $\boldsymbol{\beta}$ are p-vector and assuming $\mathbf{x}_{s^0}$ is observed. If for all samples s over the sampling space there exists a constant column vector $\boldsymbol{\lambda}$ of dimension p not depending on k such that for all $k \in s^0$

$$\begin{cases} \sigma_k^2 & = \boldsymbol{\lambda}'\mathbf{X}_k \\ \overline{\mathbf{X}}_{js} & = \overline{\mathbf{X}}_{js^0} \text{ for all } j = 1, ..., p \end{cases} \tag{2.36}$$

then $\hat{\overline{\mathbf{Y}}}_{s^0} = \overline{\mathbf{Y}}_s$.

The proof is simple. Start by noting that when $\sigma_k^2 = \boldsymbol{\lambda}'\mathbf{X}_k$ then

$$\overline{\mathbf{X}}_s = n^{-1}\sum_s \frac{\boldsymbol{\lambda}'\mathbf{X}_k\mathbf{X}_k'}{\sigma_k^2}$$

and assuming balance, that is, $\overline{\mathbf{X}}_s = \overline{\mathbf{X}}_{s^0}$ then under (2.35)

$$\begin{aligned}
\hat{\overline{\mathbf{Y}}}_{s^0} &= \overline{\mathbf{X}}_{s^0}(\sum_s \frac{\mathbf{X}_k\mathbf{X}_k^{'}}{\sigma_k^2})^{-1}\sum_s \frac{\mathbf{X}_kY_k}{\sigma_k^2} \\
&= n^{-1}\sum_s \frac{\boldsymbol{\lambda}^{'}\mathbf{X}_k\mathbf{X}_k^{'}}{\sigma_k^2}(\sum_s \frac{\mathbf{X}_k\mathbf{X}_k^{'}}{\sigma_k^2})^{-1}\sum_s \frac{\mathbf{X}_kY_k}{\sigma_k^2} \\
&= n^{-1}\sum_s \frac{\boldsymbol{\lambda}^{'}\mathbf{X}_kY_k}{\sigma_k^2} = \overline{\mathbf{Y}}_s. \qquad (2.37)
\end{aligned}$$

In other words, the m-estimator under (2.35) reduces to the survey sample average as long as (i) the sample s is balanced in the sense it satisfies the condition $\overline{\mathbf{x}}_{js} = \overline{\mathbf{x}}_{js^0}$ for all $j = 1, ..., p$, and (ii) that the correct model $m(\mathbf{x})$ variance structure can be described as a linear combination of the set (or subset) of the regression covariates $\mathbf{x}$.

This link between m-model and sample properties implies a *bias-robust strategy* where the estimator is unbiased for all members of a broad class of models. As long as the sample is balanced on all relevant covariates, within the wide class of models, the analyst may misspecified the $m-$model but still produce unbiased estimation.

To clarify this idea further it is instructive to go back to Herson and Royall (1973) who first proposed the idea of *model bias-robust strategies* under a general polynomial

$$Y_k = \sum_{j=0}^{J} \delta_j\beta_jX_k^j + \varepsilon_k v_k^{1/2}$$

where the errors are $\varepsilon_k \sim (0, \sigma^2)$ and uncorrelated, $\{\beta_j\}_{j=0}^J$ are a set of unknown parameters, and $\{\delta_j\}_{j=0}^J$ are $0-1$ variables indicating whether the jth power term is in the model or not. If this model holds, then the bias of the survey sample average is

$$E(\overline{Y}_s - \overline{Y}_{s^0}) = \sum_{j=0}^{J} \delta_j\beta_j[\overline{\mathbf{X}}_s^j - \overline{\mathbf{X}}_{s^0}^j]$$

which for any $\{\delta_j\}_{j=0}^J$ configuration will be *zero* as long as the sample s is balanced on the same configuration of moments $\overline{\mathbf{X}}_s^{(j)}$.

Their discussion was within the classic survey sampling context and as such advocated for random sampling designs to achieve (better) balance, indeed when considering the likely case where misspecification most likely occurs from omitted variables the importance of random sampling is apparent. For example when

$m(\mathbf{x})$:$E(Y_k) = \beta_0 x_k + \beta_1 h(z_k)$ where $h(z_k)$ is a function of an unknown characteristic then the bias of, say $\hat{\overline{\mathbf{Y}}} = \overline{\mathbf{X}}_{s^0}\overline{\mathbf{Y}}_s/\overline{\mathbf{X}}_s$ the ratio estimator

$$E(\hat{\overline{\mathbf{Y}}}_{s^0} - \overline{\mathbf{Y}}_{s^0}) = \beta_1(\frac{\overline{\mathbf{X}}_s}{\overline{\mathbf{X}}_{s^0}}\overline{h(\mathbf{z})}_s - \overline{h(\mathbf{z})}_{s^0})$$

which implies that for the bias to be zero we require that $\overline{\mathbf{x}}_s/\overline{\mathbf{x}}_{s^0} = \overline{h(\mathbf{z})}_s/\overline{h(\mathbf{z})}_{s^0}$.

In our context of Web panel survey sampling this bias robust strategy can be viewed as a justification to the common practice in most panel companies to sample purposively on basic population characteristics. More specifically, when a linear $m-$estimation approach is taken a bias robust strategy would be to sample from the web panel purposively so that average balance on the covariates is met.

Lastly, the question naturally arises: does it matter which model and so which estimator we calculate within the assumptions of (2.35) and (2.36) given we intend the sample to be balanced? The answer is yes for two reasons. (i) In practice, it is difficult to achieve balance exactly and the estimators ' varying levels of sensitivity to departure in balance will be closely linked to the departure from the true underlying model; (ii) closed form variance estimators will not be the same since the residuals under the different models are different, even at balance. For further discussion on these two ideas see Valliant et al (2000 sec 3.2.4, 3.2.5 and 5.5.1).

2.9 Modelling Both Selection and Outcome- $\pi m-$Estimators

In the previous section we have throughout assumed that the population distribution follows $f(d_s) = f(\mathbf{y}_s|\mathbf{x})p(\mathbf{s}|\mathbf{x})f(\mathbf{x})$, a conditional independence between $p(s|\cdot)$ and $f(\mathbf{y}|\cdot)$, by alternatively making assumption on the conditional expectations $E(\mathbf{Y}|\mathbf{x})$ and $E(\mathbf{S}|\mathbf{x})$ while ignoring the other elements of the distribution. In this last section of the chapter I discuss estimators which rely on modelling separately both conditional exceptions, while still ignoring the population distribution of $\mathbf{X}$.

We denote a specific class of such estimators as $\pi m-$estimators and discuss the strength of the method compared to π or m only estimators. Specifically, the $\pi m-$estimators can be shown to still be consistent when either π or m models are correct, while being as efficient as an $m-$model when both models are correct. From a Web panel problem perspective, these two characteristics can be seen as addressing the two key weaknesses of the single model estimators: The enhanced fear of $m-$model misspecification given that modelling is done on a non representative subset of the population, and the inefficiency of $\pi-$estimators, even when the underlying model is correctly specified.

Before describing the idea directly, it is instructive to comment that we have already touched briefly on two estimators which rely on modelling of both elements of the population distribution. The first is the idea of replacing the set $\mathbf{X}$ with the unit selection probability $\pi(\mathbf{x}_k)$ in a regression (covariance) adjustment. Rosenbaum and Rubin (1983) suggested this in the context of causal inference, with the reasoning coming naturally from the basic property of the propensity score as a balancing score. That is

$$E(Y) = EE(Y|b(\mathbf{x})) = EE(Y|b(\mathbf{x}), S = 1)$$

for balancing score $b(\mathbf{x})$ and a single variable Y. It follows that if as well the conditional expectation of Y given $\pi(\mathbf{x})$ can be explained by, for example, by a GLM type model such as that underlying the BLUP estimator (2.25) then an $m-$estimator of the population average, where the explanatory variable is $\pi(\mathbf{x})$ will have the form

$$\hat{\bar{\mathbf{Y}}}_{s^0 m} = N^{-1}\sum_{s} Y_k + \sum_{\bar{s}} m[\hat{\pi}(\mathbf{x}_k)\hat{\beta}]$$

with m^{-1} a known link function and $\hat{\beta}$ is the MLE among the set of observed units s in the GLM model with the single covariate $\hat{\pi}(\mathbf{x})$ which itself is estimated beforehand. However, as an m- estimator this approach suffers the same weaknesses of $m-$model misspecification enhanced by the self selected nature of the set s. Furthermore, the fact we are modelling the study variable with selection probabilities rather than original covariates may cause difficulty in including subject matter expertise in the modelling exercise which is not uncommon in survey based research. Rosenbaum and Rubin showed that the point estimate obtained from an $m-$estimator modelled over multivariate $\mathbf{X}$ or scalar $\pi(\mathbf{x})$ lead to the same result, and so the main argument for the estimator operational and lies in the reduction of the complexity of the $m-$model and the obvious simplified fitting and diagnostics tests over a single scalar $\pi(\mathbf{x})$. For further discussion on the approach see D'Agostino Jr (1998) who reviews empirical and theoretical work on the topic and suggests as well a variation on the method which rather than replacing $\mathbf{X}$ with $\pi(\mathbf{x})$ in the $m-$model includes both. He argues for this approach again on operational grounds but I shall show below that such a model is a $\pi m-$estimator.

The second estimation strategy which involves assumption on both elements of the joint distribution $f(d_s)$ is the family of GREG estimators and the broader idea of Calibration estimation touched on briefly in section (1.4). As a survey sampling approach both methods rely only on the $p(s)$ distribution for statistical inference such as building confidence inference or computing statistical tests, however, for reasons of increasing efficiency Cassel et al. (1976); Särndal et al. (1992) proposed initially to include an assisting linear regression model as a vehicle for incorporating sample and population level covariate information.

We shall see the significance of GREG as a $\pi m-$estimator below, but before it is useful to note a common misunderstanding when comparing GREG and BLUP estimators- that the two differ only in the use or absence of unit selection probabilities and for equal probability selection processes the two converge (see for example Valliant et al., 2000, page 40-42). Consider the simple case where $p(s)$ is such that $\pi_k = n/N$, a common assumption in practice when $p(s)$ is unknown. Then from formulation (2.4) and (2.25) assuming an identity link function

$$\begin{aligned} \hat{\overline{\mathbf{Y}}}_{GREG} &= \overline{\mathbf{Y}}_s + (\overline{\mathbf{X}} - \overline{\mathbf{X}}_s)\hat{\boldsymbol{\beta}}^{'} \\ \hat{\overline{\mathbf{Y}}}_{BLUP} &= \overline{\mathbf{X}}^{'}\hat{\boldsymbol{\beta}} + (n/N)(\overline{\mathbf{Y}}_s - \overline{\mathbf{X}}_s^{'}\hat{\boldsymbol{\beta}}). \end{aligned} \tag{2.38}$$

as $\hat{\beta}_\pi = \hat{\beta}$ the weighed and ordinary least square sample estimator of the regression coefficients are identical. What is evident is that as an $m-$estimator the BLUP estimator relies almost completely on the predicted values $\hat{m}_k = \mathbf{x}^{'}\hat{\boldsymbol{\beta}}$ (as $n/N \approx 0$) , while the model assisted $\pi-$estimator GREG is influenced by both the $m-$model based correction and the direct sample average.

More broadly, I discuss now the class of Augmented inverse probability weighted (AIPW) estimators and its specific subclass of $\pi m-$estimators. To start note that the $\pi-$estimator $\hat{\overline{\mathbf{Y}}}_\pi = \sum_{s^0} S_k\hat{\pi}_k^{-1}Y_k / \sum_{s^0} S_k\hat{\pi}_k^{-1}$ can be described as the solution to a inverse probability weighted (IPW) estimating equation system

$$\sum_{s^0} S_k\hat{\pi}_k^{-1}U_k = 0$$

where $U_k = \left(Y_k - \overline{\mathbf{Y}}_{s^0}\right)/\sigma^2$ for known σ^2, and where $\hat{\pi}_k = \pi(\mathbf{x}_k; \hat{\boldsymbol{\alpha}})$ are computed beforehand by solving the vector of normal equations $\sum_{s^0}[S_k - \pi(\mathbf{x}_k; \boldsymbol{\alpha})]\mathbf{X}_k = 0$ for the logit regression coefficients of $\pi(\mathbf{x}_k; \boldsymbol{\alpha}) = \exp(\mathbf{x}^{'}\boldsymbol{\alpha})/[1 + \exp(\mathbf{x}^{'}\boldsymbol{\alpha})]$.

However, the IPW estimation equations, although consistent and asymptotically normal (CAN) with respect to $p(\cdot)$, are defined only over the sample set s and so do not utilize available population level covariate information $\mathbf{X}_{s^0}$. Robins and Rotnitzky (1995) address this inefficiency and note that under correct $\pi-$specification one never does worse by adding function $h(\cdot)$ of the available population covariates. That is estimate the population average by the AIPW equations

$$\sum_{s^0} \left\{S_k\hat{\pi}_k^{-1}U_k + (1 - S_k\hat{\pi}_k^{-1})h(\mathbf{x}_k\right\} = \mathbf{0}.$$

When π_k are consistently estimated $E\left\{(1 - S_k\hat{\pi}_k^{-1})h(\mathbf{x}_k)\right\} \approx 0$ and so the resulting AIPW estimator, as the IPW estimator, is CAN under correct specification of the selection process (Bang and Robins, 2005). In (2.8.2) I have already touched on this idea in the discussion of the influence function for $\pi-$type estimators.

Moving on Scharfstein et al. (1999) show that by introducing a model for the outcome variable $\mathbf{Y}$, specifically by choosing $h(\mathbf{x}) = E(U_k|\mathbf{x}_{s^0})$ where the expectation is with respect to the conditional distribution of $\mathbf{Y}$ given $\mathbf{x}$, the resulting estimator solving

$$\sum_{s^0} \left\{ S_k \hat{\pi}_k^{-1} U_k + (1 - S_k \hat{\pi}_k^{-1}) E(U_k|\mathbf{x}_k \right\} = \mathbf{0} \tag{2.39}$$

is CAN when either the m or π models hold, and will attain the minimum variance bound.

To see that GREG is a member of this πm-estimator class, assume first that $m(\mathbf{x};\boldsymbol{\beta}) = E(\mathbf{Y}_k|\mathbf{X} = \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ and $V(\mathbf{Y}) = diag\{\sigma_k^2\}$; Also assume that $Pr(S_k = 1|\mathbf{X} = \mathbf{x}) = E(S_k|\mathbf{x}) = \pi_k(\mathbf{x};\boldsymbol{\alpha})$ where $0 > Pr(S_k = 1|\mathbf{x}) \geq 1$ for all $\mathbf{x}_{s^0}$. Assume parameters $\boldsymbol{\alpha}$ are estimated consistently and that $\boldsymbol{\beta}$ is estimated by the normal equations over s weighted by $\hat{\pi}_k$ $k \in s$. It is then immediate that the GREG estimator of the population mean, given here by the following equivalent forms

$$\begin{aligned}
\hat{\overline{\mathbf{Y}}}_{\pi m} &= N^{-1}\{\sum_{s^0} S_k Y_k \hat{\pi}_k^{-1} + (1 - S_k \hat{\pi}_k^{-1}) \sum_{s^0} \mathbf{X}_k' \hat{\boldsymbol{\beta}}_\pi\} \\
&= N^{-1}\{\sum_{s^0} \hat{m}_{\pi k} + \sum_{s^0} (Y_k - \hat{m}_{\pi k}) S_k \hat{\pi}_k^{-1}\}
\end{aligned} \tag{2.40}$$

is the solution to the AIPW equations (2.39) where we estimate $E(U_k|\mathbf{x}_{s^0})$ by $\hat{U}_k = \left(\hat{m}_{\pi k} - \overline{\mathbf{Y}}_{s^0}\right)/\sigma^2$ for given σ^2 and with $\hat{\boldsymbol{\beta}}_\pi = (\sum_s \sigma_k^{-2} \hat{\pi}_k^{-1} \mathbf{X}_k \mathbf{X}_k')^{-1} \sum_s \sigma_k^{-2} \hat{\pi}_k^{-1} \mathbf{X}_k Y_k$ so that $\hat{m}_{\pi k} = \mathbf{x}_k' \hat{\boldsymbol{\beta}}_\pi$. From the latter two formulation of (2.40) it is clear that $\hat{\overline{\mathbf{Y}}}_{\pi m}$ may be viewed either as a π-estimator that incorporates an $m-$model or an $m-$estimator that incorporates a $\pi-$model.

As a solution to an AIPW equation $\hat{\overline{\mathbf{Y}}}_{\pi m}$ is CAN under correct $\pi-$model specification, to see that it is also CAN when the $m-$model holds

$$B(\hat{\overline{\mathbf{Y}}}_{\pi m}) \;=\; N^{-1} \sum_{s^0} E\left\{ (Y_k - \hat{m}_{\pi k}) \left(S_k \hat{\pi}_k^{-1} - 1 \right) \right\}. \tag{2.41}$$

Estimator 2.40 is not unique and there are numerous variations the fall into the πm (or DR) class of estimators of the population average. For example, Kang and Schafer (2007) suggest an estimator identical to 2.40 but for estimating the linear regression coefficients $\boldsymbol{\beta}$ by OLS rather than the WLS on $\hat{\pi}_k$. Another $\pi m-$estimator is that discussed by Bang and Robins (2005) which model $E(\mathbf{Y}_k|\mathbf{X} = \mathbf{x})$ by $m(\mathbf{x};\boldsymbol{\beta},\phi) = m(\mathbf{x}'\boldsymbol{\beta} + \phi\hat{\pi}_k^{-1})$ where $m^{-1}(\cdot)$ here is a known link function, that is a GLM model with explanatory variables including both the set $\mathbf{X}$ and the inverse of the estimated selection probabilities $\pi(\mathbf{x})$. The resulting πm-estimator is

$$\hat{\overline{\mathbf{Y}}}_{\pi m} = N^{-1} m(\mathbf{x}'\hat{\boldsymbol{\beta}} + \hat{\phi}\pi_k^{-1}(\mathbf{x};\hat{\boldsymbol{\alpha}})) \tag{2.42}$$

where $(\hat{\boldsymbol{\beta}}, \hat{\phi})$ solves

$$\sum_{s^0} \Big\{ S_k \partial m(\mathbf{x}'\boldsymbol{\beta} + \phi\hat{\pi}_k^{-1})/\partial(\boldsymbol{\beta}, \phi)[Y_k - m(\mathbf{x};\boldsymbol{\beta},\phi)] \Big\} = \mathbf{0}. \tag{2.43}$$

It is clear the estimator is CAN when the model $m(\mathbf{x};\boldsymbol{\beta},\phi)$ (with $\hat{\pi}_k^{-1}$ replaced by probability limit) is correct for $E(\mathbf{Y}_k|\mathbf{X} = \mathbf{x})$. Bang and Robins (2005) show, by referring to the normal equations (2.43), that (2.42) is the solution to the AIPW estimation equation which is sufficient to show that it is also CAN when the $\pi-$model holds. I note that the same proof supports the estimator alluded to by (D'Agostino Jr, 1998, p. 2277) which simply replaces $\hat{\pi}_k^{-1}$ with $\hat{\pi}_k$ so that $m(\mathbf{x};\boldsymbol{\beta},\phi) = m(\mathbf{x}'\boldsymbol{\beta} + \phi\hat{\pi}_k)$ models the conditional expectation of $\mathbf{Y}$.

One possible theoretical objection to $\hat{\overline{Y}}_{\pi m}$ is that when the π is either known or correctly modelled, $\hat{\overline{Y}}_{\pi m}$ can be less efficient than $\hat{\overline{Y}}_{\pi}$ if the model for $E(Y|s = 1, \mathbf{x})$ is badly misspecified. Robins (2002, Appendix 4) has developed an alternative DR estimator, referred to as $\hat{\overline{Y}}_{IPCW}$, that, as noted by Robins, Rotnitzky, and Bonetti (2001), is always guaranteed to be at least as efficient as $\hat{\overline{Y}}_{\pi}$ when the π is either known or correctly modelled. The GREG estimator in our discussion is such an estimator. Furthermore, in practice, it would be rare for the model for $E(Y|s = 1, \mathbf{x})$ to be so badly misspecified that $\hat{\overline{Y}}_{\pi m}$ was seriously inefficient.

A final note is the advantage of estimating the the population average with all three methods. Robins and Rotnitzky (2001) note that when $\hat{\overline{Y}}_{\pi m}$, $\hat{\overline{Y}}_{\pi}$ as well as $\hat{\overline{Y}}_{m}$ are calculated a comparison of the three estimators with one another serves as a useful goodness of fit test. For example Bang and Robins (2005) sketch this the general procedure: let $\hat{\sigma}^2_{\pi-\pi m}$ and $\hat{\sigma}^2_{m-\pi m}$ be the empirical variance of $(\hat{\overline{Y}}_{\pi} - \hat{\overline{Y}}_{\pi m})$ and $(\hat{\overline{Y}}_{m} - \hat{\overline{Y}}_{\pi m})$, respectively, calculated from a large number of nonparametric bootstrap replications of the survey data. Then the tests with rejection regions $\mid (\hat{\overline{Y}}_{\pi} - \hat{\overline{Y}}_{\pi m})/\hat{\sigma}_{\pi-\pi m} \mid > 1.96$ and $\mid (\hat{\overline{Y}}_{m} - \hat{\overline{Y}}_{\pi m})/\hat{\sigma}_{m-\pi m} \mid > 1.96$ are valid large sample 0.05 level tests of the null hypotheses that the π-model and the $m-$model, respectively, are correctly specified. These tests are not consistent in the sense that they may not be rejected although estimators converge both to a constant different from $\overline{\mathbf{Y}}_{s^0}$, however this is will most likely not be common 'trap' in practice.

Chapter 3

A Sequential Framework for Web Panel Survey Estimation

3.1 Introduction

In the previous chapter I have reviewed different key practices applicable to the Web panel survey sample inference problem. We discussed broad three estimation strategies which all fit into a basic missing data framework: the π,m and πm-estimation strategies.

In this chapter I suggest a sequential framework to the same question and will argue its specific merits in the panel case. The term sequential refers mainly to the assumption that the selection process from population of interest to survey set follows a sequence of selections, each step conditional on the previous subset created. Once invoked, however, all relevant variables and parameters need to be adjusted to this framework and certain relationships need to be specified.

The following chapter has the following outline. In section 3.2 I lay out the basic set up of the sequential framework and add any necessary notation convention, and in section 3.3 I describe the assumed available or observable data. In section 3.4.1 I use the language of ignorability, here in a sequential framework, and using a likelihood inferential setting I derive and outline broadly the necessary assumptions on selection or covariates for estimation. In section 3.4.2 I discuss the important idea of independence of the covariates used for establishing ignorability (or conditional independence) from the selection process. This idea is formally stated in section 3.4.3.

In section 3.5 I discuss in detail three estimation strategies which fit into our sequential framework and discuss their properties and relative strengths. This includes a πm-type estimator which has a sequential double robust property (section 3.5.3). I conclude this chapter in 3.6 with three simulation studies. The first tests statistical properties of the three estimators under correct and incorrect model specifications. In the second I propose and inspect a bootstrap variance estimator algorithm adapted to the sequential setting. I conclude the chapter with a short discussion on the fallacy of a coding convention popular in practice- assigning automatically to all non Web users the number zero to Internet related quantitative measurements. This allows practitioners to model (or weight) survey data using these observations. However, as I show this can lead to invalid results.

3.2 Description and Set up

As before, denote a finite population of units $k = 1, ..., N$ by s^0. Now define an $N \times 1$ vector of indicator variables $\mathbf{S}^t = (S_1^t, ..., S_k^t, ..., S_N^t)'$ of selection into subset $t \in \{0, 1, ..., T\}$ where $S_k^t = 1$ if unit k selects into the t selection set and $S_k^t = 0$ otherwise. The probability of realizing a particular set is denoted $Pr(\mathbf{S}^t = \mathbf{s}^t) = p(\mathbf{s}^t)$. To clarify, upper case bold $\mathbf{S}^t$ indicates the random vector of selection indicators of phase t, lower case bold $\mathbf{s}^t$ indicates a single realization of this random vector. Also let a lower case normal font s^t indicate the set of labels k which are selected in phase t, that is $s^t = \{k : s_k^t = 1\}$. By definition let $s^t \supseteq s^{t+1}$ $\forall t = 0, 1, ..., T$ as well as that the vector $\mathbf{S}^0 \equiv \mathbf{1}_N'$ as s^0 represents the finite population of interest.

In general, each unit k is associated with a sequence of $t = 0, .., T$ selection or designed sampling phases. The history of any variable up to a period t is denoted by an underline. For notational efficiency the initial period is not normally mentioned explicitly. For example $\underline{S}_k^t = \underline{s}_k^t$ means $(S_k^1, S_k^2, ..., S_k^t) = (s_k^1, s_k^2, ..., s_k^t)$ where each $s_k^t \in \{0, 1\}$ and the sequence starts with period 1 rather than 0. A general description of the overall sequence of subsets from the population to the final survey set is described then by $\underline{\mathbf{S}}^T = (\mathbf{S}^1, \mathbf{S}^2, ..., \mathbf{S}^t, ..., \mathbf{S}^T)$, an $N \times T$ matrix which can be sorted to have a monotone zero/one structure.

Associated with each phase are random variables $\mathbf{X}^t$ so that the aggregate variables up to phase t are $\underline{\mathbf{X}}^t = (\mathbf{X}^0, \mathbf{X}^1, ..., \mathbf{X}^{t-1}, \mathbf{X}^t)$, where $\mathbf{X}^t = (\mathbf{X}_1^t, ..., \mathbf{X}_k^t, ..., \mathbf{X}_N^t)'$ is an $N \times P^t$ matrix associated with the t phase selection process. The matrix $\underline{\mathbf{X}}^t$ is thus of dimension $N \times \sum_{j=0}^t P^j$ with typical component $\underline{\mathbf{X}}_k^t = (\mathbf{X}_k^0, \mathbf{X}_k^1, ..., \mathbf{X}_k^t)'$ of dimension $1 \times \sum_{j=0}^t P^j$ defined for any $t = 0, 1, ..., T$. Over the entire T phase process I write

$$\underline{\mathbf{X}}^T \;=\; (\mathbf{X}^0, \mathbf{X}^1, ..., \mathbf{X}^t, ..., \mathbf{X}^{T-1}, \mathbf{X}^T) = (\underline{\mathbf{X}}^{T-1}, \mathbf{Y})$$

where variables associated with the last phase $\mathbf{X}^T$ are the measurement variables of interest representing the survey questions and are denoted also by $\mathbf{Y}$. The exact relationship between subset s^t and covariates $\mathbf{X}^t$ and survey measurement $\mathbf{Y}$ will be formally defined in the following two sections. In some cases, for emphasis I shall add a subscript indicating the set of population units over which a vector or matrix are defined over. For example $\underline{\mathbf{S}}^T$ or $\underline{\mathbf{X}}^T$above can be denoted $\underline{\mathbf{S}}^T_{s^0}$ and $\underline{\mathbf{X}}^T_{s^0}$ as well.

In addition, to clarify modelling assumptions, we introduce a set of unknown variables denoted here by $\mathbf{W} = (\mathbf{W}_1, ..., \mathbf{W}_k, ..., \mathbf{W}_N)'$ a matrix of unknown dimension. A similar technique is used in Smith and Sugden (1988) on which we expand in the following.

3.3 Selection, Full and Observed Population Distributions

In its most general form each phase $t = 1, .., T$ of the sequential selection process may be influenced by $\mathbf{Y}, \underline{\mathbf{X}}^{T-1}$ or $\mathbf{W}$ as well as $\underline{\mathbf{S}}^{t-1}$, the past selection pattern. So, I can write the general form of a T phase selection process by

$$Pr(\underline{\mathbf{S}}^T_{s^0} = \underline{\mathbf{s}}^T_{s^0} | \underline{\mathbf{X}}^{T-1}_{s^0} = \underline{\mathbf{x}}^{T-1}_{s^0}, \mathbf{Y}_{s^0} = \mathbf{y}_{s^0}, \mathbf{W}_{s^0} = \mathbf{w}_{s^0}) \tag{3.1}$$

described compactly by

$$p(\underline{\mathbf{s}}^T | \underline{\mathbf{x}}^T, \mathbf{w}) \;=\; \Pi_{t=1}^T p(\mathbf{s}^t | \mathbf{s}^{t-1}, \underline{\mathbf{x}}^T, \mathbf{w}) \tag{3.2}$$

where $p(s^t_k = 0 | s^{t-1}_k = 0, \underline{\mathbf{x}}^{t-1}, \mathbf{w}) = 1$ always, for any $k \in s^{t-1}$.

Similarly, under the T phase framework, write the joint distribution of the associated variables $(\underline{\mathbf{X}}^T, \mathbf{W}) = (\underline{\mathbf{x}}^T, \mathbf{w})$ by

$$\begin{aligned} f(\underline{\mathbf{x}}^T, \mathbf{w}) &= f(\mathbf{x}^T | \underline{\mathbf{x}}^{T-1}, \mathbf{w}) f(\underline{\mathbf{x}}^{T-1}, \mathbf{w}) \\ &= f(\mathbf{y} | \mathbf{w}, \underline{\mathbf{x}}^{T-1}) f(\mathbf{w}, \underline{\mathbf{x}}^{T-1}) \end{aligned} \tag{3.3}$$

where the vectors are of length N, the population size.

As before, our interest is in inference on aspects of the full finite population, specifically linear functions such as the average or total of $\mathbf{X}^T = \mathbf{Y}$ the survey measurements of interest. Denote the full population by $\underline{D}^T = (\underline{\mathbf{X}}^T, \mathbf{W}, \underline{\mathbf{S}}^T) = \{Y_k, \underline{\mathbf{X}}^{T-1}_k, W_k, \underline{S}^T_k\}_{k=1}^N$ with values $\underline{d}^T = (\underline{\mathbf{x}}^T, \mathbf{w}, \underline{\mathbf{s}}^T)$, and combining (3.2) with (3.3) gives the population joint distribution

$$\begin{aligned} f(\underline{d}^T) &= p(\underline{\mathbf{s}}^T | \underline{\mathbf{x}}^T, \mathbf{w}) f(\mathbf{x}^T | \underline{\mathbf{x}}^{T-1}, \mathbf{w}) f(\mathbf{w}, \underline{\mathbf{x}}^{T-1}) \\ &= \Pi_{t=1}^T p(\mathbf{s}^t | \underline{\mathbf{s}}^{t-1}, \underline{\mathbf{x}}^{T-1}, \mathbf{y}, \mathbf{w}) f(\mathbf{y} | \mathbf{w}, \underline{\mathbf{x}}^{T-1}) f(\mathbf{w}, \underline{\mathbf{x}}^{T-1}). \end{aligned} \tag{3.4}$$

In the web panel context take for simple illustration the case where $T = 2$. First, a relatively small subset s^1 of s^0 volunteers into a panel, while for the purpose of an ad hoc survey the panel management samples a set s^2 from the panel s^1 by a known design, which nevertheless suffers from various forms of non cooperation. Without further assumptions both selection mechanisms $p(s^1|\underline{\mathbf{x}}^1, \mathbf{y}, \mathbf{w})$ and $p(s^2|s^1, \underline{\mathbf{x}}^1, \mathbf{y}, \mathbf{w})$ may be influenced by any or all of the variables $(\mathbf{Y}, \mathbf{X}^0, \mathbf{X}^1, \mathbf{W})$. These may represent, respectively, survey variables of interest $(\mathbf{y})$, covariates such as social demographic indicators, age and gender $(\mathbf{x}^0)$, covariates associated with the panel population , for example Internet related behavioral data or panel operational tracking data, e.g. historic response rates $(\mathbf{x}^1)$. Finally the creation of the subsets may be effected by data not quantifiable or that can never be observed $(\mathbf{w})$.

Next, we need to clarify what data can (at least in theory) be assumed observed and available. In chapter 2 we assumed that $(\mathbf{x}, \mathbf{s}, \mathbf{y}_s)$, the observed values of $(\mathbf{X}, \mathbf{S}, \mathbf{Y}_s) = \{\mathbf{X}_k, S_k, \mathbf{Y}_k \cdot S_k\}_{k=1}^N$ are observed or at least possible to observe[1].

Here we restate this differently. Now I make the assumption that values $(\mathbf{x}^t_{s^t}, \mathbf{s}^t)$ of $(\mathbf{X}^t_{s^t}, S^t)$ for each $t = 0, ..., T$ are observable for drawing population inference. That means values $\mathbf{x}^0$ are observed for all members of s^0, that values $\mathbf{x}^1_{s^1}$ of $\mathbf{X}^1$ amongst members of s^1 can also be observed, and so on. For $t = T$ recall that $\mathbf{x}^T = \mathbf{y}$ and thus $\mathbf{x}^T_{s^T} = \mathbf{y}_{s^T}$ denotes the observed values of the survey sample measurements of interest $\mathbf{Y}$. As I shall argue, this structure is particularly useful for the panel inference question when the phases of selection are defined judiciously. Now, denote the observed data by

$$\begin{aligned} \underline{D}^T_{s^T} &= (\mathbf{X}^0, \mathbf{X}^1_{s^1}, ..., \mathbf{X}^{T-1}_{s^{T-1}}, \mathbf{X}^T_{s^T}, \underline{S}^T) \\ &= (\mathbf{Y}_{s^T}, \underline{\mathbf{X}}^{T-1}_{s^{T-1}}, \underline{S}^T) \end{aligned}$$

which can be written as well as $\{Y_k \cdot \underline{S}^T_k, \underline{\mathbf{X}}^{T-1}_k \cdot \underline{S}^{T-1}_k, \underline{S}^T_k\}_{k=1}^N$.

For illustration figure 3.1 gives such a description of a dataset for $T = 3$. This is an idealized scenario as in practice such a dataset most likely suggests one based on census data (thus a small set of $\mathbf{x}$) or inversely implies a very narrow definition for the finite population of interest (thus a small set s^0). Thus, in practice the analyst will in fact observe only $(\underline{\mathbf{x}}^3_{s^3}, \mathbf{s}^3)$ which makes most estimation strategies impractical without supplementary data. To anticipate later discussion, I note here that in chapter 4 I will introduce the idea of inference using data from a parallel random reference survey which will increase the applicability of the inferential approach

As in the single phase case of chapter 2, to motivate possible inference strategies and explicitly state distributional assumptions, I link the joint distribution of

[1] As our interest is usually on the average or total of the population, in many cases only a sufficient population summary statistics such as the average of $\mathbf{X}$ or an estimator of this average is necessary. This allows a much wider set of covariates $\mathbf{X}$ to be considered in estimation.

S^0	$\mathbf{X}^0$	S^1	$\mathbf{X}^1_{s^1}$	S^2	$\mathbf{X}^2_{s^2}$	S^3	$\mathbf{Y}$
1	$\mathbf{x}^0_1$	0	-	0	-	0	-
1	$\mathbf{x}^0_2$	0	-	0	-	0	-
1	.	1	$\mathbf{x}^1_1$	0	-	0	-
1	.	1	.	0	-	0	-
1	.	1	.	0	-	0	-
1	$\mathbf{x}^0_k$	1	$\mathbf{x}^1_k$	1	-	0	-
1	.	1	.	1	$\mathbf{x}^2_1$	0	-
1	.	1	.	1	.	1	$\mathbf{y}_1$
1	.	1	.	1	.	.	.
1	$\mathbf{x}^0_N$	1	$\mathbf{x}^1_{n_1}$	1	$\mathbf{x}^1_{n_2}$	1	$\mathbf{y}_{n_3}$

Figure 3.1: A description of an idealized population dataset for a $T = 3$ sequence. The hyphen $(-)$ notation indicates an unobserved or undefined value for population member $k \in s^0$.

$\underline{D}^T_{s^T}$ with values $\underline{d}^T_{s^T} = (\underline{\mathbf{x}}^T_{s^{T-1}}, \underline{\mathbf{s}}^T)$ to the theoretical population joint distribution $f(\underline{d}^T)$ by integrating out non observed elements of the population, that is

$$f(\underline{d}^T_{s^T}) = \int f(\underline{d}^T) d\mathbf{y}_{\overline{s}^T} d\underline{\mathbf{x}}^{T-1}_{\overline{s}^{T-1}} d\mathbf{w} \tag{3.5}$$

where $\overline{s}^t = s^0 - s^t$, that is the complementary set of finite population units. For $T = 2$ the full population joint distribution is

$$f(\underline{d}^2) = p(\underline{s}^2|\underline{\mathbf{x}}^1, \mathbf{y}, \mathbf{w}) f(\mathbf{y}|\underline{\mathbf{x}}^1, \mathbf{w}) f(\underline{\mathbf{x}}^1, \mathbf{w}) \tag{3.6}$$

which is linked to the observed data distribution by

$$f(\underline{d}^2_{s^2}) \quad = \quad \int f(\underline{d}^2) d\mathbf{y}_{\overline{s}^2} d\underline{\mathbf{x}}_{\overline{s}^1} d\mathbf{w}. \tag{3.7}$$

The actual observed dataset $\underline{d}^2_{s^2}$ then includes values $\mathbf{x}^0_{s^0}$ such as basic social demographic records for each member of the population, values $\mathbf{x}^1_{s^1}$ recorded only for panel members, of e.g. Internet-related measurements such as *'hours surfing'* or '*most popular websites visiting*' or panel-operational data such as past survey response rates, and values $\mathbf{y}_{s^2}$ of the panelists' Web-survey recorded responses.

3.4 Inference Under a Sequential Framework

3.4.1 A Likelihood Perspective at Conditions for Inference

I now derive the specific conditions for inference under the sequential framework taking a likelihood inference perspective. I look here at a two phase case but the conclusions are expanded to a general T phase case in subsequent sections. What is clear from the following derivation is that there are two possible sets of assumptions required for derivation. The first leaves the selection process unspecified, in the sense that it can rely on unobserved covariates $\mathbf{W}$, while the second leaves the covariate distribution unspecified.

The guiding question is what assumptions need to be made so that inference based on the joint specification of the entire finite population and selection models is identical to that over the observed data ignoring[2] the selection process. The following can be seen as a generalization of Smith and Sugden (1988) who-while invoking a two phase sequence- limit their discussion to only $\mathbf{X}^0$ type of variables for modelling.

Suppose that we observe $\mathbf{x}^0_{s^0}$, $\mathbf{x}^1_{s^1}$ and $\mathbf{y}_{s^2}$. As before $\mathbf{W}$ denotes other unobserved covariates. Also, let $\mathbf{y}_{\bar{s}^2}$ represent the unobserved values over units $\bar{s}^1 \cup (s^1 \cap \bar{s}^2)$, that is the units which were not selected in phase one ($S^1_k = 0$) or that were selected in phase one but not in the second phase ($S^2_k = 0$ and $S^1_k = 1$). Similarly $\mathbf{x}^1_{\bar{s}^1}$ represents the unobserved values over units $\bar{s}^1$, that is the units which were not selected in phase one ($S^1_k = 0$).

The joint distribution (3.3) of $\mathbf{Y}, \mathbf{W}$ and $\underline{\mathbf{X}}^1$over the entire population s^0 can then be written as

$$f(\mathbf{y}|\mathbf{w}, \underline{\mathbf{x}}^1; \boldsymbol{\beta}) f(\mathbf{w}|\underline{\mathbf{x}}^1; \boldsymbol{\alpha}) f(\mathbf{x}^1|\mathbf{x}^0; \boldsymbol{\phi}^1) f(\mathbf{x}^0; \boldsymbol{\phi}^0) \tag{3.8}$$

where we assume parameters $\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\phi}^1, \boldsymbol{\phi}^0$ are distinct as in Rubin (1976). Given the data available a *face value* analysis (Dawid and Dickey, 1977) is one that relies on the observed data

$$f(\mathbf{x}^0; \boldsymbol{\phi}^0) f(\mathbf{x}^1|\mathbf{x}^0, s^1; \boldsymbol{\phi}^1) \int \int f(\mathbf{y}|\underline{\mathbf{x}}^1; \boldsymbol{\beta}) d\mathbf{y}_{\bar{s}^2} d\mathbf{x}^1_{\bar{s}^1}. \tag{3.9}$$

A general formulation of the two phase selection process can be described by

$$p(s^1|\mathbf{w}, \mathbf{y}, \underline{\mathbf{x}}^1) p(s^2|\mathbf{w}, \mathbf{y}, \underline{\mathbf{x}}^1, s^1) \tag{3.10}$$

so a *full likelihood analysis* based on the likelihood function from (3.3) by using (3.8) and (3.10) and integrating out unobserved values $\mathbf{y}_{\bar{s}^2}$,$\mathbf{x}^1_{\bar{s}^1}$ and $\mathbf{w}$.

$$\begin{aligned} f(\mathbf{x}^0; \phi^0) \int & p(s^1|\mathbf{w}, \mathbf{y}, \underline{\mathbf{x}}^1) p(s^2|\mathbf{w}, \mathbf{y}, \underline{\mathbf{x}}^1, s^1) \\ & \times f(\mathbf{y}|\mathbf{w}, \underline{\mathbf{x}}^1; \boldsymbol{\beta}) f(\mathbf{w}|\underline{\mathbf{x}}^1; \boldsymbol{\alpha}) \quad f(\mathbf{x}^1|\mathbf{x}^0; \boldsymbol{\phi}^1) d\mathbf{y}_{\bar{s}^2} d\mathbf{x}^1_{\bar{s}^1} d\mathbf{w}. \end{aligned} \tag{3.11}$$

[2] As frequently the case, this is a misleading term. The same assumptions for ignoring the selection process are the basis for selection based inference as well.

Alternatively, if integrating out the unobservable values $\mathbf{w}$ the full likelihood is

$$f(\mathbf{x}^0;\boldsymbol{\phi}^0)\int p(s^1|\mathbf{y},\underline{\mathbf{x}}^1;\boldsymbol{\alpha})p(s^2|\mathbf{y},\underline{\mathbf{x}}^1,s^1;\boldsymbol{\alpha}) \qquad (3.12)$$
$$\times f(\mathbf{y}|\underline{\mathbf{x}}^1;\boldsymbol{\beta}) \quad f(\mathbf{x}^1|\mathbf{x}^0;\phi^1)d\mathbf{y}_{\bar{s}^2}d\mathbf{x}^1_{\bar{s}^1}.$$

Given distinct parameters the mechanisms are ignorable for likelihood inference if they can be taken outside of the integral in (3.11) or (3.12). In our two phase case, sufficient conditions for ignorability are either of type A or B

Conditions A	Conditions B
(1) $\mathbf{X}^1 \perp \mathbf{S}^1|\mathbf{x}^0,\phi^1$	(1) $\mathbf{X}^1 \perp \mathbf{S}^1|\mathbf{x}^0,\phi^1$
(2) $\mathbf{S}^1 \perp \mathbf{Y}|\mathbf{w},\mathbf{x}^0$	(2) $\mathbf{S}^1 \perp \mathbf{Y}|\mathbf{x}^0;\alpha$
(3) $\mathbf{S}^2 \perp \mathbf{Y}|\underline{\mathbf{x}}^1\mathbf{w},s^1$	(3) $\mathbf{S}^2 \perp \mathbf{Y}|\underline{\mathbf{x}}^1,s^1;\alpha$
(4) $\mathbf{Y} \perp \mathbf{W}|\underline{\mathbf{x}}^1;\beta$	
(5) $\mathbf{W} \perp \mathbf{X}^1|\mathbf{x}^0;\alpha$	

Table 3.1: Sufficient conditions for ignorability assumption to hold for likelihood inference. Conditions A refers to description (3.11) while Conditions B refers to (3.12)

where A(1) and B(1) are required to 'bring out' $f(\mathbf{x}^1|\cdot)$ from the integration; A(2) and A(3) relate to the selection mechanisms while A(4) and A(5) are required for the modelling of $\mathbf{y}$.

What one may learn from the two sets of conditions is that when the selection process is unknown, and may be influenced by unobserved and/or unknown covariates, inference based on the observed data can rely on model specification taken for the (i) selection process, or either over (ii) the observed covariates. The latter is represented in conditions set B where the two selection phases are modelled over known covariate values, while the former approach is given in set A where the observed covariates are modelled while selection is left unspecified as they rely on unobservables.

Both condition sets bring out the requirement for independence between the selection process and the covariates used for modelling. A fundamental requirement I elaborate in the following section.

3.4.2 Independence of the Conditioning Covariate

Under the influence of the remarkable popularity of the propensity score balancing approach (Abadie and Imbens, 2009), especially in the panel context (Lee, 2006), the general guidance to practitioners is to include all available covariates $\mathbf{X}$ so that conditional independence between $\mathbf{Y}$ and $\mathbf{S}$ is achieved (for example Rubin, 2004), although this recommendation has been challenged as I have reviewed in section 3.5.2. As we have just shown, a further constraint on this approach is the condition that covariates $\mathbf{X}$ must be independent of selection [3].

As Lechner and Miquel (2005) points out, this issue in not addressed in the classic formulation of Rosenbaum and Rubin (1983), and clarifications on the topic (Rosenbaum, 1984) have been, from the personal experience of this author, largely been ignored in practice. This omission to the simple formulation has had the effect that this issue is overlooked and data analysts routinely use covariates, either directly or after invalid imputation that should not be included. I give an numerical example, in the panel context, of this in (3.6.3).

The correct set $\mathbf{X}$ that achieves conditional independence is impossible to know and drives including more and more covariates. However, the high cost of survey data collection (even for Web panel data) limits the size of datasets and thus the pool of covariates we can include. This trade-off is particularly true for balancing adjustment approaches such as matching or weighting which are especially 'data hungry'. And so, in practice the analyst inevitably gravitates to the path of a parsimonious parametric model specification, searching for one that behaves well under validation tests.

The problem is that in many applications, and certainly in the panel case, the most highly informative covariates which may achieve parsimony and would otherwise be included in an estimation model cannot be used as they are not independent of selection.

To put things more generally, take an $m-$estimator which assumes a linear regression relationship between Y and (X, Z). Inference of the population average requires only that $E(Y|S = 1, \mathbf{x}, \mathbf{z}) = E(Y|S = 0, \mathbf{x}, \mathbf{z})$. Using causal inference notation denote $\mathbf{Z}|S = 1$ by $\mathbf{Z}(1)$ and $\mathbf{Z}|S = 0$ by $\mathbf{Z}(0)$. Since $\mathbf{Z}(1)$ is not observed for non panelists and $\mathbf{Z}(0)$ is not observed for panelists, we must condition on the observed $\mathbf{z}$ and thus adjustment may be on the same value, but different adjustment variables. That is $E(Y|S = 1, \mathbf{X} = \mathbf{x}, \mathbf{Z}(1) = \mathbf{z}) \neq E(Y|S = 0, \mathbf{X} = \mathbf{x}, \mathbf{Z}(0) = \mathbf{z})$. A visual representation of this is given in the left hand side panel of figure 3.2.

In the Web panel case this issue is easily understood with the use of Internet related covariates which are on the one hand highly informative practically irreplaceable - in explaining the survey process, but are also naturally correlated with it. Take for example the approach of matching on π (Rosenbaum and Rubin

[3] The widely cited Lechner (2008) uses the term endogeneity (and exogeneity) meaning that the variable is (not) influenced by the selection, which is not exactly in line with the common use of this language in econometrics (e.g. Engle et al., 1983). In observational studies the term used is post-treatment variables.

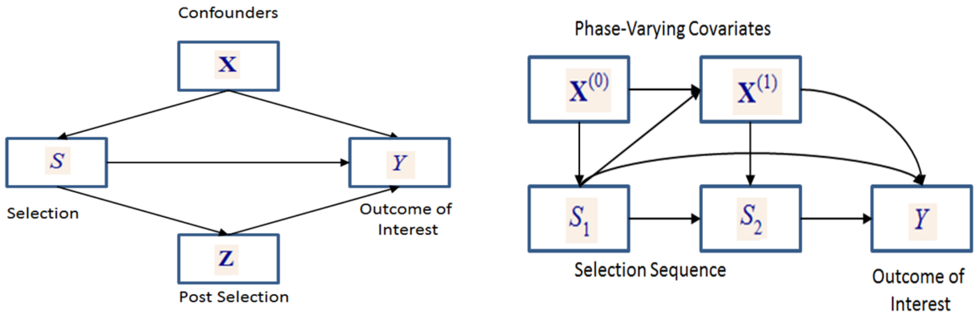


Figure 3.2: Direct Acyclic Graphs showing single phase selection vs sequential selection process. Arrows represent a causal relationship

(1983)) under a $T = 1$ framework with s denoting the Web panel survey response set and Z_k indicating personal sentiment on '*Safety in shopping Online*'. Clearly, this type of sentiment of Online behaviour is likely to be related to the process underlying the frequent use of the Internet. Ignoring covariate independence, we estimate $p(s_k = 1|Z_k(s) = z) = \pi_k$ for $S_k = 1$ and $S_k = 0$ and match panelists to general population records by proximity to π_k. However, the matching may be on the same values but measuring in effect a different variable.

A more extreme example is when an adjustment covariate $\mathbf{Z}$ is perfectly correlated with the selection indicator in the sense that $p(s_k = 1|z_k) = \begin{cases} \pi_k & \text{if } z = 1 \\ 0 & \text{if } z = 0 \end{cases}$ in which case the common support assumption is violated as well. Under the same setting an $m-$estimation approach such as BLUP will result in extrapolation of the regression model as $f(y|z = 0) = f(y|s = 0) = f(y_{\overline{s}})$ is unobserved and so the estimation model cannot be verified [4].

The usefulness of the sequential framework here is twofold. First, with the usual conditional independence between selection and the outcome variable, it naturally derives and highlights the exact covariate independence from selection necessary, as seen in the previous section. More positively, by the nature of the sequential formulation, the constraint on the conditioning covariates are fragmented and weakened which then allows to increase the potential pool of variables to include in estimation models. Covariates that are associated with the overall selection process are identified explicitly and may turn into independent within a selection process phase. Graphically, this idea is described in the right hand panel of 3.2 where $\mathbf{X}^1$ type variables are not independent of the selection process $p(\underline{s}^2)$ but are independent of the second phase of the selection process.

[4] Rosenbaum (1984) terms covariates such as $\mathbf{Z}$ as post-treatment variables and states that '... adjustments for post-treatment variables... are justified only when they are unnecessary ', that is when $f(y|z = 1) = f(y|z = 0)$ in which case adjustment is unnecessary.

This can be especially useful property for the web panel problem. The key for the analyst is to find a judicious sequential framework which maximizes the amount of available covariates within the ignorability constraint. In the panel context such a framework splits the survey process into three phases: (1) a large subset s^1 of the population s^0 randomly selects into the set of population members who are regular Web users; (2) From this population a set s^2 of the population volunteers into a Web panel, and (3) the panel management designs and selects a sample s^3 to participate in a survey study. The specific deconstruction of the panel selection into two phases, allows us to use the highly informative general Web related covariates in modelling the volunteering into the panel, which otherwise cannot be used as they are not independent of the overall process. The separation of the within-panel survey sampling mechanism allows us to minimize the chance of model misspecification as this part of the process is largely controlled and well monitor by the panel management team.

3.4.3 Restrictions on the Selection Process

Let us reflect the ideas discussed above by establishing formally the relationship between the observable covariates $\mathbf{X}^t\ ; t = 0, ..., T$ and the selection process. I continue using the concept of unobservable covariates $\mathbf{W}$ to establish the minimal sufficient adjustment sets (Pearl, 2012) which may be different for π or m estimation approaches depending on $\mathbf{Y}$ survey variables of interest studied.

I introduce now restrictions on the general selection process (3.2) of a T phase selection distribution generalizing conditions (1)-(3) in both sets (A) and (B) of table 3.1. Assume that for each $t = 1, ..., T$

$$p(s^t|\underline{s}^{t-1}, \underline{\mathbf{x}}^T, \mathbf{w}) = p(s^t|s^{t-1}, \underline{\mathbf{x}}^{t-1}_{s^{t-1}}, \mathbf{w})$$

so that overall that the the selection of set s^T follows the distribution

$$p(\underline{s}^T|\underline{\mathbf{x}}^T, \mathbf{w}) = \Pi^T_{t=1} p(s^t|s^{t-1}, \underline{\mathbf{x}}^{t-1}_{s^{t-1}}, \mathbf{w}). \tag{3.13}$$

Selection now holds the following properties: (i) each phase is conditionally independent of $\mathbf{y}$ and so overall it is independent of $\mathbf{Y}$. A broader condition which includes the independence from $\mathbf{Y}$ is that (ii) each selection of set s^t from set s^{t-1} is independent of items $\mathbf{X}^t, \mathbf{X}^{t+1}, ..., \mathbf{X}^T$ given $\underline{\mathbf{X}}^{t-1}, \mathbf{W}$ which means that selection s^t depends only on observed variables associated with earlier phases and $\mathbf{W}$. (iii) for each t and given $\mathbf{W}$, selection of s^t from set s^{t-1} may depend only on the *values* of $\underline{\mathbf{X}}^{t-1}$ that are observed or may be observed, that is $\underline{\mathbf{X}}^{t-1}_{s^{t-1}}$ the values of the $\underline{\mathbf{X}}^{t-1}$ which can be observed. Thus unobserved values $\underline{\mathbf{X}}^{t-1}_{\bar{s}^{t-1}}$ associated with unsampled units or possibly ill defined in the finite population context (e.g. Web behavioural measurements which are meaningless over s^0) can be ignored.

In our $T = 2$ case, and ignoring $\mathbf{X}$, this means that s^2 then may depend on $\mathbf{X}^1_{s^1}$ and $\mathbf{X}^0_{s^0}$, for example a selection process proportionate to the average income brackets in the general population and the average spending bracket Online in the Internet connected population. It is independent, however, of the average spending Online of non Internet users.

3.5 Three Estimation Strategies Under the Sequential Framework

In this section I propose three estimation strategies under the sequential framework, broadly mirroring my discussion within the one phase framework in chapter 2. The starting point of all three strategies is that the selection process is restricted as defined in (3.13) which means that the observed population distribution can be described by

$$\begin{aligned} f(\underline{d}^T_{s^T}) &= \int f(\underline{d}^T) d\underline{\mathbf{x}}^T_{\bar{s}^T} d\mathbf{w} \\ &= \int f(\underline{\mathbf{x}}^T_{s^T}, \mathbf{w}) \Pi^T_{t=1} p(s^t|s^{t-1}, \underline{\mathbf{x}}^{t-1}_{s^{t-1}}, \mathbf{w}) d\mathbf{w} \qquad (3.14) \end{aligned}$$

Similar to my comment on the the one phase case in chapter 2 one may posit different statistical models by making different assumptions on the components of (3.14). More specifically (3.14) suggests that estimation can rely on modelling either the distribution of $\underline{s}^T$or that of $\underline{\mathbf{X}}^T_{s^T}$ (or both) when sufficient conditions that allow taking the relevant distributions outside of the integral (3.14) are stated. In the following sections I term an π-estimation strategy as one that relies on modelling $\underline{S}^T$, I term an m-estimation strategy as one that relies on modelling $\underline{\mathbf{X}}^T_{s^T}$, and a $\pi m-$strategy as one that models the distribution of both $\underline{S}^T$and $\underline{\mathbf{X}}^T_{s^T}$.

3.5.1 Outcome-Model Based Estimation

Consider a T phase selection process where the observed data $\underline{d}^T_{s^T}$ follows the distribution described in (3.14). For $m-$estimation, further assume that the unobserved covariates $\mathbf{W}$ have no effect on the outcome measurement distribution given the observed covariates, that is

$$f(\mathbf{x}^t_{s^t}|\underline{\mathbf{x}}^{t-1}_{s^{t-1}}, \mathbf{w}) = f(\mathbf{x}^t_{s^t}|\underline{\mathbf{x}}^{t-1}_{s^{t-1}}) \text{ for } t = 1, .., T. \qquad (3.15)$$

Under (3.15) the observed data distribution is now

$$f(\underline{d}^T_{s^T}) \quad = \quad f(\underline{\mathbf{x}}^T_{s^T}) \times \int p(\underline{s}^T|\underline{\mathbf{x}}^{T-1}_{s^{T-1}}, \mathbf{w}) f(\mathbf{w}) d\mathbf{w}. \qquad (3.16)$$

A likelihood estimation approach would be to assume a parametric model $f(\underline{\mathbf{x}}^T; \boldsymbol{\theta})$ for the population covariate data of $\underline{d}^T$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with a finite-dimensional space $\boldsymbol{\Theta}$ and estimate the population parameters by maximizing $\log f(\underline{\mathbf{x}}^T; \boldsymbol{\theta})$ the log-likelihood.

To allow comparison to an alternative approach take the case where $\underline{\mathbf{X}}^T$ is a matrix of $N \times \sum_{t=0}^{T} P^t$ representing an independent and identically distributes sample of N observations from a multivariate normal distribution (MVN) with mean vector $\boldsymbol{\mu}_T = (\mu_{01}, \mu_{02}, ..., \mu_{0P^0}, \mu_{11}, ..., \mu_{1P^1}, ..., \mu_{T1}, ..., \mu_{P^T})$ and covariance matrix $\boldsymbol{\Sigma}_T$ of dimension $\sum_{t=0}^{T} P^t$ on $\sum_{t=0}^{T} P^t$. Our interest is in estimating only the mean of the survey variables $\mathbf{X}^T = (X_1^T, ..., X_{P^T}^T)$ and the relevant elements of the covariance matrix.

The observed distribution of 3.16 is

$$f(\underline{\mathbf{x}}_{s^T}^T; \boldsymbol{\theta}) = \Pi_{k=1}^{n_T} f(\underline{\mathbf{x}}_k^T; \boldsymbol{\theta}) \Pi_{k=n_T+1}^{n_{T-1}} f(\underline{\mathbf{x}}_k^{T-1}; \boldsymbol{\theta}) \cdots \cdot \Pi_{k=n_1+1}^{n_0} f(\mathbf{x}_k^0; \boldsymbol{\theta}) \tag{3.17}$$

where $n_0 = N$ the size of the finite population. Under (3.16) the log-likelihood of (3.17) is appropriate for inference[5] and has the following form

$$l(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T | \underline{\mathbf{x}}_{s^T}^T) = \sum_{t=T}^{0} \left[-\frac{n_t - n_{t+1}}{2} ln|\boldsymbol{\Sigma}_t| - \frac{1}{2} \sum_{k \in s^{t-1} - s^t} (\underline{\mathbf{x}}_k^t - \underline{\boldsymbol{\mu}}_t) \boldsymbol{\Sigma}_T^{-1} (\underline{\mathbf{x}}_k^t - \underline{\boldsymbol{\mu}}_t)' \right]$$

where $|\boldsymbol{\Sigma}_t|$ denotes the determinant of Σ_t, the scalar $n_{t+1} = 0$ for $t = T$ and $\underline{\mathbf{x}}_k^t, \underline{\boldsymbol{\mu}}_t$ represent respectively the unit value and mean of the vector of variables associated with the first t population subsets.

However, as in the one phase selection framework, the likelihood equations do not have an obvious solution. A way around this is to use the monotone data structure inherent the sequential selection process and to factorize the joint distribution (Little and Rubin, 2002, section 7.4.2) of the observed data $\underline{\mathbf{X}}_{s^T}^T$.

Given distinctness[6], the likelihood of $\boldsymbol{\theta}$ can be maximized by maximizing separately for each 'block' $t = 1, ..., T$. Specifically, this is done by (1) calculating the mean vector and covariance matrix of $\mathbf{X}^0$ over s^0, and (2) calculating the multivariate regression of $\mathbf{X}^1$ on $\mathbf{X}^0$ over s^1, (3) calculating the multivariate regression of $\mathbf{X}^2$ on $\underline{\mathbf{X}}^1$ over s^2 , and so on until step $(T+1)$ calculate the (multivariate) regression of $\mathbf{X}^T$ on $\underline{\mathbf{X}}^{T-1}$ over s^T.

[5] for likelihood inference we need as well to assume that parameters $\boldsymbol{\theta}$ are distinct of the selection mechanism parameters (which are not specified), while for Bayesian inference these two parameter sets must have an independent prior distribution.

[6] The parameter space for $\boldsymbol{\theta}$ is the standard parameter space with no prior restrictions then $(\boldsymbol{\beta}_{T-1}, \boldsymbol{\beta}_{T-2}, ..., \boldsymbol{\beta}_0)$ are distinct in the sense that the their joint parameter space is the product of the individual parameter spaces.

To ground the idea and allow empirical comparison to a competing method I return to the case $T = 2$ with data described in table 3.2. Estimation follows the following steps: (1) calculate the mean and sample variance for X_1^0 and X_2^0, (2) compute the ML estimates of the four multivariate regression coefficients of $\mathbf{X}^1$ on $\mathbf{X}^0$ and its residual covariance matrix , and (3) calculate the regression coefficient estimates for Y on $\underline{\mathbf{X}}^1$ and its residual variance. By using the sweep operator (Little and Rubin, 2002, p 112) the ML estimator for the mean vector $\hat{\boldsymbol{\mu}}$ for all 5 variables can be found as well as the covariance matrix. The ML estimator for the mean of the survey variable Y is $\hat{\mu}_y = 27.047$.

x_1^0	x_2^0	x_1^1	x_2^1	y	$\hat{m}_0$
78.5	6	7	23	50	52.8
74.3	15	1	29	52	52.8
104.3	8	11	56	20	23.6
87.6	8	11	31	47	41.8
95.9	6	7	52	33	33.8
109.2	9	11	55	22	17.8
102.7	17	3	71	*(6)*	20.8
72.5	22	1	31	*(44)*	51.1
93.1	18	2	54	*(22)*	30.7
115.9	4	*(21)*	*(47)*	*(26)*	13.0
83.8	23	*(1)*	*(40)*	*(34)*	38.3
113.3	9	*(11)*	*(66)*	*(12)*	13.3
109.4	8	*(10)*	*(68)*	*(12)*	18.1

Table 3.2: The data under a $T = 2$ framework. Values in parentheses are considered non observed. Covariates X_1^0, X_2^0 are observed over the entire dataset $N = 13$, two covariates X_1^1, X_2^1 are observed only over $n_1 = 9$ observations and the outcome of interest Y is observed only over $n_2 = 6$ observations. The ML estimator of the mean of $\mathbf{Y}$ is $\hat{\mu}_y = 27.047$, while the average of the sequential predictors $\hat{m}_0$ is 31.383, slightly closer to 29.3 the true average. (data source Draper and Smith 1981)

The same idea of factorization can be applied when $\underline{\mathbf{X}}^T$ includes non normal variables. For contingency table type data multinomial ML estimates can be found in manner analogous to the MVN case. For the case $\underline{\mathbf{X}}^T$ includes a mix of categorical and continuous data a general location model can be used where the conditional distribution of the continuous variables given the categorical are MVN with the marginal distribution of the categorical variables are multinomial (see Little and Rubin, 2002, section 14.2.3).

Some comments on this approach from a practitioner's perspective: First, in a survey context the number of covariates $\underline{\mathbf{X}}^T$ is normally large, easily many dozens, and so modelling $\underline{\mathbf{X}}^T$ even through the simplifying process of factorization leaves us with the task of multiple multivariate model fitting exercises. Note as well that this is the case even when only a single survey variable $\mathbf{Xf}^T$ is of interest. This modelling burden both hinders the regular use in a commercial environment and also increases the chance of model misspecification.

A second point is interpretation of the results to non-technical clients. Take for example a $T = 2$ case with two categorical covariates X^{0p}, X^{1q} with $p = 1, ..., P$; $q = 1, ..., Q$ levels and a continuous survey variable $\mathbf{Y}$. Assuming an additive model $Y_{pqk} = \beta + \beta_p^0 x_k^{0p} + \beta_q^1 x_k^{1q} + \varepsilon_{kpq}$; $\varepsilon_{kpq} \sim (0, \sigma^2)$ an MLE of the population average is

$$\begin{aligned}
\hat{\overline{\mathbf{Y}}}_{s^0} &= \hat{\beta} + \sum_p \frac{N_p}{N} \hat{\beta}_p^0 + \sum_p \sum_q \frac{N_p}{N} \frac{n_{1pq}}{n_{1p\cdot}} \hat{\beta}_q^1 \\
&= \sum_p \sum_q \hat{W}_{pq} \hat{y}_{pq} \qquad (3.18)
\end{aligned}$$

where the weight $\hat{W}_{pq} = \frac{N_p}{N} \frac{n_{1pq}}{n_{1p\cdot}}$ is the MLE of the proportion of the population with $(\mathbf{X}^{0p} = 1, \mathbf{X}^{1q} = 1)$, with n_{1pq} indicating the number of units $X_k^{0p} = 1$ and $X_k^{1q} = 1$ over s^1, and $n_{1p\cdot} = \sum_{q=1}^{Q} n_{1pq}$. See Vartivarian and Little (2003) for similar estimator in general missing data context.

How do we interpret $\hat{W}_{pq}$? Say X^0 indicates gender while X^1 is an Internet behavioural indicator such as hours spent on news websites. Then W_{pq} is the expected population size of units of gender p and Web news engagement q had all units been surveyed. However, this implies all of s^0 are Online. Of course under the assumptions $\hat{W}_{pq}$ indeed estimates this by implicitly imputing values X^1 to non internet population set, however, avoiding such explaining would be beneficial.

The reason for this friction is the need to define a model for distribution $f(\mathbf{y}|\underline{\mathbf{X}}^{T-1})$ over the entire s^0 which includes variables $\underline{\mathbf{X}}^{T-1}$ associated with the selection process and creation of lower phase populations. In many $\underline{\mathbf{x}}^{T-1}$ include variables not well defined over the entire population. This is not a statistical problem, but in practice I believe it may confuses both the practitioners building the estimation models and the end users which create the demand for this type of work. .

An alternative approach based on the $g-$computational algorithm (discussed in chapter 2) conveniently sidesteps the need of modelling the entire joint distribution of $\underline{\mathbf{X}}^{\mathbf{T}}$over s^0. It reduces the complexity of m-type estimation by focusing estimation on the outcome variables $\mathbf{X}^T = \mathbf{Y}$ only. The idea is simple and is best described by rearranging our estimand of interest as a series of conditional expectations. Under (3.16) the estimand $E(\overline{\mathbf{Y}}_{s^0})$ can be given by

$$\begin{aligned}
E\left(\overline{\mathbf{X}}_{s^0}^T\right) &= N^{-1}\mathbf{1}_{s^0}^{\prime} EE\left\{E\left[\cdots E\left(\mathbf{Y}_{s^0}|\underline{\mathbf{x}}_{s^0}^{T-1}\right)\cdots|\underline{\mathbf{x}}_{s^0}^{1}\right]|\mathbf{x}^0\right\} \\
&= N^{-1}\mathbf{1}_{s^0}^{\prime} EE\left\{E\left[\cdots E\left(\mathbf{X}_{s^T}^T|\underline{\mathbf{x}}_{s^{T-1}}^{T-1}, \underline{s}^T\right)\cdots|\underline{\mathbf{x}}^{1}, \underline{s}^2\right]|\mathbf{x}^0, s^1\right\} (3.19)
\end{aligned}$$

where the first equation is due to the law of total expectations and the second equation is directly inferred under (3.16).

Now, denote $m_T = \mathbf{X}^T$ and by letting $m_{T-1} = E(m_T|\underline{\mathbf{x}}^{T-1}, s^{T-1})$ and $m_{T-2} = E(m_{T-1}|\underline{\mathbf{x}}^{T-2}, s^{T-2})$ and so on ..., $m_{t-1} = E(m_t|\underline{\mathbf{x}}^{t-1}, s^{t-1})$ ..., $m_0 = E(m_1|\mathbf{x}^0)$ the expected population average is equal to the expectation of m_0 that is $E(\mathbf{Y}) = E(m_0)$ where m_0 is a function of $\mathbf{x}^0_{s^0}$, the covariates associated with the finite population of interest and which are assumed fully observed and clearly defined over s^0.

An estimation strategy suggested from the above first specifies regression models $m_{t-1}(\underline{\mathbf{x}}^{t-1}; \boldsymbol{\beta}_{t-1}) = \underline{\mathbf{x}}_k^{t-1'}\boldsymbol{\beta}_{t-1}$ for the regression functions $E(m_t|\underline{\mathbf{x}}^{t-1})$ over s^{t-1} for $t = T, ..., 1$ and then estimates the regression parameters $\boldsymbol{\beta}_{t-1}$ from the observed data s^t. This latter task will carry out recursively based on the observation that (i) by definition $m_{t-1} = E(m_t|\underline{\mathbf{x}}^{t-1}_{s^{t-1}})$, (ii) under the sequential ignorability assumption $E(m_t|\underline{\mathbf{x}}^{t-1}_{s^{t-1}}) = E(m_t|\underline{\mathbf{x}}^{t-1}_{s^{t-1}}, \underline{s}^t)$ and (iii) that m_{t-1} is a function of $\underline{\mathbf{x}}^{t-1}_{s^{t-1}}$ which is entirely observed. Note that we can generalize this approach so that $m_{t-1}(\underline{\mathbf{x}}^{t-1}; \boldsymbol{\beta}_{t-1}) = \boldsymbol{\psi}(\underline{\mathbf{x}}_k^{t-1'}\boldsymbol{\beta}_{t-1})$ with $\boldsymbol{\psi}^{-1}$ a known link function. If not stated otherwise, my discussion will focus on linear regression where $\boldsymbol{\psi}^{-1}$ is the identity link function.

More formally the sequential estimation algorithm:

1. Set $Y_k = m_{Tk}$ for all observed values of y_k ; $\quad k \in s^T$,

2. Recursively for $t = T, ..., 1$

   (a) Specify the linear regression model m_{t-1} for the conditional expectation of m_t on $\underline{\mathbf{x}}^{t-1}_{s^{t-1}}$

   $$m_{t-1}(\underline{\mathbf{x}}^{t-1}, s^{t-1}) : \begin{cases} E(m_{tk}|\underline{\mathbf{x}}^{t-1}, s^{t-1}) & = \underline{\mathbf{x}}_k^{t-1'}\boldsymbol{\beta}_{t-1} \\ V(m_{tk}|\underline{\mathbf{x}}^{t-1}, s^{t-1}) & = \sigma^2_{t-1k}. \end{cases}$$

   (b) Calculate consistent estimates $\hat{\boldsymbol{\beta}}_{t-1}$ of the regression coefficients $\boldsymbol{\beta}_{t-1}$ over the observed set $\mathbf{s}^t$

   (c) Fit predictions $\hat{m}_{t-1k}(\underline{\mathbf{X}}^{t-1}, s^{t-1}) = \underline{\mathbf{x}}_k^{t-1'}\hat{\boldsymbol{\beta}}_{t-1}$ over all of s^{t-1} using observed values of $\underline{\mathbf{x}}^{t-1}_{s^{t-1}}$.

   (d) Return to (a) now with $\hat{m}_{t-1k}(\underline{\mathbf{X}}^{t-1}, s^{t-1})$ replacing m_{t-1k}.

3. The final step, for $t = 1$ gives predictions $\hat{m}_{0k}(\mathbf{x}^0) = \mathbf{x}_k^{0'}\hat{\boldsymbol{\beta}}_0$ calculated over

the entire population s^0. Then we can estimate the population average by

$$\begin{aligned}\hat{\overline{\mathbf{Y}}}_m &= N^{-1}\sum_{s^0}\mathbf{X}_k^{0'}\hat{\boldsymbol{\beta}}_0 \\ &= N^{-1}\sum_{s^0}\hat{m}_{0k} \ .\end{aligned}$$

To demonstrate the estimation strategy, and compare to the likelihood approach I return to a $T=2$ example applied to our empirical data in table 3.2. Start by defining $m_2 = \mathbf{Y}$, $m_1 = E(m_2|\underline{\mathbf{x}}^1, s^1)$ and $m_0 = E(m_1|\mathbf{x}^0)$; Under the conditional independence assumptions

$$\begin{aligned}E(\overline{\mathbf{Y}}_{s^0}) &= N^{-1}\mathbf{1}_{s^0}^{'}E\left\{E\left[E(m_2|\underline{\mathbf{x}}^1, s^2)\right]|\mathbf{x}^0, s^1\right\} \\ &= N^{-1}\mathbf{1}_{s^0}^{'}EE(m_1|\mathbf{x}^0, s^1) \\ &= N^{-1}\mathbf{1}_{s^0}^{'}E(m_{0s^0}). \qquad (3.20)\end{aligned}$$

Let $\mathbf{W}_{s^t}^{t-1}$ denote a diagonal matrix with typical element $\{\sigma_{t-1k}^{-2}\}$ for $k \in s^t$. Estimation starts by fitting model $m_1(\underline{\mathbf{x}}^1): \ \ E(\mathbf{Y}|\underline{\mathbf{X}}^1, s^1; \boldsymbol{\beta}_1) = \underline{\mathbf{X}}_k^{1'}\boldsymbol{\beta}_1$ over s^2 by regressing $\mathbf{Y}_{s^2}$ on $\underline{\mathbf{X}}_{s^2}^1$ giving the coefficient estimator $\hat{\boldsymbol{\beta}}_1 = (\underline{\mathbf{X}}_{s^2}^{1'}\mathbf{W}_{s^2}^1\underline{\mathbf{X}}_{s^2}^1)^{-1}\underline{\mathbf{X}}_{s^2}^{1'}\mathbf{W}_{s^2}^1\mathbf{Y}_{s^2}$ and calculate $\hat{m}_{1k} = \underline{\mathbf{X}}_k^{1'}\hat{\boldsymbol{\beta}}_1$ over higher set s^1. Then fit $m_0(\mathbf{x}^0): \ \ E(m_1|\mathbf{x}^0; \boldsymbol{\beta}_0)$ over s^1 by regressing $\hat{m}_{1k}$ on $\mathbf{X}_k^0$ giving the coefficient estimator $\hat{\boldsymbol{\beta}}_0$ equal to $(\mathbf{X}_{s^1}^{0'}\mathbf{W}_{s^1}^0\mathbf{X}_{s^1}^0)^{-1}\mathbf{X}_{s^1}^{0'}\mathbf{W}_{s^1}^0\hat{\mathbf{Y}}_{m_1s^1}$ and calculate $\hat{m}_{0k} = \mathbf{X}_k^{0'}\hat{\boldsymbol{\beta}}_0$ over population s^0 as $\mathbf{X}^0$ is observed for all units. These are the sequential predictors stated in the right hand side column of table 3.2. The population average estimator is then

$$\begin{aligned}\hat{\overline{\mathbf{Y}}}_m &= N^{-1}\mathbf{1}_{s^0}^{'}\mathbf{X}_{s^0}^0\hat{\boldsymbol{\beta}}_0 \\ &= N^{-1}\mathbf{1}_{s^0}^{'}\mathbf{X}_{s^0}^0(\mathbf{X}_{s^1}^{0'}\mathbf{W}_{s^1}^0\mathbf{X}_{s^1}^0)^{-1}\mathbf{X}_{s^1}^{0'}\mathbf{W}_{s^1}^0\hat{\mathbf{m}}_{1s^1} \\ &= N^{-1}\mathbf{1}_{s^0}^{'}\mathbf{H}_{s^0s^1}^0\mathbf{H}_{s^1s^2}^1\mathbf{Y}_{s^2}\end{aligned}$$

where $\mathbf{H}_{s^1s^2}^1 = \underline{\mathbf{X}}_{s^1}^1(\underline{\mathbf{X}}_{s^2}^{1'}\mathbf{W}_{s^2}^1\underline{\mathbf{X}}_{s^2}^1)^{-1}\underline{\mathbf{X}}_{s^2}^{1'}\mathbf{W}_{s^2}^1$ is parallel to the hat matrix in normal linear regression. The estimated population average is $N^{-1}\sum_{s^0}\hat{m}_{0k} = 31.383$, slightly closer to the true population average 27.047 than the MLE which was found to be 29.231. Unbiasedness of the estimator can be shown by

$$\begin{aligned}E(\hat{\overline{\mathbf{Y}}}_m) &= N^{-1}\mathbf{1}_{s^0}^{'}E\left\{E\left[\underline{\mathbf{H}}_{s^1s^2}^1E(m_{2s^2}|\underline{\mathbf{x}}_{s^1}^1, s^2)\right]|\mathbf{x}^0, s^1\right\} \\ &= N^{-1}\mathbf{1}_{s^0}^{'}EE(\mathbf{H}_{s^0s^1}\underline{\mathbf{X}}_{s^1}^1\boldsymbol{\beta}_1|\mathbf{x}^0, s^1) \\ &= N^{-1}\mathbf{1}_{s^0}^{'}E(\mathbf{H}_{s^0s^1}E(m_{1s^1}|\mathbf{x}^0, s^1)) \\ &= N^{-1}\mathbf{1}_{s^0}^{'}E(\mathbf{H}_{s^0s^1}m_{0s^1}) = N^{-1}\mathbf{1}_{s^0}^{'}E(m_{0s^0})\end{aligned}$$

which as shown in (3.20) is equal to $E(\overline{\mathbf{Y}}_{s^0})$, our estimand of interest. The key to solving the above equations is to note that $\mathbf{H}_{s^{t-1}s^t}^{t-1}\underline{\mathbf{X}}_{s^t}^{t-1}\boldsymbol{\beta}_{t-1} = \underline{\mathbf{X}}_{s^{t-1}}^{t-1}\boldsymbol{\beta}_{t-1}$ for any $t = 1, ..., T$.

Note that by letting $\underline{h}_k^1 = \mathbf{1}'_{s^0}\mathbf{X}^0_{s^0}(\mathbf{X}^{0'}_{s^1}\mathbf{W}^0_{s^1}\mathbf{X}^0_{s^1})^{-1}\mathbf{X}^{0'}_{s^1}\mathbf{W}^0_{s^1}\underline{\mathbf{X}}^1_{s^1}(\underline{\mathbf{X}}^{1'}_{s^2}\mathbf{W}^1_{s^2}\underline{\mathbf{X}}^1_{s^2})^{-1}\underline{\mathbf{X}}^1_k\mathbf{W}^1_k$ the estimator can take the shape of a weighted sample based estimator

$$\hat{\bar{\mathbf{Y}}}_m = N^{-1}\sum_{k\in s^2}\underline{h}_k^1 Y_k. \tag{3.21}$$

This is essentially a sequential ($m-$model type) application of the 'g-computational algorithm' approach advocated by Robins (1986, 1987) developed in the context of causal inference in longitudinal studies. Robins (1999) discusses theoretical results, namely that an estimator constructed by this algorithm is consistent and asymptotic normal for the $E(\mathbf{Y})$ under the union of parametric models $m_{t-1}(\underline{\mathbf{x}}^{t-1};\boldsymbol{\beta}_{t-1})$ for $t = 1,..,T$. This algorithm and the theoretical properties can be expanded to any GLM link function Robins (1999).

Some comments on benefits of this approach are worth noting. First, similar to likelihood estimation under a judicious definition of the selection sets we can include covariates of type $\mathbf{X}^1, \mathbf{X}^2, .., \mathbf{X}^{T-1}$ which could not be used in the single phase framework. Also, there are several characteristics which make this method arguably better than the likelihood approach. Clearly for a single survey measurement of interest, the recursive estimation has a lower modelling burden lending to a higher chance of correct specification. The practitioner here can also test more effectively model misspecification and avoid issues such as extrapolation by inspecting separately for each model m_t over s^{t-1} the distribution of $\underline{\mathbf{X}}_{s^{t-1}-\bar{s}^t}$ and $\underline{\mathbf{X}}_{s^t}$ which for each t will have a less imbalanced common support (that is a very low incidence level for certain combinations of covariates) than we would expect in a single phase framework. Furthermore, by averaging out covariates associated with the specific sub population which are clearly defined for each phase $t = T,..,1$ we are avoiding interpretation problems such as in (3.18) making it is easier to integrate expert knowledge and contextual understanding.

I conclude this section by generalizing (3.21), a formulation fundamental in my later discussion of panel estimation using a random reference survey sample and purposive sampling designs. For T sequential framework

$$\begin{aligned}
\hat{\bar{\mathbf{Y}}}_m &= \sum_{s^0}\hat{m}_{0k} \\
&= N^{-1}\mathbf{1}'_{s^0}\mathbf{X}^0_{s^0}\hat{\boldsymbol{\beta}}_0 \\
&= N^{-1}\mathbf{1}'_{s^0}\mathbf{X}^0_{s^0}(\sum_{s^1}\mathbf{X}^0_k\mathbf{X}^{0'}_k/\sigma^2_{0k})^{-1}\sum_{s^1}\mathbf{X}^0_k\hat{m}_{1k}/\sigma^2_{0k} \\
&= N^{-1}\mathbf{1}'_{s^0}\mathbf{X}^0_{s^0}(\mathbf{X}^{0'}_{s^1}\mathbf{W}^0_{s^1}\mathbf{X}^0_{s^1})^{-1}\mathbf{X}^{0'}_{s^1}\mathbf{W}^0_{s^1}\hat{\mathbf{m}}_{1s^1}
\end{aligned}$$

where $\mathbf{W}^0_{s^1} = diag(\sigma^2_{0k})^{-1}$ is a $n_1 \times n_1$ diagonal matrix of model weights used in the estimation equations of m_o , and $\hat{\mathbf{m}}_{1s^1}$ is the vector of fitted values $\hat{m}_{1k} = \underline{\mathbf{X}}^{1'}_k\hat{\boldsymbol{\beta}}_1$ over the set s^1.

Now, let $\mathbf{H}^{t-1}_{s^{t-1}s^t} = \underline{\mathbf{X}}^{t-1}_{s^{t-1}}\mathbf{A}^{t-1}_{s^t}\underline{\mathbf{X}}^{t-1'}_{s^t}\mathbf{W}^{t-1}_{s^t}$ be a t-phase sequential framework analogue to the linear regression 'hat' matrix, where $\mathbf{A}^{t-1}_{s^t} = (\underline{\mathbf{X}}^{t-1'}_{s^t}\mathbf{W}^{t-1}_{s^t}\underline{\mathbf{X}}^{t-1}_{s^t})^{-1}$ with $\mathbf{W}^{t-1}_{s^t}{=}diag(\sigma^2_{t-1k})^{-1}$ a $n_t \times n_t$ diagonal matrix of model weights used in the estimation equations of model m_{t-1}. Then a T phase outcome model

$$\begin{aligned}
\bar{\bar{\hat{\mathbf{Y}}}}_m &= N^{-1}\mathbf{1}'_{s^0}\mathbf{H}^0_{s^0s^1}\hat{\mathbf{m}}_{1s^1} \\
&= N^{-1}\mathbf{1}'_{s^0}\mathbf{H}^0_{s^0s^1}\cdots\mathbf{H}^{t-1}_{s^{t-1}s^t}\cdot\cdot\mathbf{H}^{T-1}_{s^{T-1}s^T}\mathbf{Y}_{s^T} \\
&= N^{-1}\mathbf{1}'_{s^0}\underline{\mathbf{H}}^{T-1}_{s^{T-1}s^T}\mathbf{Y}_{s^T} = N^{-1}\sum_{\underline{s}^T}\underline{h}^{T-1}_k\mathbf{Y}_k
\end{aligned} \tag{3.22}$$

where $\underline{h}^{T-1}_k = \mathbf{1}'_{s^0}\underline{\mathbf{H}}^{T-2}_{s^{T-2}s^{T-1}}\mathbf{X}^{T-1}_{s^{T-1}}\mathbf{A}^{T-1}_{s^T}\mathbf{X}^{T-1}_k\mathbf{W}^{T-1}_k$ which achieves our aim of representing $\bar{\bar{\hat{\mathbf{Y}}}}_m$ as a linear function of $\mathbf{y}_{s^T}$ the observed set of outcome variables. It is then simple to show unbiasedness by plugging (3.22) into equation (3.19).

3.5.2 Selection-Model Based Estimation

Now I turn to a sequential version of the π-estimator. As discussed in chapter 2 the π-estimator is by far the most popular approach in practice, either in its role at the centre of survey sampling theory (Horvitz and Thompson, 1952) or in observational studies where Rosenbaum and Rubin (1983) introduced the balancing properties of a consistently estimated unit selection probability. The most important drivers of this popularity is its simplicity and universality - a single simple to estimate model can be applied to any survey measurement of interest.

Start again with the observed joint distribution. We leave the multivariate covariate distribution $f(\underline{\mathbf{x}}^T_{s^T}, \mathbf{w})$ completely unspecified and assume that selection is independent of unobserved covariates or values, so that

$$f(\underline{d}^T_{s^T}) \;=\; \int f(\underline{\mathbf{x}}^T_{s^T}, \mathbf{w})d\mathbf{w} \times \Pi^T_{t=1}p(s^t|\underline{s}^{t-1}, \underline{\mathbf{x}}^{t-1}_{s^{t-1}}). \tag{3.23}$$

As in chapter 2, for simplicity we assume a parametric approach and model the selection process by logistic regression, and assume as well that selection is individualistic and probabilistic. Under these assumptions, which can be relaxed,

$$\begin{aligned}
\Pi^T_{t=1}p(s^t|\underline{s}^{t-1}, \underline{\mathbf{x}}^{t-1}_{s^{t-1}}) \;=\; & \\
& \Pi^T_{t=1}\Pi^{n_{s^{t-1}}}_{k=1}p(s^t_k = 1|\underline{\mathbf{x}}^{t-1}_k, \underline{s}^{t-1}_k)^{s^t_k}\left(1 - p(s^t_k = 1|\underline{\mathbf{x}}^{t-1}_k, \underline{s}^{t-1}_k)\right)^{1-s^t_k}
\end{aligned} \tag{3.24}$$

where $p(s^t_k = 1|\underline{\mathbf{x}}^{t-1}_k, \underline{s}^{t-1}_k)$ denotes the selection probability of unit $k \in s^{t-1}$ into subset s^t. This selection process is equivalent to a multiphase Poisson sample design which are common practice (see for example Kott and Fetter 1999; Lohr 2009) in survey sampling theory. Next assume that

$$\pi_t(\underline{\mathbf{x}}^{t-1}, s^{t-1}) \;:\; \begin{cases} p(s^t_k = 1|\underline{\mathbf{x}}^{t-1}_k, \underline{s}^{t-1}_k) & = \pi_t(\underline{\mathbf{x}}^{t-1}_k, \underline{s}^{t-1}_k; \boldsymbol{\alpha}^{t-1}) \\ p(s^t_k = 1|\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1}) \geq 0 & \text{for all } t = 1, ..., T \text{ and } k \in s^0 \end{cases}$$

where $\pi_t(\underline{\mathbf{x}}_k^{t-1};\boldsymbol{\alpha}^{t-1}) = exp(\underline{\mathbf{x}}_k^{t-1'}\boldsymbol{\alpha}^{t-1})/1 + exp(\underline{\mathbf{x}}_k^{t-1'}\boldsymbol{\alpha}^{t-1})$ are specified by a logistic model over the sub population s^{t-1}.

Under (3.23) each set of model coefficients $\underline{\boldsymbol{\alpha}}^{T-1}$ can be estimated consistently over the appropriate sub population. The population average is then estimated with the sequential π- estimator

$$\hat{\overline{\mathbf{Y}}}_{\pi} = N^{-1}\sum_{s^T} Y_k \hat{\underline{\pi}}_{Tk}^{-1} \tag{3.25}$$

where $\underline{\pi}_T(\underline{\mathbf{x}}_k^{T-1}, \underline{s}_k^{T-1}; \hat{\boldsymbol{\alpha}}^{T-1}) = \pi_1(\mathbf{x}_k^0; \hat{\boldsymbol{\alpha}}^0) \times .... \times \pi_T(\underline{\mathbf{x}}_k^{T-1}, \underline{s}_k^{T-1}; \hat{\boldsymbol{\alpha}}^{T-1})$.

To see that the π-estimator is approximately unbiased, assume $\pi_t(\underline{\mathbf{x}}^{t-1}, \underline{s}_k^{T-1}; \hat{\boldsymbol{\alpha}}^{T-1})$ are estimated consistently so that $\pi_t(\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1}; \hat{\boldsymbol{\alpha}}^{t-1})\xrightarrow[n_t \to \infty]{}\pi_t(\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1}; \boldsymbol{\alpha}^{t-1})$ for all $t = 1, .., T$; $k \in s^0$. If as well models $\underline{\pi}_T(\cdot; \boldsymbol{\alpha}^{t-1})$ are correctly specified, then by rearranging the expectation of our estimand as a series of conditional expectation

$$N^{-1}\sum_{s^0} EE\left\{E\left[\cdots E\left(\underline{S}_k^T Y_k \underline{\pi}_{Tk}^{-1}(\underline{\mathbf{x}}^{T-1}, \underline{s}^{T-1}; \boldsymbol{\alpha}^{t-1})|\underline{\mathbf{x}}_{s^{T-1}}^{T-1}, \underline{s}^T\right)\cdots|\underline{\mathbf{x}}^1, \underline{s}^2\right]|\mathbf{x}^0, s^1\right\}$$

is equal to $E(\overline{\mathbf{Y}}_{s^0})$ as under the assumptions $E(S_k^t|\underline{\mathbf{x}}^{t-1}, \underline{s}_k^{t-1}, y_k; \boldsymbol{\alpha}^{t-1}) = Pr(S_k^t = 1|\underline{\mathbf{x}}^{t-1})$ for each $t = 1, ..., T$.

As in the one phase formulation, the role of the unit selection probability in estimation can be understood by its strength as balancing score. In the sequential framework balancing properties of $\pi_{tk}(\underline{\mathbf{x}}_{s^{t-1}}^{t-1}, \underline{s}^{t-1})$ follows from the statement

$$\text{if } \; Y_k \perp S_k^t|\underline{\mathbf{x}}_k^{t-1}, \underline{s}_k^{t-1} \text{ then } Y_k \perp S_k^t|\pi_t(\underline{\mathbf{x}}_k^{t-1}, \underline{s}_k^{t-1}), \underline{s}_k^{t-1} \; \text{ as well.} \tag{3.26}$$

For (3.26) to hold, it is sufficient to show that

$$p(s^t|\mathbf{y}, \pi_t(\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1}), \underline{s}^{t-1}) = p(s^t|\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1}).$$

which holds for any $t = 1, ..., T$

$$\begin{aligned} p(s^t|\mathbf{y}, \pi_t(\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1}), \underline{s}^{t-1}) &= E\left(E(S^t|\mathbf{y}, \pi_t(\cdot), \underline{s}^{t-1}, \underline{\mathbf{x}}^{t-1})|\mathbf{y}, \pi_t(\cdot), \underline{s}^{t-1}\right) \\ &= E\left(E(S^t|\underline{s}^{t-1}, \underline{\mathbf{x}}^{t-1})|\mathbf{y}, \pi_t(\cdot), \underline{s}^{t-1}\right) \\ &= E\left(\pi_t(\cdot)|\mathbf{y}, \pi_t(\cdot), \underline{s}^{t-1}\right) = \pi_t(\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1}). \end{aligned}$$

A similar proof is given in Lechner and Miquel (2005); Lechner (2009); Lechner and Miquel (2010) in the context of potential outcomes in two phase treatment effect studies of labour market policy changes. Robins (1986) uses sequential estimating equations to develop statistical properties of the π-estimator which take into account the model coefficient estimation. As in the one phase framework, the sequential π-estimator will be inferior to the $m-$estimator in terms of efficiency; the reason is identical and rests on fact that $\hat{\overline{\mathbf{Y}}}_{\pi}$ is defined only over the final observed set s^T while the m-estimator is defined over the entire population s^0 (Carpenter et al., 2006).

Fundamentally, estimation relies on correct specification of all $t = 1, ...T$ models $\pi_t(\underline{\mathbf{x}}^{t-1})$. This may be seen as an increase in modelling burden compared to the single phase π-estimator. However I would argue that given our discussion in chapter 2, it is questionable that the panel survey process can be reduced to a single model. Compared to (3.24) a single phase π-estimator assumes that

$$\begin{aligned}\Pi_{t=1}^{T} p(s^t|\underline{s}^{t-1}, \underline{\mathbf{x}}_{s^{t-1}}^{t-1}) &= \Pi_{t=1}^{T} p(s^t|\underline{s}^{t-1}\mathbf{x}^0; \boldsymbol{\alpha}) \\ &= \Pi_{k=1}^{N} \pi(\mathbf{x}_k^0; \boldsymbol{\alpha})^{s_k^T} (1 - \pi(\mathbf{x}_k^0; \boldsymbol{\alpha}))^{1-s_k^T}\end{aligned}$$

a single model capturing the product of all T phases using only a limited pool of $\mathbf{X}^0$ type of potential variables. This constraint on directly observed covariates $\mathbf{X}^1, ..., \mathbf{X}^{T-1}$ and the use of a single model is exactly the reason for practitioners resorting to made up metrics, so called 'Webographics', that supposedly (Schonlau et al., 2007; Terhanian et al., 2000) are uniquely informative in explaining the differences between the panel and general population. If I take a model driven perspective, the sequential framework, by allowing more data and contextual input to be used, should improve estimation.

Similar to previous comments, and even more so here, the choice of defining the Web panel survey process into an Internet connection/usage phase, a panel selection phase and a within panel survey sampling phase is beneficial. Separately, these selection phases are much better understood.

For the first phase process we can use Information and communications technology (ICT) and Internet take up models which have been researched extensively Robertson et al. (2007); Kridel et al. (2002). Conditional on Web usage the analyst may model self-selection into the panel $p(s^2|s^1, \cdot)$ using crucial Internet behaviour information and general respondent participation models (e.g. Groves and Couper 1998; Groves et al. 2004, 1992) while avoiding the common support constraint. The third process to model, survey sampling mechanism $p(s^3|s^2, \cdot)$, which is a mixture of known design with an increasingly large non response element (Baker and Brick, 2013) can be modelled now separately over time with the panel management team's ongoing tracking of panelists response record.

3.5.3 Selection and Outcome-Model Based Estimation: A Double Robust Strategy

In the preceding sections we suggested estimation strategies relying on modelling of either the sequential selection process $p(s^t|\underline{s}^{t-1}, \underline{\mathbf{x}}^{t-1}, \mathbf{w})$ or the sequence of covariate distributions $f(\mathbf{x}^t_{s^t}|\underline{\mathbf{X}}^{t-1}_{s^{t-1}}, \mathbf{w})$. In turn validity of the estimators relies on correct specification of the set of π-models or m-models respectively. An alternative is to build an estimator which takes into account both the π set of models and m set of models. I introduce one such estimator, a πm-estimator, under a general T phase selection framework.

Under the assumptions separately taken in the m and π approaches, first denote $m_{Tk} = \mathbf{x}^T_k$ and let $m_{t-1k} = E(m_{tk}|\underline{\mathbf{x}}^{t-1}_k, s^{t-1})$ for $t = 1, ..., T$ and any $k \in s^0$. We assume then that

$$\begin{cases} E(S^t_k|\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1}) & = \pi_t(\underline{\mathbf{x}}^{t-1}_k; \boldsymbol{\alpha}^{t-1}) > 0 \\ E(m_{tk}|\underline{\mathbf{x}}^{t-1}, s^{t-1}; \boldsymbol{\beta}_{t-1}) & = \psi[m_{t-1}(\underline{\mathbf{x}}^{t-1}_k; \boldsymbol{\beta}_{t-1})] \end{cases} \tag{3.27}$$

where ψ^{-1} is the link function of a given GLM, $m_{t-1}(\underline{\mathbf{x}}^{t-1}_k; \boldsymbol{\beta}_{t-1})$ is a known regression function with unknown parameter $\boldsymbol{\beta}_{t-1}$ and $\pi_t(\underline{\mathbf{x}}^{t-1}_k; \boldsymbol{\alpha}^{t-1})$ are specified by a logistic model. Under (3.27) constructing the πm follows these steps

1. Independently for each $t = 1, .., T$
   (a) Specify logistic regression models $\pi_t(\underline{\mathbf{x}}^{t-1})$ of selection S^t_k's over the population subset s^{t-1}.
   (b) Calculate consistent estimators $\hat{\boldsymbol{\alpha}}^{t-1}$ of model coefficients $\boldsymbol{\alpha}^{t-1}$ and predict selection probabilities $\pi_t(\underline{\mathbf{x}}^{t-1}_k; \hat{\boldsymbol{\alpha}}^{t-1}) = \hat{\pi}_{tk}$ for all $k \in s^t$.

2. Recursively for $t = T, ..., 1$
   (a) Specify a parametric regression model $\psi[m_{t-1}(\underline{\mathbf{x}}^{t-1}_k; \boldsymbol{\beta}_{t-1})]$ for the conditional expectation of m_t on $\underline{\mathbf{X}}^{t-1}_{s^{t-1}}$ over population subset s^{t-1} .
   (b) Fit models m_{t-1} over s^t and compute consistent estimators $\hat{\boldsymbol{\beta}}_{\pi t-1}$of $\boldsymbol{\beta}_{t-1}$ by weighting the estimation equations on weights $\hat{\pi}_{tk} \cdot \sigma^2_{t-1k}$ the product of unit selection weight $\hat{\pi}_{tk}(\underline{\mathbf{x}}^{t-1}, s^{t-1})$ and model weights σ^2_{t-1k}.
   (c) Predict values m_{t-1k} for all $k \in s^{t-1}$ using the observed values $\underline{\mathbf{x}}^{t-1}$ and $\hat{\boldsymbol{\beta}}_{\pi t-1}$ the coefficient estimators, that is $\hat{m}_{\pi t-1k}(\underline{\mathbf{x}}^{t-1}) = \underline{\mathbf{x}}^{t-1\prime}_k \hat{\boldsymbol{\beta}}_{\pi t-1}$.

3. The final step for $t = 1$ gives predictions $\hat{m}_{\pi 0k}(\mathbf{X}^0) = \psi(\mathbf{X}^{0\prime}_k \hat{\boldsymbol{\beta}}_{\pi 0})$, which allows us to estimate the population average by

   $$\hat{\bar{\mathbf{Y}}}_{\pi m} \quad = \quad N^{-1} \sum_{s^0} \hat{m}_{\pi 0k}$$

   as the sum of $\mathbf{X}^0_{s^0}$ is assumed known over the population of interest s^0.

It is convenient to describe some of the properties of this πm estimator when ψ^{-1} is the identity link function. This format will also guide us in the next chapter when we consider the application to real panel settings.

For any $t = 1, .., T$ the coefficients of model $m_{t-1}(\underline{\mathbf{x}}^{t-1}, s^{t-1}; \boldsymbol{\beta}_{\pi t-1})$ are estimated by

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\pi t-1} &= (\underline{\mathbf{X}}_{s^t}^{t-1'} \mathbf{W}_{s^t}^{t-1} \hat{\boldsymbol{\pi}}_{ts^{t-1}}^{-1} \underline{\mathbf{X}}_{s^t}^{t-1})^{-1} \underline{\mathbf{X}}_{s^t}^{t-1'} \mathbf{W}_{s^t}^{t-1} \hat{\boldsymbol{\pi}}_{ts^{t-1}}^{-1} \hat{\mathbf{m}}_{ts^t} \\
&= \mathbf{A}_{\pi s^t}^{t-1} \underline{\mathbf{X}}_{s^t}^{t-1'} \mathbf{W}_{s^t}^{t-1} \hat{\boldsymbol{\pi}}_{ts^{t-1}}^{-1} \hat{\mathbf{m}}_{ts^t}
\end{aligned}$$

where $\underline{\mathbf{X}}_{s^t}^{t-1}$ is the $n_t \times \underline{p}^{t-1}$ matrix of covariates $\underline{\mathbf{X}}^{t-1}$ over s^t ; $\mathbf{W}_{s^t}^{t-1}$ is the $n_t \times n_t$ diagonal matrix of model weights with typical component $W_{t-1k}^{-1} = \sigma_{t-1k}^2$; $\hat{\boldsymbol{\pi}}_{ts^{t-1}}^{-1} = diag\{\hat{\pi}_t^{-1}(\underline{\mathbf{X}}_k^{t-1}; \hat{\boldsymbol{\alpha}}_{t-1})\}$ an $n_t \times n_t$ matrix , and $\hat{\mathbf{m}}_{ts^t}$ is the n_t vector of unit predictions $\hat{m}_{tk}$. This allows us to describe $\hat{\bar{\mathbf{Y}}}_{\pi m}$ as

$$\begin{aligned}
\hat{\bar{\mathbf{Y}}}_{\pi m} &= N^{-1} \mathbf{1}_{s^0}' \underline{\mathbf{H}}_{\pi s^{T-1} s^T}^{T-1} \mathbf{Y}_{s^T} \\
&= N^{-1} \sum_{\underline{s}^T} \underline{h}_{\pi_T k}^{T-1} Y_k \qquad (3.28)
\end{aligned}$$

where $\underline{\mathbf{H}}_{\pi s^{T-1} s^T}^{T-1} = \mathbf{H}_{\pi s^0 s^1}^{0} \cdots \mathbf{H}_{\pi s^{t-1} s^t}^{t-1} \cdot \cdot \mathbf{H}_{\pi s^{T-1} s^T}^{T-1}$ the product of the π-weighted hat matrices $\mathbf{H}_{\pi s^{t-1} s^t}^{t-1} = \underline{\mathbf{X}}_{s^{t-1}}^{t-1} \mathbf{A}_{\pi s^t}^{t-1} \underline{\mathbf{X}}_{s^t}^{t-1'} \mathbf{W}_{s^t}^{t-1} \hat{\boldsymbol{\pi}}_{ts^{t-1}}^{-1}$ for $t = 1, ..., T$ with $\mathbf{A}_{\pi s^t}^{t-1} = \underline{\mathbf{X}}_{s^{t-1}}^{t-1} (\underline{\mathbf{X}}_{s^t}^{t-1'} \mathbf{W}_{s^t}^{t-1} \hat{\boldsymbol{\pi}}_{ts^{t-1}}^{-1} \underline{\mathbf{X}}_{s^t}^{t-1}) \underline{\mathbf{X}}_{s^t}^{t-1'} \mathbf{W}_{s^t}^{t-1} \hat{\boldsymbol{\pi}}_{ts^{t-1}}^{-1}$ and

$$\begin{aligned}
\underline{h}_{\pi_T k}^{T-1} &= \mathbf{1}_{s^0}' \underline{\mathbf{H}}_{\pi s^{T-2} s^{T-1}}^{T-2} \underline{\mathbf{X}}_{s^{T-1}}^{T-1} \mathbf{A}_{\pi s^T}^{T-1} \underline{\mathbf{X}}_k^{T-1} \mathbf{W}_k^{T-1} \underline{\hat{\pi}}_{Tk}^{-1} \qquad (3.29) \\
&= \mathbf{1}_{s^0}' \underline{\mathbf{H}}_{\pi s^{T-2} s^{T-1}}^{T-2} \underline{\mathbf{X}}_{s^{T-1}}^{T-1} \underline{\mathbf{X}}_k^{T-1} \mathbf{W}_k^{T-1} \underline{\hat{\pi}}_{Tk}^{-1} / \sum_{\underline{s}^T} \underline{\mathbf{X}}_k^{T-1} \mathbf{W}_k^{T-1} \underline{\hat{\pi}}_{Tk}^{-1} \underline{\mathbf{X}}_k^{T-1'}
\end{aligned}$$

is the estimation weight we attach each observed unit k in the panel survey sample.

As in the single phase discussion $\hat{\bar{\mathbf{Y}}}_{\pi m}$ is double robust, that is robust to misspecification of either m or π set of models.

Robustness to misspecification **of m models** can be found by conditioning $E(\hat{\bar{\mathbf{Y}}}_{\pi m})$ on the outcome variable $\mathbf{Y}$ and then sequentially on the pairs $(\mathbf{X}^{t-1}, S^{t-1})$ for all phases $t = 1, ...T$. This leads to an identity which for each phase is a function only of the selection process S^t. To show unbiasedness start with the identity

$$\begin{aligned}
E(\hat{\bar{\mathbf{Y}}}_{\pi m}) &= N^{-1} \mathbf{1}_{s^0}' E \left\{ \cdots EE \left(\underline{\mathbf{H}}_{\pi s^{T-1} s^T}^{T-1} \mathbf{y}_{s^T} | \underline{\mathbf{x}}^{T-1}, \mathbf{y}, \underline{s}^{T-1} \right) \cdots | \mathbf{x}^0, \mathbf{Y}, s^0 \right\} \qquad (3.30) \\
&= N^{-1} E \left\{ \cdots EE \left(\sum_{\underline{s}^T} \underline{h}_{\pi_T k}^{T-1} y_k | \underline{\mathbf{x}}^{T-1}, \mathbf{y}, \underline{s}^{T-1} \right) \cdots | \mathbf{x}^0, \mathbf{Y}, s^0 \right\}
\end{aligned}$$

Note that from (3.29)

$$\sum_{\underline{s}^T} \underline{h}^{T-1}_{\pi_T k} Y_k = \mathbf{1}'_{s^0} \underline{\mathbf{H}}^{T-2}_{\pi s^{T-2} s^{T-1}} \underline{\mathbf{X}}^{T-1}_{s^{T-1}} \frac{\sum_{\underline{s}^T} \underline{\mathbf{X}}^{T-1}_k \mathbf{W}^{T-1}_k \underline{\hat{\pi}}^{-1}_{Tk} Y_k}{\sum_{\underline{s}^T} \underline{\mathbf{X}}^{T-1}_k \mathbf{W}^{T-1}_k \underline{\hat{\pi}}^{-1}_{Tk} \underline{\mathbf{X}}^{T-1'}_k}$$

and so looking only at the inner expectation of (3.31), explicitly indicating the last phase random indicator and taking out $\mathbf{1}'_{s^0} \underline{\mathbf{H}}^{T-2}_{\pi s^{T-2} s^{T-1}} \underline{\mathbf{X}}^{T-1}_{s^{T-1}}$ from the expectation we get

$$\mathbf{1}'_{s^0} \underline{\mathbf{H}}^{T-2}_{\pi s^{T-2} s^{T-1}} \underline{\mathbf{X}}^{T-1}_{s^{T-1}} E\Big(\frac{\sum_{\underline{s}^{T-1}} S^T_k \underline{\mathbf{X}}^{T-1}_k \mathbf{W}^{T-1}_k \underline{\hat{\pi}}^{-1}_{Tk} Y_k}{\sum_{\underline{s}^{T-1}} S^T_k \underline{\mathbf{X}}^{T-1}_k \mathbf{W}^{T-1}_k \underline{\hat{\pi}}^{-1}_{Tk} \underline{\mathbf{X}}^{T-1'}_k} | \underline{\mathbf{x}}^{T-1}, \mathbf{y}, \underline{s}^{T-1}\Big)$$

which - focusing only on the expectation now - under ignorability and assuming correct specification and consistent estimation of selection model $\pi_T(\underline{\mathbf{x}}^{T-1})$ includes the ratio of two correctly specified single phased π-estimators and so , similar to fixed population inference of sampling theory logic, these are approximately unbiased (Särndal et al., 1992, page 176) for the corresponding ratio of (sub)population quantities $\mathbf{A}^{T-1}_{\pi s^{T-1}} \underline{\mathbf{X}}^{T-1'}_{s^{T-1}} \underline{\mathbf{W}}^{T-1}_{s^{T-1}} \underline{\pi}^{-1}_{T-1 s^{T-1}} \mathbf{y}_{s^{T-1}}$. Furthermore, as $\underline{\mathbf{X}}^{T-1}_{s^{T-1}} \mathbf{A}^{T-1}_{\pi s^{T-1}} \underline{\mathbf{X}}^{T-1'}_{s^{T-1}} \underline{\mathbf{W}}^{T-1}_{s^{T-1}} \underline{\hat{\boldsymbol{\pi}}}^{-1}_{T-1 s^{T-1}} = \mathbf{H}^{T-1}_{s^{T-1} s^{T-1}}$ and for any t , the hat matrix $\mathbf{H}^{t-1}_{s^{t-1} s^{t-1}} = \mathbf{I}_{s^{T-1}}$ the n_{T-1} identity matrix. Thus,

$$\begin{aligned} E(\hat{\bar{\mathbf{Y}}}_{\pi m}) &= N^{-1} \mathbf{1}'_{s^0} E\left\{\cdots EE\left(\underline{\mathbf{H}}^{T-1}_{\pi s^{T-1} s^T} \mathbf{Y}_{s^T} | \underline{\mathbf{x}}^{T-1}, \mathbf{y}, \underline{s}^{T-1}\right) \cdots | \mathbf{x}^0, \mathbf{y}, s^0\right\} \\ &\approx N^{-1} \mathbf{1}'_{s^0} E\left\{\cdots EE\left(\underline{\mathbf{H}}^{T-2}_{\pi s^{T-2} s^{T-1}} \mathbf{Y}_{s^{T-1}} | \underline{\mathbf{x}}^{T-2}, \mathbf{y}, \underline{s}^{T-2}\right) \cdots | \mathbf{x}^0, \mathbf{y}, s^0\right\} \\ &= \cdots \\ &= N^{-1} \mathbf{1}'_{s^0} EE\left(\underline{\mathbf{H}}^0_{\pi s^0 s^1} \mathbf{Y}_{s^1} | \mathbf{x}^0, \mathbf{y}\right) = E\left(\bar{\mathbf{Y}}_{s^0}\right) \end{aligned}$$

reaching the final line by repeating the above derivation for all sequences under the assumption that models $\pi_t(\underline{\mathbf{x}}^{t-1})$ $t = T, T-1, ..., 1$ hold.

Robustness to misspecification of the π models can be found by conditioning $E(\hat{\bar{\mathbf{Y}}}_{\pi m})$ sequentially on the pairs $(\mathbf{X}^{t-1}, \mathbf{S}^t)$ for all phases $t = 1, ...T$. This allows us to 'condition out' the selection process for each phase and derivation relies only on correct $m-$models ' specification. To show unbiasedness start with

$$E(\hat{\bar{\mathbf{Y}}}_{\pi m}) = N^{-1} E\left\{\cdots EE\left(\sum_{\underline{s}^T} \underline{h}^{T-1}_{\pi_T k} Y_k | \underline{\mathbf{x}}^{T-1}, \underline{s}^T\right) \cdots | \mathbf{x}^0, s^1\right\}$$

and by first letting $m_T = \mathbf{X}^T$ and under model $m_{T-1}(\underline{\mathbf{x}}^{T-1}, s^{T-1})$ which in our case states that $E(m_{Tk} | \underline{\mathbf{x}}^{T-1}, \underline{s}^T) = \underline{\mathbf{X}}^{T-1'}_k \boldsymbol{\beta}_{T-1}$

$$\begin{aligned} E(\hat{\bar{\mathbf{Y}}}_{\pi m}) &= N^{-1} E\left\{\cdots EE\left(\sum_{\underline{s}^T} \underline{h}^{T-1}_{\pi_T k} Y_k | \underline{\mathbf{x}}^{T-1}, \underline{s}^T\right) \cdots | \mathbf{x}^0, s^1\right\} \qquad (3.31) \\ &= N^{-1} E\left\{\cdots E\left(\sum_{\underline{s}^T} \underline{h}^{T-1}_{\pi_T k} \underline{\mathbf{X}}^{T-1'}_k \boldsymbol{\beta}_{T-1} | \underline{\mathbf{x}}^{T-2}, \underline{s}^{T-1}\right) \cdots | \mathbf{x}^0, s^1\right\} \end{aligned}$$

by noting that

$$\begin{aligned}
\sum_{\underline{s}^T} \underline{h}_{\pi_T k}^{T-1} \underline{\mathbf{X}}_k^{T-1'} \boldsymbol{\beta}_{t-1} &= \mathbf{1}_{s^0}^{'} \underline{\mathbf{H}}_{\pi s^{T-2} s^{T-1}}^{T-2} \underline{\mathbf{X}}_{s^{T-1}}^{T-1} \mathbf{A}_{\pi s^T}^{T-1} \sum_{\underline{s}^T} \underline{\mathbf{X}}_k^{T-1} \mathbf{W}_k^{T-1} \hat{\underline{\pi}}_{Tk}^{-1} \underline{\mathbf{x}}_k^{T-1'} \boldsymbol{\beta}_{t-1} \\
&= \mathbf{1}_{s^0}^{'} \underline{\mathbf{H}}_{\pi s^{T-2} s^{T-1}}^{T-2} \underline{\mathbf{X}}_{s^{T-1}}^{T-1} \boldsymbol{\beta}_{t-1}
\end{aligned}$$

then continuing (3.31) we get

$$\begin{aligned}
E(\hat{\overline{\mathbf{Y}}}_{\pi m}) &= N^{-1} E \left\{ \cdots E E \left(\sum_{\underline{s}^T} \underline{h}_{\pi_T k}^{T-1} m_{Tk} | \underline{\mathbf{x}}^{T-1}, \underline{s}^T \right) \cdots | \mathbf{x}^0, s^1 \right\} \\
&= N^{-1} \mathbf{1}_{s^0}^{'} E \left\{ \cdot \cdot E \left(\underline{\mathbf{H}}_{\pi s^{T-2} s^{T-1}}^{T-2} \underline{\mathbf{X}}_{s^{T-1}}^{T-1} \boldsymbol{\beta}_{t-1} | \underline{\mathbf{x}}^{T-2}, \underline{s}^{T-1} \right) \cdots | \mathbf{x}^0, s^1 \right\} \\
&= N^{-1} E \left\{ \cdot \cdot E \left(\sum_{\underline{s}^{T-1}} \underline{h}_{\pi_{T-1} k}^{T-2} m_{T-1k} | \underline{\mathbf{x}}^{T-2}, \underline{s}^{T-1} \right) \cdots | \mathbf{x}^0, s^1 \right\} \\
&= \cdots \\
&= N^{-1} E E (\sum_{s^1} \underline{h}_{\pi_1 k}^{0} m_{1k} | \mathbf{x}^0, s^1) = N^{-1} \sum_{s^0} E(m_{ok}) = E(\overline{\mathbf{Y}}_{s^0})
\end{aligned}$$

by assuming ignorability and that all linear regression models $m_{t-1}(\underline{\mathbf{x}}^{t-1}, s^{t-1})$ hold over the rest $t = T-1, T-2, ..., 1$ phases.

It is useful to contrast this πm-estimator to others that use both parts of the joint distribution under a sequential framework. One frequently cited estimator is described by Bang and Robins (2005) in the context of longitudinal observational studies where monotone missing data assumption is appropriate. Its estimation algorithm is similar to ours but in its use of $\underline{\pi}_{tk}$, which it includes (its inverse) as an additional covariate in the regression model rather than a weights. Robins et al. (2000) discusses this model in more detail giving its asymptotic properties and shows it is also a double robust estimator.

As noted in the previous chapter, this idea, of unit selection as a covariate, is popular in the single phase setting[7], however from a model driven perspective, I find this method unintuitive and that there is little justification or diagnostics regarding the appropriateness of such practice given it requires a very strong assumption: the conditional expectation of response is linear in the selection probability. In the following section I show by simulation that these two πm-estimators behave similarly in terms of efficiency and bias correction.

Another estimator is the regression estimator (GREG) under a two phase setting, popularized by (Särndal et al., 1992, section 9.7), but which I here expand

[7] Hade and Lu (2014) found that 24% of literature reviewed using propensity score estimation takes this approach.

to a T phase setting.

The fundamental departure of the GREG estimator is in the modelling of the outcome variable of interest Y. While the πm-estimator models sequentially the conditional expectation m_t on $\underline{\mathbf{X}}^{t-1}$ over the population s^{t-1}, the GREG estimator suggests a set of models of the conditional distribution of Y on $\underline{\mathbf{X}}^{t-1}$ each over the entire finite population s^0. This difference leads to an estimator which has only a weak form of double robustness. In the case of Web survey inference this difference can be of significance.

More specifically, GREG estimation starts by assuming for each $t = 1, .., T$ that the point scatter $(Y_k, \underline{\mathbf{X}}_k^{t-1})$ in the finite population s^0 can be modelled by

$$\begin{cases} E(Y_k) & = \underline{\mathbf{X}}_k^{t-1'} \boldsymbol{\beta}_{t-1} \\ V(Y_k) & = \sigma_{t-1k}^2 \end{cases}$$

and estimates the population average by

$$\hat{\bar{\mathbf{Y}}} \;=\; N^{-1} \sum_{s^0} \left\{ \hat{Y}_{1\pi k} + \sum_{t=1}^{T-1} S_k^t (\hat{Y}_{t+1\pi k} - \hat{Y}_{t\pi k}) \hat{\underline{\pi}}_{tk}^{-1} + S_k^T (Y_k - \hat{Y}_{T\pi k}) \hat{\underline{\pi}}_{Tk}^{-1} \right\} \quad (3.32)$$

with $\hat{y}_{t\pi k} = \underline{\mathbf{x}}_k^{t-1'} \hat{\boldsymbol{\beta}}_{\pi(s^T)}^{t-1}$. The estimators $\hat{\boldsymbol{\beta}}_{\pi(s^T)}^{t-1}$ of the unknown regression coefficients $\boldsymbol{\beta}_{t-1}$ are derived by the following $\pi-$estimation logic: If the y_k-values were known for the whole set s^0, an estimator of the unknown $\boldsymbol{\beta}_{t-1}$vector could be formed, over the entire population , namely

$$\hat{\boldsymbol{\beta}}^{t-1} \;=\; \left(\sum_{s^0} \frac{\underline{\mathbf{X}}_k^{t-1'} \underline{\mathbf{X}}_k^{t-1}}{\sigma_{t-1k}^2} \right)^{-1} \left(\sum_{s^0} \frac{\underline{\mathbf{X}}_k^{t-1} Y_k}{\sigma_{t-1k}^2} \right).$$

What can actually be calculated from the available data is π - weighted regression coefficient vector

$$\hat{\boldsymbol{\beta}}_{\pi(s^T)}^{t-1} = \left(\sum_{\underline{s}^T} \frac{\underline{\mathbf{X}}_k^{t-1'} \underline{\mathbf{X}}_k^{t-1}}{\sigma_{t-1k}^2 \hat{\underline{\pi}}_{Tk}} \right)^{-1} \left(\sum_{\underline{s}^T} \frac{\underline{\mathbf{X}}_k^{t-1} Y_k}{\sigma_{t-1k}^2 \hat{\underline{\pi}}_{Tk}} \right)$$

computed over the final set s^T.

As a ratio of two unbiased $\pi-$estimators, under ignorability $\hat{\boldsymbol{\beta}}_{\pi(s^T)}^{t-1}$ estimates consistently the finite population quantity $\hat{\boldsymbol{\beta}}^{t-1}$, however, it does not necessarily estimate the model parameter $\boldsymbol{\beta}^{t-1}$. This can be easily shown

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}) &= EE\left\{\left(\sum_{\underline{s}^T} \frac{\underline{\mathbf{X}}_k^{t-1'}\underline{\mathbf{X}}_k^{t-1}}{\sigma^2_{t-1k}\underline{\hat{\pi}}_{Tk}}\right)^{-1} \sum_{\underline{s}^T} \frac{\underline{\mathbf{X}}_k^{t-1}Y_k}{\sigma^2_{t-1k}\underline{\hat{\pi}}_{Tk}}|\underline{\mathbf{x}}^{T-1}, \underline{s}^T\right\} \\ &= E\left\{\left(\sum_{\underline{s}^T} \frac{\underline{\mathbf{X}}_k^{t-1'}\underline{\mathbf{X}}_k^{t-1}}{\sigma^2_{t-1k}\underline{\hat{\pi}}_{Tk}}\right)^{-1} \sum_{\underline{s}^T} \frac{\underline{\mathbf{X}}_k^{t-1'}\underline{\mathbf{X}}_k^{T-1}}{\sigma^2_{t-1k}\underline{\hat{\pi}}_{Tk}}\right\}\boldsymbol{\beta}^{T-1} \end{aligned}$$

and so when modelling the point scatter $(Y_k, \underline{\mathbf{X}}_k^{t-1})$ for the final the set of $t = T$ is $E(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}) = \boldsymbol{\beta}^{t-1}$, however, for the rest $t = 1, ..., T-1$ models this is not the case. The reason for this result is that as only $\mathbf{Y}_{s^T}$ is observed, for ignorability to hold we must condition on $\underline{\mathbf{X}}^{T-1}$ regardless of which subpopulation level regression coefficient we are estimating.

More generally, for $t = 1, .., T$, the estimator $\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}$ is approximately unbiased of $\boldsymbol{\beta}^{t-1}$ if selection models $\pi_t(\underline{\mathbf{X}}^{t-1})$ up to t *and* the specific model $E(Y_k) = \underline{\mathbf{x}}_k^{t-1'}\boldsymbol{\beta}_{t-1}$ are both correctly specified and consistently estimated. This can be shown by first noting that

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}) &= EE\left\{E\left[\cdot\cdot E\left(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}|y, \underline{\mathbf{x}}^{T-1}, \underline{s}^{T-1}\right)\cdot\cdot|y, \underline{\mathbf{x}}^t, \underline{s}^t\right]|\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1}\right\} \\ &= E(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^t)}) \end{aligned}$$

as $E(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}|y, \underline{\mathbf{x}}^{T-1}, \underline{s}^{T-1}) = \hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{T-1})}$, $E(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{T-1})}|y, \underline{\mathbf{x}}^{T-2}, \underline{s}^{T-2}) = \hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{T-2})}, ...,$ and so on up to phase t where $E(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{t+1})}|y, \underline{\mathbf{x}}^t, \underline{s}^t) = \hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^t)}$ holds under ignorability when selection models for phases $T, ..., t$ hold. Next if the specific model $E(Y_k) = \underline{\mathbf{X}}_k^{t-1'}\boldsymbol{\beta}_{t-1}$ is also correct then

$$E(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^t)}) = EE\left\{\left(\sum_{s^0} \underline{S}_k^t\underline{\mathbf{X}}_k^{t-1'}\underline{\mathbf{X}}_k^{t-1}\underline{\hat{\pi}}_{tk}^{-1}\right)^{-1}\left(\sum_{s^0} \underline{S}_k^t\underline{\mathbf{X}}_k^{t-1'}Y_k\underline{\hat{\pi}}_{tk}^{-1}\right)|\underline{\mathbf{x}}^{t-1}, \underline{s}^t\right\} = \boldsymbol{\beta}^{t-1}$$

as $E(Y_k|\underline{\mathbf{x}}^{t-1}, \underline{s}^t) = \underline{\mathbf{X}}^{t-1'}\boldsymbol{\beta}^{t-1}$.

This π-leaning set of assumptions for the estimator $\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}$ carry over when considering the properties of the GREG estimator. Similarly to the derivation on the model coefficients it can be shown that the GREG estimator $\hat{\overline{\mathbf{Y}}}_{\pi y}$ under an ignorable T phase selection distribution is unbiased of the population mean $\overline{\mathbf{Y}}_{s^0}$ if *both* (i) the selection model $\pi_1(\mathbf{x}^0)$ or the measurement model $E(Y_k) = \mathbf{X}_k^{0'}\boldsymbol{\beta}_0$ holds, *and that* (ii) all selection models $\pi_t(\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1})$ for $t = 2, ..., T$ are correctly specified and consistently estimated. This can be seen by algebraically rearranging (3.32) so that

$$\hat{\overline{\mathbf{Y}}} = N^{-1}\sum_{s^0}\left\{Y_k + \sum_{t=1}^{T}\underline{S}_k^{t-1}(\hat{\pi}_{tk} - S_k^t)(\hat{Y}_{t\pi k} - Y_k)\underline{\hat{\pi}}_{tk}^{-1}\right\}$$

and following a similar derivation to that of the regression coefficient above (see section 5.2).

In the context of survey sampling where the selection process is assumed known this form of robustness is useful. However, when the selection process is not known, as is our case, comparing the two sequential estimators which incorporate both m and π type models the πm estimator has a much stronger form of double robustness.

3.5.4 A Short Note on Variance Estimation Under a Multi-Phase Framework

For the single phase estimators discussed in the previous chapter, closed-form estimators for the variances of $\hat{\bar{\mathbf{Y}}}_{\pi}$,$\hat{\bar{\mathbf{Y}}}_{m}$ and $\hat{\bar{\mathbf{Y}}}_{\pi m}$ are available . For the $\pi-$estimator, a classic design variance estimator (for example see Särndal et al. 1992, chapter 9) can be suggested by replacing the unknown selection probabilities by their estimated values. This can be justified by noting that the variance under the joint $(\mathbf{Y}, \mathbf{S})$ distribution is approximately the same as the $\mathbf{Y}-$expectation of the $\mathbf{S}$-variance when the sample fraction is small (Pfeffermann, 1993). This gives a conservative estimate as it ignores the estimation of selection probabilities over the sample data.[8] Williamson et al. (2012); Lunceford and Davidian (2004) and others have proposed adjusted estimators to take this into account. For the m and πm-estimators discussed in the previous chapter closed-form estimators for the unconditional variance under an ignorable one phase framework have also been suggested. Specifically, for these estimators, when a linear m- model is assumed, a sandwich type estimators can be easily derived. Valliant et al. (2000) discuss this idea for the $m-$estimators and in Valliant (2002) you can find a similar estimator proposed for $\pi m-$estimators in a one phase case.

Neither of these approaches translate smoothly to the sequential framework. The 'conservativeness' of the fixed population approach for variance estimator of $\hat{\bar{\mathbf{Y}}}_{\pi}$ becomes non negligible, while a direct derivation of a variance estimator for $\hat{\bar{\mathbf{Y}}}_{m}$ and $\hat{\bar{\mathbf{Y}}}_{\pi m}$, which can still be expressed in linear form, is not trivial. However, relevant work on closed-form variance estimators has been done in more recent years. In the setting of dynamic treatment regimes in a multi-interval setting, Robins (2004) derives the variance estimator for doubly robust g-estimators, which should be parallel to the πm estimation under our sequential framework,[9]

[8] Estimating the selection probability results in gains in efficiency as they capture the additional random sampling error and so an estimator using the estimated rather than the true selection probability will be less variable

[9] As I've discussed briefly in earlier chapters G-estimation, is a generalization of $M-$estimation Stefanski and Boos (2002) using as well selection probabilities to give the DR

by linearising the associated system of generalized estimating equations. Moodie (2009) proposed to estimate the parameters of this linearised variance recursively. But, adopting this estimation system to our specific problem is outside of the scope of this work.

Apart from closed form solutions, another popular approach to variance estimation is resampling and replication methods (such as jackknifing, balanced repeated replication, bootstrapping). A description of these methods can be found in many books on survey sampling, for example in Lohr (2009, chapter 9). These methods are relatively easy to implement, regardless of the form of the point estimator, and there is a general move towards them and away from the analytical approach (Binder et al., 2004). I suggest now a simple Bootstrap approach (Efron, 1979) to our inferential question.

For the survey sampling case, a direct extension for *i.i.d.* samples (Shao and Tu, 1996, see) is to apply the standard bootstrap, and if necessary independently in each stratum. However, such an estimator may be inconsistent. For example, the estimator $\hat{\overline{Y}} = \sum_h W_h \overline{Y}_{sh}$ under stratified random sampling, has variance $V(\hat{\overline{Y}}) = \sum_h W_h^2 s_{yh}^2$ where W_h are population strata sizes and s_{yh}^2 is the strata sample variance. A standard Bootstrap estimator resamples with replacement independently from each strata and the expectation of the estimator is $\sum_h W_h^2 (\frac{n_h-1}{n_h}) s_{yh}^2$, inconsistent when stratum sample sizes are bound. To deal with these inconsistencies many modified bootstrap methods have been proposed Mach et al. (2005). Still, for the case of unequal probability selection with independent draws a naive bootstrap is valid (Lohr, 2009, chapter 6, section 1).

In our sequential framework the values of the variables of interest are not independent given the hierarchical structure, and so a simple bootstrap may misrepresent variability in the sampling distribution of statistics Ramasubramanian et al. (2002). In order to preserve the correlation structure in resamples, we may mimic the data generating mechanism by resampling in a nested fashion. Most published work in the survey sampling literature consider multi stage samples so that resampling follows a top (clusters) to bottom (units) approach. However, in our case this is not applicable - the hierarchy is not clustered but is rather through a sequential linkage.

A simple solution is to reverse the process and start at the bottom of the hierarchy instead of the top. Ramasubramanian et al. (2002) discusses such a bootstrap approach under the two phase simple random sample framework which can be easily generalized to T phases and shows its consistency for the ratio estimator. The key idea is to keep the proportion of units belonging to s^t and

property.

$\overline{s}^t \cap s^{t-1} = (s^{t-1} - s^t)$, the selected and unselected populations, appearing in each Bootstrap resample the same as in the original sample. As I shall now show,such an approach seems to give useful results and is worth exploring further in future work.

The algorithm for $T = 2$ is as follows:

1. Draw a simple random sample of size n_2 with replacement from the set s^2 of size n_2. Denote the set s_b^2 .
2. Draw another simple random sample of size $n_1 - n_2$ with replacement from the set $(s^1 - s^2)$ of size $n_1 - n_2$. Denote this set by $\overline{s}_b^2$ and let $s_b^1 = s_b^2 \cup \overline{s}_b^2$ and similarly define $s_b^0 = (s^0 - s^1) \cup s_b^1$.
3. Repeat the steps (i) and (ii) independently B times and let the b_{th} $(b = 1, 2, ..., B)$ Bootstrap resampled set be $\underline{s}_b^2 = (s_b^0, s_b^1, s_b^2)$.

For any of the three estimator types we discuss, the Bootstrap variance estimator for this method are given by

$$\hat{V}(\hat{Y}) = \frac{1}{B-1} \sum_b (\hat{Y}_b - \overline{\hat{Y}}_B)^2$$

where $\hat{Y}$ is the π, m or $\pi m-$estimator, $\hat{Y}_b$ is the same estimator calculated over the b_{th} simulated population $\underline{s}_b^2 = (s_b^0, s_b^1, s_b^2)$ and $\overline{\hat{Y}}_B = B^{-1} \sum_b \hat{Y}_b$.

3.6 A Simulation Study

In the following section I study numerically, over three separate simulation studies, some statistical properties and specific behaviours of the three sequential estimators described in this chapter.

In the first study I examine the basic properties of the three estimators under different model specifications. This includes the bias and efficiency of the estimators under the correct specification of the set of $\underline{\pi}_T$ and $\underline{\mathbf{x}}^T$ models as well as the different combinations possible for misspecifications. Under these setting I also try and address the question raised by Kang and Schafer (2007) who suggested that a πm-type of estimation which hold the valuable double robust property perform, however, worse than a π or m estimator when *both* sets of model are misspecified. This issue and its subsequent discussion (Ridgeway and McCaffrey, 2007; Tsiatis and Davidian, 2007; Robins et al., 2007) was under a single phase framework and it is interesting to put the question in a sequential setting as well.

In the second study I test the behaviour of a bootstrap variance estimation approach outlined in section 3.5.4 for the three sequential estimators. After outlining again the estimation algorithm, I perform a simulation study to test this algorithm under different population settings. I add also a short example on the increase and decrease in estimator variance moving from a one phase to a multi phase framework

In the final study I give what is intended to be a simple cautionary tale for the average practitioner estimating population estimands in a Web panel setting. The problem is that in many cases in the estimation procedure practitioners 'impute' a zero to missing values of Internet related quantitative measurements for example the number of daily visits to a certain web site. Intuitively, and certainly from a fixed population perceptive, this is a logical approach. However, I show that this assumption is an imputation which relies on certain assumptions.

3.6.1 Basic Examination of the Three Sequential Estimators

As noted above, one of the objectives in this section is to test the bias of the estimators under the misspecification of both π and m model sets. One of the critiques of the Kang and Schafer (2007) study, which dealt with this question, was that the authors picked an extreme simulated population to prove their point of the superiority of m-only estimation. To avoid this and provide a certain objectivity, I chose a sequential dataset available in literature not related to the misspecification issue. Thus I have recreated[10] the population used in Bang and Robins (2005, table 3). The use of this dataset in particular allows me also an immediate comparison to the competing πm-algorithm where the the selection probabilities are included as an additional covariate.

From finite population s^0, two phases of selection define two population subsets denoted by $\underline{s}^2 = (s^1, s^2)$. The variable of interest and relevant covariates over the population are $\underline{\mathbf{X}}^2_{s^0} = (\mathbf{X}^{0\prime}_{s^0}, \mathbf{X}^1_{s^0}, \mathbf{X}^2_{s^0})'$ where $\mathbf{X}^0_{s^0} = (\mathbf{X}^0_{1s^0}, \mathbf{X}^0_{2s^0}, \mathbf{X}^0_{3s^0})'$ are variables associated with the finite population s^0 , the vector $\mathbf{X}^1_{s^0}$ is the single variable associated with the subpopulation s^1 and the vector $\mathbf{X}^2_{s^0} = \mathbf{Y}_{s^0}$ is the variable of interest.The size of the finite population we study is $10,000$.

The population $\underline{\mathbf{X}}^2_{s^0}$ follows a multivariate normal (MVN) distribution. Values

[10] The exact parameter details were kindly supplied by Dr Bang through private communication.

$\mathbf{x}^0_{s^0} = (\mathbf{x}^0_{1s^0}, \mathbf{x}^0_{2s^0}, \mathbf{x}^0_{3s^0})'$ were generated independently from a standard normal distribution, the values $\mathbf{x}^1_{s^0}$ were generated by $N(m_0(\mathbf{x}^0, \boldsymbol{\beta}_0), 1)$ which denotes a normal distribution with mean a function of $\mathbf{X}^0$ and variance one; Values of the outcome measurement of interest $\mathbf{y}_{s^0}$ were generated from $N(m_1(\underline{\mathbf{x}}^1, \boldsymbol{\beta}_1), 1)$, where the mean $(m_1(\underline{\mathbf{x}}^1, \boldsymbol{\beta}_1)$ is a function of variables associated with both s^0 and s^1. Table 3.3 describes the structure and parameters of the mean function. Note that a MVN distribution linking $\mathbf{X}^1$ to $\mathbf{X}^0$ and the two to the $\mathbf{Y}$ is not required for either of our estimators - a strength compared to likelihood estimation. In sections 3.6.2 and 3.6.3 I simulate alternative populations where $\mathbf{X}^0$ and $\mathbf{X}^1$ have no obvious parametric relationship.

The subpopulation s^1 is generated by the selection model $\pi_1(\mathbf{x}^0, \boldsymbol{\alpha}_0)$ defined over s^0 which suggests that each member of the population has a positive probability, a function of $\mathbf{X}^0$ of being a member of s^1. The sample set s^2 is generated by the function $\pi_2(\underline{\mathbf{x}}^1, \boldsymbol{\alpha}_1)$ defined over the subpopulation s^1. The two selection models are described in table 3.3. Under this data configuration, the expected size of subset s^1 is $n_1 \approx 3,300$ and the expected size of sample s^2 is $n_2 \approx 700$. Our interest is in estimating $E(\overline{\mathbf{Y}}) = 11$, and I assume we observe the set of of values $(\mathbf{x}^{0'}_{s^0}, \mathbf{x}^1_{s^1}, \mathbf{y}_{s^2})'$.

	Models	Parameters
$m_0(\mathbf{x}^0, \boldsymbol{\beta}_0)$	$\boldsymbol{\beta}_0[1, \mathbf{x}^0_1, \mathbf{x}^0_1 \cdot \mathbf{x}^0_3]'$	$\boldsymbol{\beta}_0 = [0, 3, -2]$
$m_1(\underline{\mathbf{x}}^1, \boldsymbol{\beta}_1)$	$\boldsymbol{\beta}_1[1, (\mathbf{x}^0_1)^2, \mathbf{x}^0_2, (\mathbf{x}^1)^2), \mathbf{x}^0_2 \cdot \mathbf{x}^1]'$	$\boldsymbol{\beta}_1 = [0, -3, 3, 1, -2]$
$\pi_1(\mathbf{x}^0, \boldsymbol{\alpha}_0)$	$\boldsymbol{\alpha}_0[1, \mathbf{x}^0_1, \mathbf{x}^0_2, \mathbf{x}^0_3, \mathbf{x}^0_1 \cdot \mathbf{x}^0_2]'$	$\boldsymbol{\alpha}_0 = [0, -1, 1, 1, -1, -1]$
$\pi_2(\underline{\mathbf{x}}^1, \boldsymbol{\alpha}_1)$	$\boldsymbol{\alpha}_1[1, \mathbf{x}^0_1, \mathbf{x}^0_2, \mathbf{x}^0_3, \mathbf{x}^0_1 \cdot \mathbf{x}^0_2, \mathbf{x}^0_3 \cdot \mathbf{x}^1]'$	$\boldsymbol{\alpha}_1 = [0, 1, 1, 0, -1, 0, -2]$

Misspecification	Model Estimated
$m_0(\mathbf{x}^0, \boldsymbol{\beta}_0)$	$\boldsymbol{\beta}_0[1, \mathbf{x}^0_1, \mathbf{x}^0_2]'$
$m_1(\underline{\mathbf{x}}^1, \boldsymbol{\beta}_1)$	$\boldsymbol{\beta}_1[1, \mathbf{x}^0_1, (\mathbf{x}^0_2)^2, (\mathbf{x}^0_3)^2, \mathbf{x}^1]'$
$\pi_1(\mathbf{x}^0, \boldsymbol{\alpha}_0)$	$\boldsymbol{\alpha}_0[1, \mathbf{x}^0_2, \mathbf{x}^0_3]'$
$\pi_2(\underline{\mathbf{x}}^1, \boldsymbol{\alpha}_1)$	$\boldsymbol{\alpha}_1[1, \mathbf{x}^1]'$

Table 3.3: Simulation scenarios for estimation models- Both correctly specified and misspecified. Parameters for outcome and selection models represent linear and logistic regression model coefficients respectively.

Of interest here is the behaviour of the three estimators under different levels of misspecification. I use four false models described in table 3.3. Under this set up, we calculate $\hat{\overline{\mathbf{Y}}}_\pi$,$\hat{\overline{\mathbf{Y}}}_m$and $\hat{\overline{\mathbf{Y}}}_{\pi m}$ using both the correct model specifications and all fifteen possible misspecification combinations.

The $\pi-$estimator $\hat{\overline{\mathbf{Y}}}_\pi$ is constructed based on the sequential π-estimation procedure described in section 3.5.2: (a) fit a logistic regression of S^1_k on $\mathbf{x}^0$ for all $k \in s^0$ the full population , and separately model, also by logistic regression, S^2_k on $\underline{\mathbf{x}}^1$ over the subpopulation s^1 and obtained the regression coefficient estimates; (b) compute estimated selection probabilities $\hat{\pi}_1$ and $\hat{\pi}_2$ for all units in s^2 where both $\mathbf{x}^0$ and $\mathbf{x}^1$ are fully observed; (c) estimate the population mean by the π-estimator (3.25) with probability weights $\hat{\underline{\pi}}_{2k}^{-1} = \hat{\pi}_{1k}^{-1} \cdot \hat{\pi}_{2k}^{-1}$.

The $m-$estimator $\hat{\bar{\mathbf{Y}}}_m$ is constructed following the algorithm described in section 3.5.1; I regress $\mathbf{y}$ on $\underline{\mathbf{x}}^1 = (\mathbf{x}^0, \mathbf{x}^1)$ over the survey sample set s^2 and compute the corresponding predicted values for all units in s^1. Next, these predicted values were regressed on $\mathbf{x}^0$. Hence, the new predicted values were obtained as a function of $\mathbf{x}^0$ only, which is known for the full population s^0. The average of this quantity is the final estimate.

The πm- estimator $\hat{\bar{\mathbf{Y}}}_{\pi m}$ was calculated by a similar process described in section 3.5.3 which in essence follows the m-estimator computation, but for each t the $m-$estimation equation is weighted by the estimated selection probabilities π_t.

We examine these three estimators by summarizing over 500 simulations when all π and m models are correctly specified, as well as for all 15 possible cases of misspecification which include: four single model misspecification, six double model specification, four triple model misspecification and the case where all four models are misspecified. A summary of these scenarios is presented in table 3.4 with the associated bias, variance and inter quantile range of each estimator.

The first thing to notice from the results in table 3.4 is that in terms of point estimation, the performance of the estimators under comparison are in agreement with our discussion and theory. The π-estimator and m-estimator are unbiased when either $\pi_2(\underline{\mathbf{x}}^1), \pi_1(\mathbf{x}^0)$ or $m_2(\underline{\mathbf{x}}^1), m_1(\mathbf{x}^0)$ respectively are correctly modelled. These scenarios are indicated by an underline in the table. On the other hand in all cases of misspecification the mean bias is non negligible. In our specific settings the $m-$estimator $\hat{\bar{\mathbf{Y}}}_m$ has a significantly larger bias under misspecification relative to the π-estimator estimators. The bias of $\hat{\bar{\mathbf{Y}}}_m$ is particularly large when the model $m_2(\underline{\mathbf{x}}^1)$ over s^1 is misspecified and is a result of the specific functional form - the inclusion of higher order terms for both $(\mathbf{x}_2^0, \mathbf{x}_3^0)$- given to the population in this study. The πm-estimator $\hat{\bar{\mathbf{Y}}}_{\pi m}$ displays the double robust property and is unbiased in all seven scenarios where either selection or outcome models or both are correctly specified. These combinations are also indicated in the table with an underline. A convenient way of comparing the relative behaviour of the three estimators in terms of bias is by the dot plot given in figure 3.3 which clearly shows the dominance of the πm-estimator compared to the π and m- estimators, underlying it's DR property.

Table 3.4: Simulation Results: Estimating $E(\overline{\mathbf{Y}}_{s^0})$ in the two phase selection model

	Bias			Variance			IQR		
Misspecification	$B(\hat{\overline{\mathbf{Y}}}_{\pi})$	$B(\hat{\overline{\mathbf{Y}}}_{m})$	$B(\hat{\overline{\mathbf{Y}}}_{\pi m})$	$V(\hat{\overline{\mathbf{Y}}}_{\pi})$	$V(\hat{\overline{\mathbf{Y}}}_{m})$	$V(\hat{\overline{\mathbf{Y}}}_{\pi m})$	$IQR(\hat{\overline{\mathbf{Y}}}_{\pi})$	$IQR(\hat{\overline{\mathbf{Y}}}_{m})$	$IQR(\hat{\overline{\mathbf{Y}}}_{\pi m})$
—-	$\underline{-0.076}$	$\underline{-0.002}$	$\underline{0.001}$	$\underline{1.077}$	$\underline{0.09}$	$\underline{0.092}$	1.373	0.397	0.429
$\pi_2(\underline{\mathbf{x}}^1)$	2.034	$\underline{0.007}$	$\underline{0.005}$	0.87	$\underline{0.094}$	$\underline{0.096}$	1.284	0.407	0.423
$\pi_1(\mathbf{x}^0)$	0.681	$\underline{0.003}$	$\underline{0.004}$	1.095	$\underline{0.102}$	$\underline{0.105}$	1.331	0.437	0.459
$m_2(\underline{\mathbf{x}}^1)$	$\underline{-0.01}$	6.792	$\underline{0.214}$	$\underline{1.021}$	1.811	$\underline{1.379}$	1.372	1.784	1.603
$m_1(\mathbf{x}^0)$	$\underline{0.062}$	3.641	$\underline{-0.015}$	$\underline{1.213}$	0.355	$\underline{0.184}$	1.518	0.842	0.573
$\pi_2(\underline{\mathbf{x}}^1)$,$\pi_1(\mathbf{x}^0)$	3.068	$\underline{0.004}$	$\underline{0.003}$	1.095	$\underline{0.097}$	$\underline{0.095}$	1.436	0.45	0.442
$m_2(\underline{\mathbf{x}}^1)$,$m_1(\mathbf{x}^0)$	$\underline{0.044}$	6.809	$\underline{0.238}$	$\underline{1.235}$	2.109	$\underline{1.47}$	1.447	1.789	1.537
$\pi_2(\underline{\mathbf{x}}^1)$,$m_1(\mathbf{x}^0)$	1.996	3.637	-0.013	0.938	0.332	0.194	1.255	0.771	0.578
$\pi_1(\mathbf{x}^0)$,$m_2(\underline{\mathbf{x}}^1)$	0.653	6.823	0.782	0.877	1.805	1.359	1.301	1.818	1.565
$\pi_2(\underline{\mathbf{x}}^1)$,$m_2(\underline{\mathbf{x}}^1)$	1.963	6.774	3.834	1.043	1.951	1.549	1.401	1.861	1.702
$\pi_1(\mathbf{x}^0)$,$m_1(\mathbf{x}^0)$	0.644	3.666	0.475	1.064	0.345	0.215	1.355	0.76	0.592
$\pi_1(\mathbf{x}^0)$,$m_2(\underline{\mathbf{x}}^1)$,$m_1(\mathbf{x}^0)$	0.604	6.766	0.686	0.895	1.746	1.416	1.16	1.801	1.471
$\pi_2(\underline{\mathbf{x}}^1)$,$m_2(\underline{\mathbf{x}}^1)$,$m_1(\mathbf{x}^0)$	2	6.771	3.821	0.946	1.909	1.395	1.216	1.867	1.604
$\pi_2(\underline{\mathbf{x}}^1)$,$\pi_1(\mathbf{x}^0)$,$m_1(\mathbf{x}^0)$	2.998	3.677	0	1.24	0.37	0.098	1.444	0.798	0.414
$\pi_2(\underline{\mathbf{x}}^1)$,$\pi_1(\mathbf{x}^0)$,$m_2(\underline{\mathbf{x}}^1)$	3.032	6.822	3.956	1.034	1.71	1.374	1.309	1.647	1.469
$\pi_2(\underline{\mathbf{x}}^1)$,$\pi_1(\mathbf{x}^0)$, $m_2(\underline{\mathbf{x}}^1)$,$m_1(\mathbf{x}^0)$	2.998	6.74	3.881	1.069	1.73	1.349	1.374	1.54	1.493

The true mean $E(\mathbf{Y})$ is 11. Bias, Variance denote the bias in the mean, variance of the estimates from 500 simulations, respectively. IQR denotes the interquartile range, that is, upper quartile (75%)-lower quartile (25%) . Each simulation is based on a population size of 10,000. Estimates in the table with an underline indicate scenarios where the estimator is expected to be unbiased under model and distribution assumptions.

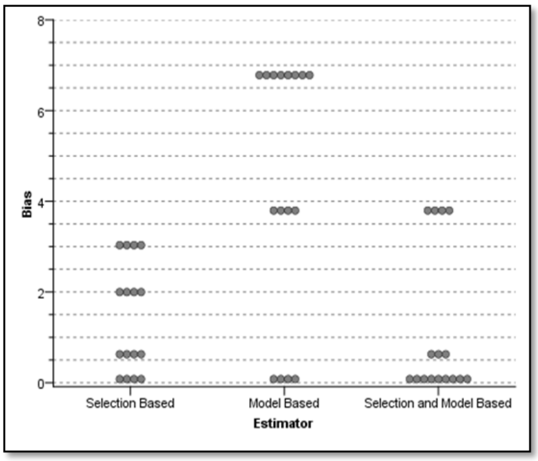


Figure 3.3: A simple dot plot summarizing the bias of the three estimators. Bias denotes the empirical deviation from the mean of the estimates from 500 simulations. Each simulation is based on a population size of 10,000.

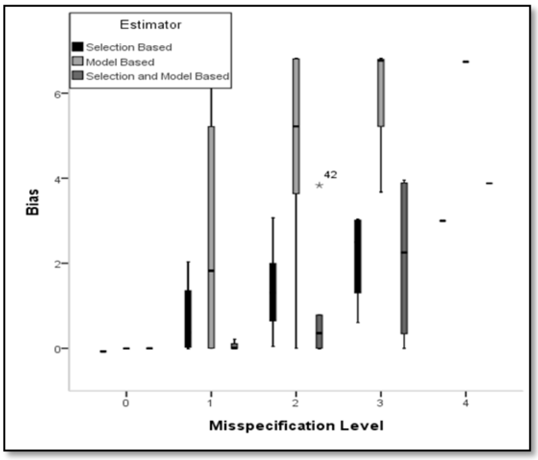


Figure 3.4: A Boxplot summarizing the bias distribution of the three estimators conditional on the number of misspecifications in the estimation models. The bias and number of simulations are identical to that of figure 3.3 on the left panel.

A lively debate has taken place on the behaviour of DR estimators under a single phase setting (as mentioned as well in the previous chapter) where *both* π and m model assumptions are misspecified. Robins (2000) suggested that even under complete misspecification, the bias of the πm-estimators will be no worse than a π only or m only estimation strategy, implying an 'extra' double robustness. Kang and Schafer (2007) on the other hand show a scenario where in fact a πm-estimator is significantly inferior to an m only estimator when both models are misspecified. Several authors have followed the issue and suggested DR type estimators (or at least paths towards such estimators) which under misspecification will give estimates which are at least no worse than an m-estimator, see Ridgeway and McCaffrey (2007); Robins et al. (2007); Tsiatis and Davidian (2007).

Ours is a sequential two phase setting and it is interesting to consider this idea of an 'extra' robustness here. Looking again at figure 3.3 the error plot does show that $\hat{\bar{\mathbf{Y}}}_{\pi m}$ has what seems to be a negligible bias for 8 out of the 15 scenarios studied, while our expectations are that the πm-estimator will be unbiased only under seven combinations. The m or π-estimators have negligible bias for only the four combinations where the relevant models are correctly specified. Furthermore, it is clear from the graph that the inclusion of π models in the $\pi m-$algorithm caps (here to under 4) the maximum level of bias which potentially we may incur if modelling only on m-models where the maximum bias (shared in eight scenarios) is 7. The box plot in figure 3.4 breaks down the results into five misspecification categories - zero to four misspecified models. From the graph, and the raw results in table 3.4 we clearly see that (a) across all specification scenarios the πm-estimator is superior to the m-estimator, while (b) in comparison to the the

	Bias Ratio		
Misspecification	$BR(\hat{\bar{\mathbf{y}}}_\pi)$	$BR(\hat{\bar{\mathbf{y}}}_m)$	$BR(\hat{\bar{\mathbf{y}}}_{\pi m})$
—-	$\underline{-0.073}$	$\underline{-0.007}$	$\underline{0.004}$
$\pi_2(\underline{\mathbf{x}}^1)$	2.18	$\underline{0.021}$	$\underline{0.018}$
$\pi_1(\mathbf{x}^0)$	0.65	$\underline{0.01}$	$\underline{0.013}$
$m_2(\underline{\mathbf{x}}^1)$	$\underline{-0.01}$	5.047	$\underline{0.182}$
$m_1(\mathbf{x}^0)$	$\underline{0.056}$	6.106	$\underline{-0.034}$
$\pi_2(\underline{\mathbf{x}}^1)$,$\pi_1(\mathbf{x}^0)$	2.932	$\underline{0.012}$	$\underline{0.011}$
$m_2(\underline{\mathbf{x}}^1)$,$m_1(\mathbf{x}^0)$	$\underline{0.04}$	4.689	$\underline{0.196}$
$\pi_2(\underline{\mathbf{x}}^1)$,$m_1(\mathbf{x}^0)$	2.061	6.311	-0.031
$\pi_1(\mathbf{x}^0)$,$m_2(\underline{\mathbf{x}}^1)$	0.697	5.079	0.671
$\pi_2(\underline{\mathbf{x}}^1)$,$m_2(\underline{\mathbf{x}}^1)$	1.922	4.85	3.081
$\pi_1(\mathbf{x}^0)$,$m_1(\mathbf{x}^0)$	0.624	6.239	1.025
$\pi_1(\mathbf{x}^0)$,$m_2(\underline{\mathbf{x}}^1)$,$m_1(\mathbf{x}^0)$	0.638	5.12	0.576
$\pi_2(\underline{\mathbf{x}}^1)$,$m_2(\underline{\mathbf{x}}^1)$,$m_1(\mathbf{x}^0)$	2.056	4.901	3.235
$\pi_2(\underline{\mathbf{x}}^1)$,$\pi_1(\mathbf{x}^0)$,$m_1(\mathbf{x}^0)$	2.692	6.043	0.002
$\pi_2(\underline{\mathbf{x}}^1)$,$\pi_1(\mathbf{x}^0)$,$m_2(\underline{\mathbf{x}}^1)$	2.981	5.217	3.375
$\pi_2(\underline{\mathbf{x}}^1)$,$\pi_1(\mathbf{x}^0)$,$m_2(\underline{\mathbf{x}}^1)$,$m_1(\mathbf{x}^0)$	2.901	5.124	3.342

Table 3.5: Simulation Results: The Bias Ratio of the three estimators. Here $BR(\cdot)$ denotes the Bias ratio calculated by $BR = \frac{B(\hat{\bar{\mathbf{y}}})}{V(\hat{\bar{\mathbf{y}}})^{1/2}}$. The variance is that of the estimates from 500 simulations.

π-estimator the πm-estimator is superior as well but for the single case where all four models are misspecified. In that sense one may argue that this study gives anecdotal evidence to the 'extra' robustness of the πm-estimation approach.

Consider next the variance and interquartile range of the estimators given in table 3.4. The m-estimator is the most efficient under correct specification, while the $\pi-$estimator is in comparison highly inefficient. As I have discussed in the previous chapter, this relative weakness of the $\pi-$estimator is well known and is exasperated in a sequential framework. What is interesting is that whenever the m-model was correctly specified, $\hat{\bar{\mathbf{Y}}}_{\pi m}$ was nearly as efficient as $\hat{\bar{\mathbf{Y}}}_m$. Thus a very small price is paid in terms of efficiency loss by using $\hat{\bar{\mathbf{Y}}}_{\pi m}$ in place of $\hat{\bar{\mathbf{Y}}}_m$, and yet as we've discussed, when the $\pi-$models were correct, huge benefits were obtained in terms of robustness against misspecification of the m-models. When both set of models are correct $\hat{\bar{\mathbf{Y}}}_{\pi m}$ and $\hat{\bar{\mathbf{Y}}}_m$ have almost identical variance. In a similar study Bang and Robins (2005) notes that following from the theory of semiparametric efficiency bounds, when assuming correct specification, $\hat{\bar{\mathbf{Y}}}_{\pi m}$ based on correct $m-$models is asymptotically more efficient than $\hat{\bar{\mathbf{Y}}}_{\pi m}$ based on an incorrect model for the $m-$models. This is born out here; indeed we see that $\hat{\overline{\mathbf{Y}}}_{\pi m}$ based on an incorrect $m-$models have variance over ten times that of $\hat{\overline{\mathbf{Y}}}_{\pi m}$ based on a correct $m-$models.

I conclude this section with a brief discussion on the behaviour of the estimators as tools of inference. I use the bias ratio statistic as an indicator to the expected nominal coverage of confidence intervals. As Särndal et al. (1992, Table 5.1, p165) discuss, under the assumption that $(\hat{\overline{\mathbf{Y}}} - E(\overline{\mathbf{Y}}))/V(\hat{\overline{\mathbf{Y}}})^{1/2}$ follows the standard normal distribution, a bias ratio of less 0.5 has only a small effect on the nominal coverage (92.1%) on a confidence interval with $\alpha = 5\%$, and even a bias ratio of 0.65 still is expected to have a coverage of 90% which may be considered still tolerable. The bias ratio of the estimators over the 16 specifications studied are given in table 3.5, a summary of these results in the form of a dot plot is given in figure 3.5.

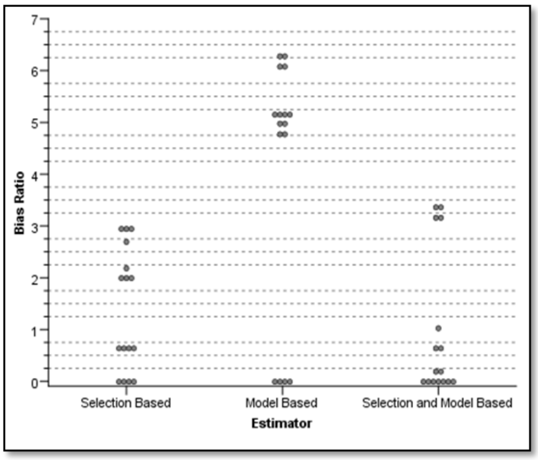


Figure 3.5: A dot plot of the Bias ratio (BR) of the three estimators. BR denotes the empirical deviation from the estimates' over 500 simulations divided by the empirical standard deviation. Each simulation is based on a population size of 10,000.

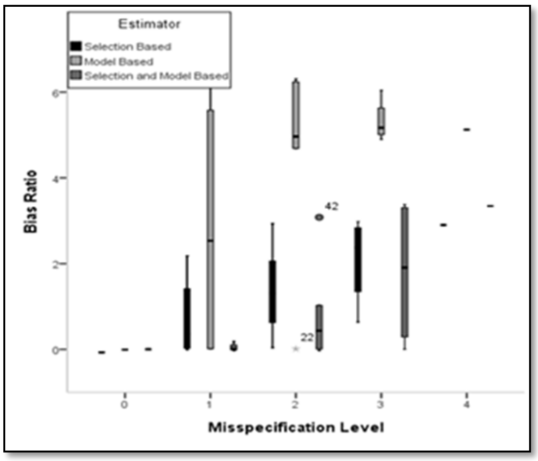


Figure 3.6: A Boxplot of the bias ratio distribution of the three estimators conditional on the number of misspecifications in the estimation models. The bias ratio and number of simulations are identical to that of figure 3.5 on the left panel.

Looking at the dot plot it is immediately clear that the m-estimator is useful as an inferential (that is at least 90% nominal coverage) tool only when the $m-$models are both correctly specified. On the other hand, the $\pi-$estimator gives what may be considered useful inference in 8 out of the 16 scenarios, and the πm estimator in 11 (or even 12) scenarios. From the box plot in figure 3.6 the same picture emerges - πm gives better inferential results over all levels of misspecification but the case where both sets of model specification are incorrect in which case neither of the three approaches yield meaningful results. As a final remark, note that when looking at the nine misspecification scenarios (meaning I ignore the first seven cases in table 3.4) πm is superior to the m-estimator in all scenarios, and is equal to the π-estimator.

3.6.2 A study of Variance Estimation by Bootstrap and on the Increase or Decrease in Variance when Moving from a Single to a Multi-Phase Framework

In the following section I start by testing the bootstrap variance estimator algorithm which I discussed in section 3.5.4. For convenience I give the algorithm again for the case $T = 2$:

1. Draw a simple random sample of size n_2 with replacement from the set s^2 of size n_2. Denote the set s_b^2 .
2. Draw another simple random sample of size $n_1 - n_2$ with replacement from the set $(s^1 - s^2)$ of size $n_1 - n_2$. Denote this set by $\overline{s}_b^2$ and let $s_b^1 = s_b^2 \cup \overline{s}_b^2$ and similarly define $s_b^0 = (s^0 - s^1) \cup s_b^1$.
3. Repeat the steps (i) and (ii) independently, say, B times and let the b_{th} $(b = 1, 2, ..., B)$ Bootstrap resampled set be $\underline{s}_b^2 = (s_b^0, s_b^1, s_b^2)$.

For any of the three estimator types we discuss, the Bootstrap variance estimator for this method are given by

$$\hat{V}(\hat{Y}) = \frac{1}{B-1} \sum_b (\hat{Y}_b - \overline{\hat{Y}}_B)^2$$

where $\hat{Y}$ is the π, m or $\pi m-$estimator, $\hat{Y}_b$ is the same estimator calculated over the b_{th} simulated population $\underline{s}_b^2 = (s_b^0, s_b^1, s_b^2)$ and $\overline{\hat{Y}}_B = B^{-1} \sum_b \hat{Y}_b$.

To test the method I suggest the following simulation study. I start with an identical population as in section 3.6; A two phase process described by the unit selection indicators $\underline{\mathbf{s}}^2 = (\mathbf{s}^1, \mathbf{s}^2)$ with distribution over a population defined by $\underline{\mathbf{X}}_{s^0}^2 = (\mathbf{X}_{s^0}^{0\prime}, \mathbf{X}_{s^0}^1, \mathbf{X}_{s^0}^2)'$ with $\mathbf{X}_{s^0}^0 = (\mathbf{X}_{1s^0}^0, \mathbf{X}_{2s^0}^0, \mathbf{X}_{3s^0}^0)'$ and $\mathbf{X}_{s^0}^2 = \mathbf{Y}_{s^0}$. The population variables $\mathbf{X}_j^0$ $(j = 1, 2, 3)$ were generated independently from a standard normal, the distribution of $\mathbf{x}^1$ from $N(m_0(\mathbf{x}^0, \boldsymbol{\beta}_0), 1)$ and $\mathbf{y}$ from $N(m_1(\underline{\mathbf{x}}^1, \boldsymbol{\beta}_1), 1)$ are presented in table 3.3. The population size is fixed to $N = 30,000$.

The Bootstrap variance estimator was tested on seven different two phase selection processes which define s^1 and s^2. Both selection mechanisms follow a logistic regression model such as in section 3.6 and represent a wide variety of possibilities, see table 3.6. Note that the first two cases represent in fact a one phase scenario as all members of s^1 have a probability of selection equal de facto to one. The other five settings include the entire spectrum of relationship between the first and the second phases of selection. In all cases I aimed to achieve a survey sample size of $n_2 \approx 200$. I denote the seven scenarios by capital letters A to G with average achieved subset sizes of $A : n_1 = n_2 = 200$; $B : n_1 = n_2 = 225$; $C : n_1 = 2,074$, $n_2 = 207$; $D : n_1 = 3,548$, $n_2 = 222$; $E : n_1 = 3,548$, $n_2 = 206$; $F :$

Selection Models		
$\pi_1(\mathbf{x}^0, \boldsymbol{\alpha}_0)$	$\boldsymbol{\alpha}_0[1, \mathbf{x}^0_1, \mathbf{x}^0_2, \mathbf{x}^0_3, \mathbf{x}^0_1 \cdot \mathbf{x}^0_2]'$	
$\pi_2(\underline{\mathbf{x}}^1, \boldsymbol{\alpha}_1)$	$\boldsymbol{\alpha}_1[1, \mathbf{x}^0_1, \mathbf{x}^0_2, \mathbf{x}^0_3, \mathbf{x}^0_1 \cdot \mathbf{x}^0_2, \mathbf{x}^0_3 \cdot \mathbf{x}^1]'$	
Scenario	Parameters	Description
A	$\boldsymbol{\alpha}_0 = [-5, 0, 0, 0, 0]$	$p(s^1) = srs$
		$p(s^2\|s^1) = census$
B	$\boldsymbol{\alpha}_0 = [-5.5, 1, 0, 0, 0)]$	$p(s^1) = \pi_1(\mathbf{x}^0)$
		$p(s^2\|s^1) = census$
C	$\boldsymbol{\alpha}_0 = [-2.6, 0, 0, 0, 0]$	$p(s^1) = srs$
	$\boldsymbol{\alpha}_1 = [-2.2, 0, 0, 0, 0, 0, 0]$	$p(s^2\|s^1) = srs$
D	$\boldsymbol{\alpha}_0 = [-2.6, 1, 0, 0, 0]$	$p(s^1) = \pi_1(\mathbf{x}^0)$
	$\boldsymbol{\alpha}_1 = [-3.5, 1, 0, 0, 0, 0, 0]$	$p(s^2\|s^1) = \pi_2(\mathbf{x}^0)$
E	$\boldsymbol{\alpha}_0 = [-2.6, 1, 0, 0, 0]$	$p(s^1) = \pi_1(\mathbf{x}^0)$
	$\boldsymbol{\alpha}_1 = [-3.4, 0, 1, 0, 0, 0, 0]$	$p(s^2\|s^1) = \pi_2(\mathbf{x}^1)$
F	$\boldsymbol{\alpha}_0 = [-3.2, 1, 1, 0, 0]$	$p(s^1) = \pi_1(\mathbf{x}^0)$
	$\boldsymbol{\alpha}_1 = [-4.3, 0, 1, 0, 1, 0, 0]$	$p(s^2\|s^1) = \pi_2(\underline{\mathbf{x}}^1)$
G	$\boldsymbol{\alpha}_0 = [-2.4, 1, 1, -1, -1]$	$p(s^1) = \pi_1(\mathbf{x}^0)$
	$\boldsymbol{\alpha}_1 = [-3.4, 1, 1, 0, -1, 0, -2]$	$p(s^2\|s^1) = \pi_2(\underline{\mathbf{x}}^1)$

Table 3.6: A description of the seven simulation scenarios testing. The two selection models are given in the top of the table, each representing a logistic regression model. The seven scenarios differ in the coefficient parameter values given in the middle column.

$n_1 = 3,516$, $n_2 = 205$; $G : n_1 = 3,590$, $n_2 = 212$. The Bootstrap estimators were calculated over 500 runs. The number of simulations to test these estimators was chosen to be 500 as well.

Looking at the results in table 3.7, it is useful first to comment on the three estimators properties over the scenarios. The three estimators give, as expected, consistent estimates, although this is slightly less evident for the $\pi-$estimator which can be attributed to the relatively small number of simulations and relatively large variance. The variance of the $\pi-$estimator is substantially larger than the πm-estimator which in turn is slightly larger (or equal) to that of the $m-$estimator.

More subtle and interesting is the significant difference in the variance across all three estimators between the one phase (A,B) and two phase (C to G) cases. For the $\pi-$estimator the two phase scenarios have a slightly higher variance which is related to the additional selection phase; note here that for the π-estimator scenario C is also a one phase case- the two consecutive srs designs can be considered also a one phase srs. On the other hand for both m and πm-estimators there is a notable drop in the variance from the one phase to the two phase case. This however, can be predicted from the linear presentation of the πm and m-estimators. Start with the general formulation of the πm-estimator $\hat{\bar{\mathbf{y}}}_{\pi m} = N^{-1}\mathbf{1}'_{s^0}\underline{\mathbf{H}}^{T-1}_{\pi s^{T-1} s^T}\mathbf{Y}_{s^T}$, for a one phase case the estimator is $N^{-1}\mathbf{1}'_{s^0}\mathbf{H}^0_{\pi s^0 s^1}\mathbf{Y}_{s^1}$ while for a two phase case the estimator is $N^{-1}\mathbf{1}'_{s^0}\mathbf{H}^0_{\pi s^0 s^1}\hat{\mathbf{m}}_{1s^1}$.

	$B(\hat{\bar{\mathbf{y}}}_{\pi})$	$B(\hat{\bar{\mathbf{y}}}_{m})$	$B(\hat{\bar{\mathbf{y}}}_{\pi m})$	$V(\hat{\bar{\mathbf{y}}}_{\pi})$	$B(\hat{V}_{\pi})$	$V(\hat{\bar{\mathbf{y}}}_{m})$	$B(\hat{V}_{m})$	$V(\hat{\bar{\mathbf{y}}}_{\pi m})$	$B(\hat{V}_{\pi m})$
A	0.8%	0.3%	0.3%	1.75	-2.1%	0.58	-1.6%	0.57	-2.0%
B	-1.3%	0.3%	0.1%	1.88	-3.6%	0.59	-2.3%	0.58	-2.7%
C	3.9%	-0.2%	-0.2%	1.84	1.5%	0.18	-4.7%	0.18	-4.6%
D	3.0%	0.1%	-0.1%	2.60	0.7%	0.15	-5.6%	0.17	-7.2%
E	2.8%	0.1%	-0.1%	2.09	-2.3%	0.15	-0.4%	0.16	-10.4%
F	-0.6%	-0.1%	-0.2%	3.90	1.9%	0.17	2.1%	0.20	-2.3%
G	1.0%	-0.4%	0.0%	2.14	-0.4%	0.15	-1.8%	0.20	3.7%

Table 3.7: Simulation Results: Estimating the variance of the three Estimators. In the first three columns, $B(\cdot)$ indicates the average difference percentage between the estimator and the true population average over 500 simulations. In the remaining columns the bias denotes the average difference between the variance estimator and the variance calculated over the 500 simulations for each specification scenario. For each simulation 500 resampling with replacement we used to calculate the Bootstrap variance estimator. Bias percentage is the percentage average bias in terms of the simulated variance. $A : n_1 = n_2 = 200$; $B : n_1 = n_2 = 225$; $C : n_1 = 2,074,\ n_2 = 207$; $D : n_1 = 3,548,\ n_2 = 222$; $E : n_1 = 3,548,\ n_2 = 206$; $F : n_1 = 3,516,\ n_2 = 205$; $G : n_1 = 3,590,\ n_2 = 212$;

As the hat matrix $\mathbf{H}^0_{\pi s^0 s^1}$ and size of s^1 is the same for both cases the fundamental property of regression to the means results in the predicted values $\hat{m}_1$over the set s^1 having a smaller variation than the direct observations y we use in the one phase case.

Finally, looking at the the bias of the bootstrap variance estimators, the results are promising and merit further inspection in future work. As I note earlier, a direct bootstrap estimator does give consistent results for finite population statistics when the selection process is one where individual unit selection probabilities are independent - with unequal or equal probabilities. Although in our case there is a hierarchical structure, still within each phase of the sequential process here the unit selection probabilities are independent and so it seems the resampling algorithm I suggest above is sufficient to capture the variability in the sampling distribution of the statistics.

3.6.3 The Fallacy of Coding Non Internet Users with Zero

Throughout the discussion over this and the preceding chapter, a reader may be tempted to suggest using Internet related covariates by simply coding a value zero for members of the 'non Internet' population. They then treat these covariates as general population covariates which can be used in any of the estimation strategies we have discussed. This is common practice in research organizations. This sort of imputation is simple and intuitively appealing. However, in the following I give a stylized example to show that even when the true values of these x^1 covariates are indeed close to zero for non Internet members estimators can be badly biased.

Consider a survey estimating the population average of Y_k, unit k's answer to '*How engaged are you in the current US presidential campaign?*'. Assume we observe the replies $\mathbf{Y}_{s^2}$ as well as $\mathbf{X}^1_{s^1}$ and $\mathbf{X}^0_{s^0}$

$-X^0_k$ the annual income of member k of the population, standardized.
$-X^1_k$ frequency (hours) of accesing news or entertainment websites.

where $s^2 \subseteq s^1$ are the Internet and the Panel population subsets respectively and are realized each following a logit regression model

$$\begin{aligned}
\pi_1(\mathbf{x}^0, s^0):\ p(s^1_k = 1|\mathbf{x}^0, s^0) &= (1 + e^{-0.5x^0_k})^{-1} \quad \text{and} \\
\pi_2(\underline{\mathbf{x}}^1, s^1):\ p(s^2_k = 1|\underline{\mathbf{x}}^1, s^1) &= \begin{cases} (1 + e^{5-0.5x^0_k-0.5x^1_k})^{-1} & \text{for } s^1_k = 1 \\ 0 & \text{for } s^1_k = 0 \end{cases}
\end{aligned}$$

To keep the example within a fixed population setting[11], I introduce W^0_k, a set of covariates (here univariate) not measured but associated with the use of Internet. Let $(X^0_k, W^0_k) \sim N\begin{pmatrix} 0 & 1 & 0.3 \\ 0 & 0.3 & 1 \end{pmatrix}$ for all $k \in s^0$. And $X^1_k = e^{W^0_k}$;$\forall k \in s^1$ which means that X^1 over s^1 has a Log-normal distribution. Further, assume (as is frequently the case in practice) that the analyst wrongly imputes the value $X^1_k = 0$ rather than assigning $X^1_k = NA$ for all non internet members of the population $\overline{s}^1$. Finally, let

$$Y_k = 1 + x^0_k + 2w^0_k + \varepsilon_k \ ; \varepsilon_k \sim (0, 1).$$

I compare 6 estimators, three $\pi-$estimators and three m estimators. The π estimators include one sequential π- estimator which models separately s^1 and s^2 using $\mathbf{X}^0_{s^0}$ and $\underline{\mathbf{X}}^1_{s^1}$ respectively, and two single phase estimators: The first $\hat{\overline{\mathbf{Y}}}_{\pi x^0}$ models s^2 over s^0 using $\mathbf{X}^0$ only; The second $\hat{\overline{\mathbf{Y}}}_{\pi \underline{X}^1}$ models s^2 over s^0 using both $\mathbf{X}^1$ and $\mathbf{X}^0$. These are described in the following table.

Estimators	Notes
$\hat{\overline{\mathbf{Y}}}_{\pi\underline{2}} = \sum_{s^2} Y_k(\hat{\pi}_{1k} \cdot \hat{\pi}_{2k})^{-1}$	$\hat{\pi}_{1k}$, $\hat{\pi}_{2k}$ are fitted by logistic regression
$\hat{\overline{\mathbf{Y}}}_{\pi x^0} = \sum_{s^2} Y_k \hat{\pi}_{x^0 k}{}^{-1}$	$\hat{\pi}_{x^0 k} : logit^{-1} \left[p(s^2_k = 1|\mathbf{x}^0, s^0)\right].$
$\hat{\overline{\mathbf{Y}}}_{\pi \underline{x}^1} = \sum_{s^2} Y_k \hat{\pi}_{\underline{x}^1 k}{}^{-1}$	$\hat{\pi}_{\underline{x}^1 k} : logit^{-1} \left[p(s^2_k = 1|\underline{\mathbf{x}}^1, s^0)\right].$

While the sequential estimator $\hat{\overline{\mathbf{Y}}}_{\pi\underline{2}}$ is clearly unbiased, we expect the second $\hat{\overline{\mathbf{Y}}}_{\pi x^0}$ to be only partially beneficial- as $\mathbf{X}^0$ will explain some of the selection

[11] By which I mean here that selection (is random, but) does not effect population quantities such as $\overline{\mathbf{Y}}^0_{s^0}$or $\overline{\mathbf{X}}^0_{s^0}$. In other words $\overline{\mathbf{Y}}^0_{s^0}$or $\overline{\mathbf{X}}^0_{s^0}$will not vary with different selection patterns of the population.

process. Estimator $\hat{\bar{\mathbf{Y}}}_{\pi \underline{X}^1}$ uses $\mathbf{X}^1$ which over s^0 is misspecified, so we expect also a biased estimator. In our set up this is reflected by a failure to meet the common support requirement which is best shown by the non overlapping distribution of $\mathbf{X}^1$ in the final graph panel below[12].

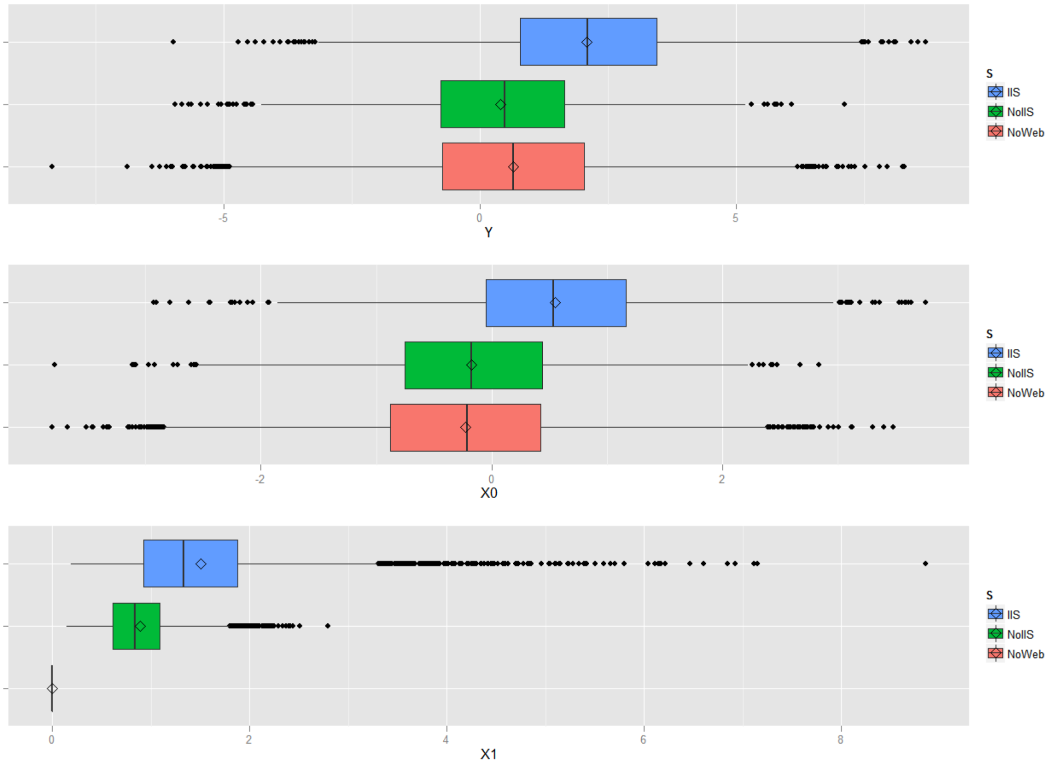


Figure 3.7: Boxplots representing the distribution of Y,X^0 and X^1over the achieved sample, here denoted IIS, the non-sampled population denoted $NoIIS$, and the subpopulation $NoWeb$ of non-Internet connected units of the population respectively. From the bottom panel describing X^1it is clear that it is defined only over the Internet-connected population.

Similarly, I calculate three m-estimators. A two phase regression estimator which first models $\mathbf{Y}_{s^2}$ against $\underline{\mathbf{X}}^1$, and then regresses the fitted values of the model against $\mathbf{X}^0$ over s^1. The other two estimators are simple regression estimators. Over s^2 the first models Y against X^0 and the second models Y against $\underline{\mathbf{X}}^1$. These three estimators are described in the table below.

Estimators	Notes
$\hat{\bar{\mathbf{Y}}}_{\underline{m2}} = N^{-1}\mathbf{1}'_{s^0}\mathbf{X}^0_{s^0}\hat{\boldsymbol{\beta}}_0$ $= N^{-1}\mathbf{1}'_{s^0}\mathbf{H}_{s^0 s^1}\underline{\mathbf{X}}^1_{s^1}\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_1 = (\underline{\mathbf{X}}^{1'}_{s^2}\underline{\mathbf{X}}^1_{s^2})^{-1}\underline{\mathbf{X}}^{1'}_{s^2}\mathbf{Y}_{s^2}$ and letting $\hat{Y}_{m_1 k} = \underline{\mathbf{X}}^1_k\hat{\boldsymbol{\beta}}_1$ $\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}^{0'}_{s^1}\mathbf{X}^0_{s^1})^{-1}\mathbf{X}^{0'}_{s^1}\hat{\mathbf{Y}}_{m_1 s^1}$
$\hat{\bar{\mathbf{Y}}}_{mx^0} = N^{-1}\sum_{s^0}\hat{Y}_{mx^0 k}$	$\hat{Y}_{mx^0 k} = \hat{\alpha} + \hat{\boldsymbol{\beta}}\mathbf{X}^0_k + \hat{\gamma}(\mathbf{X}^0_k)^2$
$\hat{\bar{\mathbf{Y}}}_{m\underline{x}^1} = N^{-1}\sum_{s^0}\hat{Y}_{m\underline{x}^1 k}$	$\hat{Y}_{m\underline{x}^1 k} = \hat{\alpha} + \hat{\boldsymbol{\beta}}\mathbf{X}^0_k + \hat{\gamma}(\mathbf{X}^0_k)^2 + \hat{\boldsymbol{\beta}}\mathbf{X}^1_k + \hat{\gamma}(\mathbf{X}^1_k)^2$

[12] if $(\mathbf{X}^0, \mathbf{X}^1)$ were factors, this lack of common support would result in estimated probabilities equal to zero for all members of $\overline{s}^1$- immediately eliminating π- estimation strategy.

The models for $\hat{\overline{\mathbf{Y}}}_{mx^0}$ and $\hat{\overline{\mathbf{Y}}}_{m\underline{x}^1}$ were found by simple goodness of fit measurements given available covariates. Equivalent to π-estimator we expect $\hat{\overline{\mathbf{Y}}}_{\pi x^0}$ to be only partially effective in removing the bias. As for $\hat{\overline{\mathbf{Y}}}_{\underline{m2}}$ and $\hat{\overline{\mathbf{Y}}}_{mx^1}$ the difference is more subtle. In the graph below the blue dots represent the observed marginal scatter between Y and $\mathbf{X}^1$. The best fit of the observed data is the green line which is a logarithmic regression and this will be the first model in the two phase $m-$estimator. However, when taking a one phase $m-$estimator approach the analyst must consider that $\mathbf{X}^1_k$ is equal to zero for non Internet members of the population. To make this idea clear I added in the plot red dots representing possible range of values y for the non internet population members.

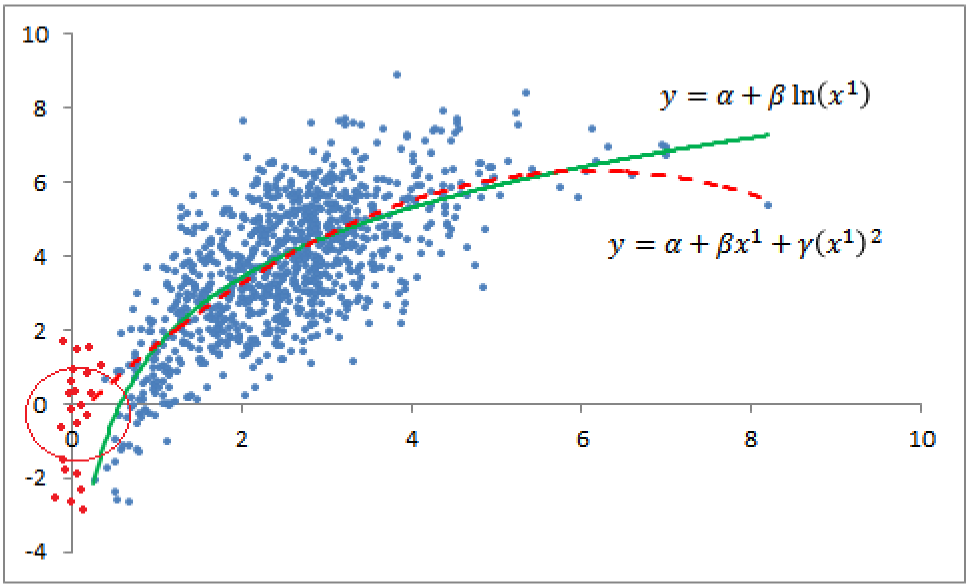


Figure 3.8: A scatter plot describing the stylized example where an analyst taking a one phase modelling perspective and coding zero values X^1 of non internet units will miss-specify the true model and assume a polynomial relationship, here denoted by the red dots.

Given $\mathbf{x}^1_k = 0$ for $\overline{s}^1$ the analyst will give up the logarithmic model which is thus undefined. The best alternative is a very reasonable polynomial model which is displayed by the red dotted line above.

To test these ideas we created a a population of size $N = 20,000$. And simulated a two phase selection process. The relevant population profiles and averages of 1,000 simulations are $\overline{X}^0_{s^0} = 0.01$, $\overline{Y}_{s^0} = 1.01$, $p(s^1) = 0.51$, $p(s^2|s^1) = 0.05$, $\overline{Y}_{s^1} = 1.39$ and $\overline{Y}_{s^2} = 3.90$.

the selection process causes a strong positive bias in the observed sample average of the outcome variable The following two density plots and associated summary tables confirm our earlier discussion: The sequential estimators are unbiased, while the two other sets of estimators are biased to different degrees.

Estimators	$E(\cdot)$	$V(\cdot)$
$\hat{\bar{\mathbf{Y}}}_{\underline{\pi 2}}$	1.01	0.07
$\hat{\bar{\mathbf{Y}}}_{\underline{\pi x^0}}$	2.22	0.06
$\hat{\bar{\mathbf{Y}}}_{\underline{\pi \underline{x}^1}}$	0.90	0.06
$\hat{\bar{\mathbf{Y}}}_{\underline{m 2}}$	0.99	0.00
$\hat{\bar{\mathbf{Y}}}_{\underline{m x^0}}$	2.12	0.01
$\hat{\bar{\mathbf{Y}}}_{\underline{m x^1}}$	0.47	0.01

Table 3.8: The estimated expectation and variance of the three $\pi-$estimators and three $m-$type estimators based on 1,000 simulation over a population of 20,000 units. Additional grpahs of the distribution of the six estimators are presented in section 5.3.

The short simulation study above intended to show in a simple example the error of using Internet related adjustment covariates in a one phase framework. While it is true that in very specific scenarios such estimators are unbiased, Rosenbaum (1984) puts it nicely that in general these estimators are justified only when they are unnecessary- such as a when a simple linear relationship exists between Y and $\mathbf{W}^0$ and between $\mathbf{X}^1$ and $\mathbf{W}^0$ or under unrealistic restrictions. Under a sequential framework this issue is completely avoided.

Chapter 4

The Application of the Sequential Framework in Practice

4.1 The Reference Survey and the Role of its Sampling Weights

4.1.1 Introduction, Basic Set- up and Introductory Notation

The panel of persons who volunteer to participate in Web surveys are used to make estimates for entire populations, including segments who have only a small theoretical chance to participate in the surveys. In our case these are the population subset of rare or infrequent Web users- A segment which in some populations may be very large. A successful estimation method will allow inference based on panel data to be representative of the entire target population. In the previous chapter I have outlined a general framework under which three competing estimation procedures- a π, m and πm estimator class type have been proposed. As discussed, a key strength in the underlying sequential structure of the framework allows to directly use covariates in the estimation process which otherwise would not be utilized as they invalidate the conditional independence assumption. However, throughout this discussion has been underpinned by the unrealistic assumption that the practitioner observes data which in practice is unlikely to be available. In this chapter I will translate these theoretical ideas into a realistic setting of a finite population survey sampling inferential problem based on survey data collected from a Web access panel and a reference random sample.

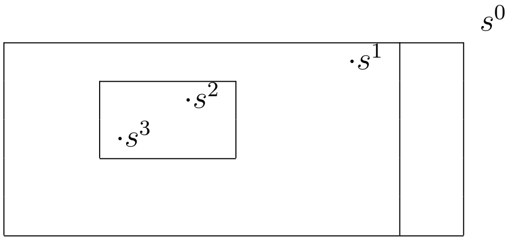
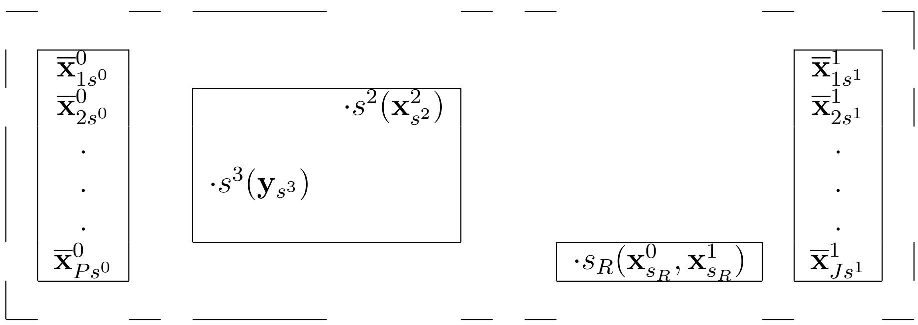


Figure 4.1: A graphical description of the three subsets of interest s^1, s^2 and s^3 within the frame of the finite population s^0. In the Web panel context these represent respectively the finite population (s^0), the Internet active subpopulation (s^1), the Web access panel volunteers (s^2) and the survey sample (s^3) .

The sequential framework on which the three sets of estimators I proposed can be visualized by the diagram in figure 4.1 where in the Web panel survey sample context s^0, s^1, s^2 and s^3 represent respectively the unit sets of the finite population members of interest (s^0), the Internet active members in the population (s^1), the Web access panel volunteers (s^2) and the survey sample respondents (s^3) from which measurement data is recorded. Throughout we have assumed that the individual membership for all units $k \in s^0$ in each set s^t, $t = 0, .., 3$ are fully observed along with the unit values of of variables $\mathbf{X}^t$ denoted $\mathbf{x}^t_{s^t}$. Earlier I outlined the different assumptions on the selection process leading to $\underline{s}^3 = (s^0, s^1, s^2, s^3)$ or the distribution of the covariates $\underline{\mathbf{X}}^2_{s^2} = (\mathbf{X}^0_{s^0}, \mathbf{X}^1_{s^1}, \mathbf{X}^2_{s^2})$ which allow a sequential version of the conditional independence assumption to hold. Under these assumptions inference on general population quantities of the measurements of interest $\mathbf{Y}$ can be derived from the survey sample collected data $\mathbf{y}_{s^3}$ by any of the three estimation approaches.

In practice however, the practitionaire is likely to observe only the subset $\underline{\mathbf{x}}^2_{s^2} \cap s^3 = (\mathbf{x}^0_{s^3}, \mathbf{x}^1_{s^3}, \mathbf{x}^2_{s^3})$ of the covariate information, that is the values of covariates $\underline{\mathbf{x}}^2$ over the survey sample members. In more specific cases where the analyst is a member of the panel owner the set of values $\underline{\mathbf{x}}^2$ over the entire panel members s^2 may be available to process as well. To estabilish a link to the population of interest the practionaire may use available population statistics (averages or totals) of some *but not necessarily all* of $\mathbf{x}^0_{s^0}$and $\mathbf{x}^1_{s^1}$, taken from external sources such as census data or established large social surveys. Such a constraint on the available data undermines the already strong theoretical assumptions and ultimately limits the range of estimators and model fitting process. A good example is the strong reliance in survey sampling practice on the GREG estimator and the linear models underlying it, in large parts because it requires only population level statistics for estimation.

To strengthen the plausibility of the conditional independence assumption and broaden the range of estimation approaches, practitioners and academics (Taylor et al. (2001); Terhanian et al. (2000); Varedian and Forsman (2003)) suggest combining the panel sample survey data and available population statistics with information collected over a reference sample denoted by s_R randomly selected from the target population.



$\overline{\mathbf{x}}^0_{1s^0}$
$\overline{\mathbf{x}}^0_{2s^0}$
$\overline{\mathbf{x}}^0_{Ps^0}$
$\cdot s^2(\mathbf{x}^2_{s^2})$
$\cdot s^3(\mathbf{y}_{s^3})$
$\cdot s_R(\mathbf{x}^0_{s_R}, \mathbf{x}^1_{s_R})$
$\overline{\mathbf{x}}^1_{1s^1}$
$\overline{\mathbf{x}}^1_{2s^1}$
$\overline{\mathbf{x}}^1_{Js^1}$

Figure 4.2: A description of the possible data available to the statistician analyzing Web panel survey data. At best the analyst will observe the within panel collected data $(\mathbf{x}^0_{s^2}, \mathbf{x}^1_{s^2}, \mathbf{x}^2_{s^2}, \mathbf{y}_{s^3})$, reference survey data $(\mathbf{x}^0_{s_R}, \mathbf{x}^1_{s_R})$ and population statistics $\overline{\mathbf{x}}^0_{1s^0}, ..., \overline{\mathbf{x}}^0_{Ps^0}$and $\overline{\mathbf{x}}^1_{1s^1}, ..., \overline{\mathbf{x}}^1_{Js^1}$.

Thus under a more realistic limited information case the optimum an analyst in practice has at his disposal are the within panel collected data $(\mathbf{x}^0_{s^2}, \mathbf{x}^1_{s^2}, \mathbf{x}^2_{s^2}, \mathbf{y}_{s^3})$, the reference survey collected data $(\mathbf{x}^0_{s_R}, \mathbf{x}^1_{s_R})$ and population statistics $\overline{\mathbf{x}}^0_{1s^0}, ..., \overline{\mathbf{x}}^0_{Ps^0}$ and $\overline{\mathbf{x}}^1_{1s^1}, ..., \overline{\mathbf{x}}^1_{Js^1}$ a diagram summarizing the available data is in figure 4.2. Ideally the random reference sample is designed to collect specific data which is considered relevant to the Web panel survey problem- This can be from an $\pi-$estimation perspective such as the one taken by Harris Interactive which coined the phrase *webographics* (Schonlau et al., 2007) for such covariates that explain the process of panel volunteering, or from an m-estimation view where then covariate will change on a survey-by-survey basis. However, in general any large well designed random survey of the entire general population may be considered potentially useful. The benefit in obtaining such a dataset is clear- the range of models $\pi_t(\underline{\mathbf{x}}^{t-1})$ and $m_t(\underline{\mathbf{x}}^{t-1})$ which may be tested now are not constraint by the limited pool of potential covariates available from secondary sources such as national statistic databases, nor are we limited in applying only models consistent with the use of (only) aggregate level data for inference.

To allow a model to be estimated, the same set of covariates must be observed for both the web panel and reference samples. As noted by Lee and Valliant (2009), the covariates should be measured in similar way in both surveys (e.g., identical question wording). Measurement differences for some factual insensitive covariates such as age, sex, and perhaps race may be inconsequential even if different modes are used in the surveys. But, for more complex socioeconomic measures like income and assets, incompatibilities may be serious. This can lead to biased estimates.

As noted in previous section, the choice of an Internet or Web available segment of the population is crucial in enhancing the inference power of the estimation procedure when assuming a sequential framework. Thus the definition of membership into the Web available segment of s^0 and panel membership needs to be considered carefully so that any discrepancies between these definitions between the two survey samples may cause serious measurement error in the estimation. The interpretation of the term “Web available” is arbitrary and can vary among surveys depending on operational decisions such as the survey measurements used and their exact wordings. “Web available” could mean that a person has Web access at home or it could mean access anywhere, including home, work, libraries, or other locations. In developed countries with high Internet penetration “Web Usage” or “Web Active” can be used alternatively assuming that the practical barrier for panel recruitment requires a certain minimum level of web browsing or Online usage. The analyst must specify the exact term used which is largely a function of the specific set of data available.

In the following I cover several issues related to the application of the structured Web panel estimation framework through. I start in sections 4.1.2 and 4.1.3 by looking at the coverage of the reference survey and the use of the reference surveys's sample design weights in estimating model parameters and population statistics. I then cover in length over section 4.2 the mechanics of calculating estimators over the combined data of the reference and Web panel survey samples. In section 4.3 I describe several simulation studies to demonstrate the basic algorithm and show some possible implications of not using the Web panel or reference sample survey weights. In section 4.3.2 I consider the question of sampling s_R and s^3, specifically the sample sizes of each survey and show that conclusions differ when considering m or π-based estimators. In section 4.4 I combine the previous estimation procedure with a sequential $\pi-$balanced and calibration sampling design. I show that together with a $\pi m-$estimator, this offers an estimation strategy which is double robust as well as robust to $m-$model misspecification. I conclude this chapter with some conclusions, recommendations for future work and some final words on this research.

4.1.2 A Look at the Universe Coverage of the Reference and Web Panel Survey Samples

The reference survey's utility as a surrogate for the unobserved general population distribution is directly linked to its design and any deficiencies it has vis a vis the target population (Couper et al., 2001; Couper, 2008). The assumption a reference survey covers probabilistically all segments of the population is especially questionable for data collected by commercial entities given questionable design practice and low response rates. Valliant and Dever (2011) give a useful taxonomy of the problem in the context of propensity score adjustments for Web panel surveys which I build on to our sequential framework.

The target population s^0 can be devided into subsets covered and not covered by the reference survey sampling frame, the Web active subpopulation, the Web panel and the final set of sampled survey participants. This means that to identify how much of a population we can validly make inferences for, we need to clarify this partition from the outset. Let s^1 and $\overline{s}^1$ be the set of population members defined by their availability (or not) to Internet web browsing. The set of Web panelists is denoted by s^2 and the reference survey sample is denoted by s_R drawn from s^0_R a subset of s^0. Together, these describe the group of relevant subpopulations. For example $s^0 = s^0_R \cup s^0_{\overline{R}}$, where $s^0_{\overline{R}}$ is the population not covered by the reference sample, while the Web connected subpopulation which is not covered by the reference survey sampling frame is $s^1_{\overline{R}}$. A diagram describing the entire multitude of sub populations is given in 4.3

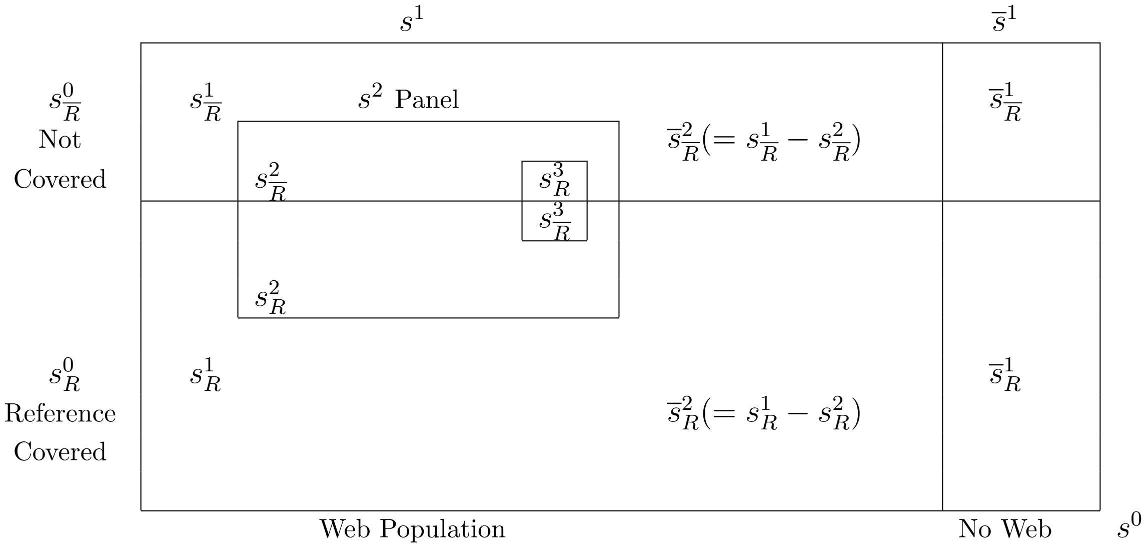


Figure 4.3: A description of the different sub populations defined by the two selection phases leading to panel participation indicated by the numerical superscript and the reference survey sampling frame coverage indicated by the subscript R. For example: The general population covered by the reference survey sampling frame is s^0_R while the Internet connected subpopulation which is not covered by the reference survey sampling frame is $s^1_{\overline{R}}$. The non Internet connected population which is not covered by the reference survey sampling frame is $\overline{s}^1_{\overline{R}}$ while that which is covered by reference sampling frame is $\overline{s}^1_R$.

Of course the discrepancy between s^0_R the covered population of s_R compared with s^0 is a function of the survey sampling process applied, the population of interest, the survey topic and even the institution conducting the survey Groves et al. (2004). In principle, area or face to face (F2F) sampling gives the best coverage of s^0, however, as Valliant and Dever (2011) note, even in well designed and executed area probability samples certain groups by as much as 25 percent

(U.S. Census Bureau, 2002,). In the UK, official government social surveys such as Labour Force Survey (LFS) and the British Crime Survey (BCS) use as a sampling frames the Postcode Address File (PAF)Lynn et al. (1999) which entirely excludes certain parts of the population[1] Pickering (2008). More broadly there is a steadily increase of non cooperation with survey interviewers- The LFS reports response rates decline from 80.02% in 1993 to 57% in 2012 Steel (2012), with some governmental regions - London and West Midlands Metropolitan Council- achieving response rates of just over 50% Barnes et al. (2008). For large reputable commercial surveys the state is even worse. For example the National Readership survey (NRS) reports a proportion of sample successfully contacted in the first place is 84% (in 2010) and the proportion of identified respondents who complete an interview is 61% (in 2010) which suggests an overall response rate, which in 2010 was 52%. In London the response rate is just over 40% (NRS and Ipsos MORI official website, 2010) .

More concerning is that area or F2F sampling is dramatically more expensive than other survey sampling methods. The American Community Survey which includes census-type demographic questions (household and person level) is collected by a mixed mode design allowing a good basis for comparison. The cost for a completed mail form is $10, a completed computer-assisted telephone interviewing (CATI) survey coasts $18 per interview while a computer-assisted personal interviewing (CAPI) survey costs $140 per interview. In the UK the BCS reports a unit cost of 100 ($162) per achieved case (Betts, 2010).

However, basing the reference survey on a cheaper sample of landline telephones introduces substantial coverage problems as $s^0_{\overline{R}}$ will be the segment of the population which cannot be reached by a landline phone either because they do not have a phone at all or, as is increasingly the case are members of a house hold which use only mobile/wireless phones. This segment can be large. For example, in the UK National Readership Survey (NRS), in the period January through June 2010, found that 45.7 percent of persons 25 to 29 years old lived in households with only mobile phones, while as a whole 7% of UK households are without a fixed landline Betts (2010).

The coverage level of s_R is reflected in its unit selection probability denoted here π_{Rk}. This probability may be unequal due to clustering, stratification, post-stratification, attrition, purposive oversampling and other non-response adjustments. Depending on the survey objective, data and modeling constraints π_R are constructed to project the sample to s^0 or limited to the subset s^0_R (see 4.3). Still, to focus the discussion on the Web panel question I assume throughout that π_R are known design probabilities that allow total population s^0 coverage. Clearly,

[1] For example, the communal establishment establishment resident population made up about 2.1 per cent of the adult population at the time of the 2001 Census

that the reference survey itself needs adjustment is problematic, however, a mitigating argument may be that problem of non sampling errors in classic survey designs are well researched. Nevertheless, when discussing the calculation of the estimators I will occasionally make suggestions on the composition of the auxiliary data used for the reference survey adjustment models to increase the overall robustness of the combined approach.

Lastly note that strictly speaking the reference survey probabilities should be adjusted to inflate to $s^0 - s^2$ or $s_R^0 - s_R^2$, that is respectively, the covered or general population excluding the Web panel segment. This adjustment removes the overlap of persons who are in panel s^2 as well as the reference sample which is necessary for selection models to be estimated cleanly (Valliant and Dever, 2011). Similar to the situation in an observational study where the treated and the non treated cases are disjoint. Typically, however, s^2 is so small relative to s_R^0 or s^0 that this further adjustment is unnecessary. For example, if population frame consists of the approximately 49 million adults (16+) in the UK who have either a landline or a mobile phone, and a large volunteer panel has 300 thousand adult panelists, all of whom have a telephone, then the panel would constitute[2] only 0.0057 percent of the population.

4.1.3 The Role of Sampling Weights When Modeling Survey Data

The mechanically combined data of panel and reference survey data are used to implement in practice the different estimation strategies introduced in the previous chapter. The basic idea I follow for each of the estimation procedures is of estimating by inverse probability weighting (IPW) over the available dataset the estimation algorithms which assumed full information over the three populations subsets $\underline{s}^3$.

To understand the role of the weights take for simplicity a two phase π-estimator, while conclusions can be then applied easily to a three phase π, m or πm approach. Consider a two phase framework where first s^2 of panel members is drawn from s^0, followed by a selection of an ad hoc survey sample s^3 from the panel. In addition a parallel sample s_R is selected from the population. Throughout assume that labels s^2, s^3 and s_R and their associated measurements are fully observed. For estimation we combine the datasets of panelists and the random survey respondents into one set denoted s, that is $s = s^2 \cup s_R$. The dataset s is used primarily for estimating $Pr(S_k^2 = 1|\mathbf{X}_k^0 = \mathbf{x}_k, s^0)$ as I assume $p(s^3|s^2, \underline{\mathbf{x}}^2)$

[2] Note, however, that there will be overlap between the volunteer subuniverse and the reference universe; that is, s_R^1 in Figure 4.3 is not empty. For example, if s_R^0 consists of persons with landline phones, there will be some who volunteer to be part of a panel.

the within panel survey selection process, where $\underline{\mathbf{X}}^2 = (\mathbf{X}^0, \mathbf{X}^2)$ and $\mathbf{X}^2$ measuring panel related information, either follows closely its planned design or that the panelists non response element is well understood- a reasonable assumption. Under such setting the $\pi-$Estimator is

$$\hat{\overline{\mathbf{Y}}}_\pi = \sum_{k\in s^3} Y_k \hat{\pi}_{2k}(\mathbf{x}^0)^{-1}\pi_{3k}^{-1} / \sum_{k\in s^3} \hat{\pi}_{2k}(\mathbf{x}^0)^{-1}\pi_{3k}^{-1} \tag{4.1}$$

where $\hat{\pi}_2$ is the estimated unit probability of joining the panel and π_{3k} is assumed known for all panelists. The question here is the role of reference survey weights π_R in building $\hat{\overline{\mathbf{Y}}}_\pi$.

If pairs $(s_k^2, \mathbf{x}_k^0)$ were observed over all s^0 then by considering a logistic model $\pi_2(\mathbf{x}_k^0; \boldsymbol{\alpha}) = exp(\mathbf{x}_k^{0'}\boldsymbol{\alpha})/[1 + exp(\mathbf{x}_k^{0'}\boldsymbol{\alpha})]$ the parameters $\boldsymbol{\alpha}$ could be estimated through the equations

$$\mathbf{U}_{s^2}(s^0) = \sum_{k\in s^0}[S_k^2 - \pi_2(\mathbf{x}_k^0; \boldsymbol{\alpha})]\mathbf{x}_k^0 = \mathbf{0} \tag{4.2}$$

resulting with consistent estimators $\pi_2(\mathbf{x}_k^0; \hat{\boldsymbol{\alpha}}) = \hat{\pi}_{2k}$ of the unit participation probability. In practice however, we are limited to $s = s^2 \cup s_R$ the pooled sample assembled for estimation. Define a new indicator

$$S_k = \begin{cases} 1 & \text{if } k \in s \\ 0 & \text{if } k \notin s \end{cases}$$

defined over s^0 of whether unit k is available to the analyst. When s_R and s^2 are processed independently - possibly by separate organizations - this leaves the possibility that a unit k may be included in both surveys, however, as discussed earlier, the sizes of s_R and s^2 are most likely small relative to the general population so that

$$\begin{aligned} E(S_k) &= Pr(S_k^2 = 1) + Pr(S_{Rk} = 1) - Pr(S_k^2 \cdot S_{Rk} = 1) \\ &\approx \pi_{2k} + \pi_{Rk} \end{aligned} \tag{4.3}$$

which is understood that $Pr(S_k^2 \cdot S_{Rk} = 1) \approx 0$, that is for practical purposes no person can be selected for both panel and survey.

A naive approach is to estimate π_{2k} the unit selection probability into the Web panel over the set s directly without the use of the reference weights

$$\hat{\mathbf{U}} = \sum_{k\in s}[S_k^2 - \pi_{2s}(\mathbf{x}_k^0; \boldsymbol{\alpha}_s)]\mathbf{x}_k^0 = 0 \tag{4.4}$$

which in fact models S_k^2 in the combined sample s by a logistic model, that is $Pr(S_k^2 = 1|, s_k = 1, \mathbf{X}_k^0 = \mathbf{x}_k^0) = exp(s_k\mathbf{x}_k^{0'}\boldsymbol{\alpha}_s)/[1 + exp(s_k\mathbf{x}_k^{0'}\boldsymbol{\alpha}_s)]$. Denote the consistent selection probability estimates by $\hat{\pi}_{2sk} = \pi_{2s}(\mathbf{x}^0_k; \hat{\boldsymbol{\alpha}}_s)$.

To see the potential bias resulting in using $\hat{\pi}_{2sk}$ to estimate π_{2k} when calculating the π-estimator (4.1) assume that ignorability of s given $\mathbf{x}^0$ holds, and denote $E(Y_k|\mathbf{x}^0) = m_k(\mathbf{x}^0)$ then

$$\begin{aligned} E(\hat{\overline{\mathbf{Y}}}_\pi) &= E\left(\sum_{k\in s^0} S_k Y_k \underline{S}^3_k \hat{\pi}^{-1}_{2sk}\pi^{-1}_{3k} / \sum_{k\in s^0} S_k \underline{S}^3_k \hat{\pi}^{-1}_{2sk}\pi^{-1}_{3k}\right) \\ &\approx E_s\left\{n_s^{-1} E_{\mathbf{x}^0}\left(\sum_{k\in s} m_k(\mathbf{x}^0)|s\right)\right\} \\ &= n^{-1}E_s\left\{\sum_{k\in s^0} S_k m_k\right\} = n^{-1}\sum_{k\in s^0}(\pi_{Rk}+\pi_{2k})E(Y_k) \qquad (4.5) \end{aligned}$$

where the second line is arrived by iterative expectations and assuming $\hat{\pi}_{2sk}$ are consistent. Specifically, for fixed s

$$\begin{aligned} E\left(\sum_s Y_k \underline{S}^3_k \hat{\pi}^{-1}_{2sk}\pi^{-1}_{3k} / \sum_s \underline{S}^3_k \hat{\pi}^{-1}_{2sk}\pi^{-1}_{3k}\right) &\approx \frac{\sum_s E\{EE(Y_k \underline{S}^3_k \pi^{-1}_{2sk}\pi^{-1}_{3k}|\mathbf{x}^0, \mathbf{y}, s^2)|\mathbf{x}^0\}}{\sum_s E\{EE(\underline{S}^3_k \pi^{-1}_{2sk}\pi^{-1}_{3k}|\mathbf{x}^0, s^2)|\mathbf{x}^0\}} \\ &= \frac{\sum_s EE(Y_k S^2_k \pi^{-1}_{2sk}|\mathbf{x}^0)}{\sum_s EE(S^2_k \pi^{-1}_{2sk}|\mathbf{x}^0\}} = n_s^{-1}\sum_s EE(Y_k|\mathbf{x}^0_k). \end{aligned}$$

The last line in (4.5) is arrived by fixing $n_s = n_2 + n_R$ to be n which is reasonable in large sample sizes and noting from (4.3) that $E(S_k) = \pi_{2k} + \pi_{Rk}$.

To make things more tangible, assume that $p(s_R)$ is a simple random sample over s^0 and that $s^2_R = s^2 \cap s_R$ is negligible relative to s^0 so then even after an adjustment to the reference survey weights the unit selection probabilities are $\pi_{Rk} \approx n_R/N$ and so

$$n^{-1}\sum_{s^0}(\pi_{Rk}+\pi_{2k})E(Y_k) = \frac{n_R}{n}E(\overline{\mathbf{Y}}_{s^0}) + n^{-1}\sum_{s^0}\pi_{2k}E(Y_k).$$

Now, by noting that

$$E(n_2\overline{\mathbf{Y}}_{s^2}) = EE(\sum_{s^0} S^2_k Y_k|\mathbf{x}^0) = \sum_{s^0}\pi_{2k}E(Y_k)$$

and approximating n_2 to be constant, then the expectation of the π-estimator is

$$E(\hat{\overline{\mathbf{Y}}}_\pi) \approx \frac{n_R}{n}E(\overline{\mathbf{Y}}_{s^0}) + \frac{n_2}{n}E(\overline{\mathbf{Y}}_{s^2}). \qquad (4.6)$$

Thus for the common case where the Web panel set s^2 is negligible relative to s^0 but is much larger than s_R the reference survey $E(\hat{\overline{\mathbf{Y}}}_\pi) \approx E(\overline{\mathbf{Y}}_{s^2})$, the panel unadjusted average. Only when the reference survey is by a large order bigger than the panel survey then we can expect valid estimates. This is an example of the bias from the use of unadjusted estimating equations ignoring the reference survey selection weights. In the literature of case control studies this bias is corrected by a post hoc manual adjustment of the estimated intercept parameter.

As implied from the above discussion, the substantial bias underlying 4.6 can be removed by weighting the estimating equations (4.4) with the reference survey weights. To see this, for a given panel set s^2 and a randomly selected reference survey sample s_R define the individual inclusion probabilities

$$\pi_k = \begin{cases} 1 & k \in s^2 \\ \pi_{Rk} & k \in s_R. \end{cases} \tag{4.7}$$

Again, the definition of π_k is partial as it ignores the possibility that $k \in (s^2 \cap s_R)$, however, ignoring this small overlap the underlying idea is that by including π_k in the estimation equations, the available data s covers[3] s^0. Specifically, we estimate now π_{2k} by

$$\hat{\mathbf{U}}_\pi \;=\; \sum_{k \in s} \pi_k^{-1}[S_k^2 - \pi_2(\mathbf{x}_k^0; \boldsymbol{\alpha})]\mathbf{x}_k^0 = \mathbf{0} \tag{4.8}$$

which are approximately unbiased for the estimating equations over the entire populations given in (4.2). To see this note that

$$\begin{aligned} E(\hat{\mathbf{U}}_\pi - \mathbf{U}_{s^2}(s^0)) &= E\left(\sum_{k \in s^0}(S_k\pi_k^{-1} - 1)(S_k^2 - \pi_2(\mathbf{x}_k^0; \boldsymbol{\alpha}))\mathbf{x}_k^0\right) \\ &= EE\left(\sum_{k \in s^0 - s^2}(S_{Rk}\pi_{Rk}^{-1} - 1)(0 - \pi_2(\mathbf{x}_k^0; \boldsymbol{\alpha}))\mathbf{x}_k^0|s^2\right) \end{aligned}$$

which is equal to zero as $E(S_{Rk}\pi_{Rk}^{-1}|s^2) = 1$ given (i) independence between the processes of s^2 and s_R and (ii) ignoring the possible small overlap between s_R and s^2. Thus given conditional independence on $\mathbf{x}^0$ and under model $\pi_2(\mathbf{x}^0; \boldsymbol{\alpha})$

$$\begin{aligned} E(\hat{\overline{\mathbf{Y}}}_\pi) \;&=\; E_{\mathbf{x}^0}E\left(\sum_{k \in s^0} S_k^3 Y_k \hat{\pi}_{2sk}^{-1}\pi_{3k}^{-1} // \sum_{k \in s^0} S_k^3 \hat{\pi}_{2sk}^{-1}\pi_{3k}^{-1} /|\mathbf{x}^0\right) \\ \;&\approx\; E_{\mathbf{x}^0}\left(\sum_{k \in s^0} E(Y_k|\mathbf{x}^0)N^{-1}\right) = E(\overline{\mathbf{Y}}) \end{aligned}$$

as $\hat{\pi}_{2k}/\pi_{2k} \approx 1$ which holds as the size of the panel is large and that we assume π_3 is the known unit sampling probability of a random panelist into the specific

[3] In practice let $s^0 - s^2 \approx s^0$ and directly use the reference survey probabilities with no adjustment. If the analyst wishes nevertheless to adjust the reference survey weights so that it appears the reference survey sampling frame excludes the panel population s^2 ,simply calculate a new weight $\pi'_{Rk} = \pi_{Rk}(\frac{N-n_2}{N})$. Then $\sum_{s_R} \pi'^{-1}_{Rk} + \sum_{s^2} \pi_{2k}^{-1} = N$.

survey sample.

Two comments: Our premise above is that s_R is processed independently from s^2, however, when the same organization controls the two survey sample processes we can avoid the overlap $s^2 \cap s_R$ by simply including a screening question in the reference survey protocol. If this is applied the reference survey sampling frame excludes the panel population s^2 and the new weight $\pi^{'}_{Rk}$ are such that $\sum_{s_R} \pi^{'-1}_{Rk} + \sum_{s^2} \pi^{-1}_{2k} = N$. A second point is that when $\mathbf{x}^0$ include both design and selection covariates for s_R and s^2 respectively the discussion above implies that we can in fact ignore the weights π_k^{-1} in estimating π_{2k} and proceed with finding $\hat{\pi}_{2k}$ directly from $\hat{\mathbf{U}}_\pi$ and inflating the estimates by $\frac{n_R}{N}$ or if taking into account the overlap $\frac{n_R}{N-n_2}$.

It is also instructive to clarify on the common case where the analyst is an end user of the dataset independent of the panel management team and thus not exposed to the entire panel data set s^2, but rather only to the final survey sample set s^3. This means the dataset available is $s = s_R \cup s^3$ with known weights π_{Rk} for units s_R and weights π_{3k} for the panel respondents s^3 to the survey. Still, even here the analyst can draw valid inference. The only adjustment necessary is in the definition of the combined selection probability π_k defined in (4.7). In this case, for a given panel set s^2 the unit selection probability attached to each member of the combined dataset s is now

$$\pi_k = \begin{cases} \pi_{3k} & k \in s^3 \\ \pi_{Rk} & k \in s^0 - s^2. \end{cases}$$

Relying on the same conditional independence assumptions stated above, the estimation equations (4.8) defined over $s = s^3 \cup s_R$ and weighted by π_k^{-1} also give consisted estimates of the panel selection probabilities.

Interestingly, the same bias manifests itself when calculating our proposed m-type estimators, that is estimation by sequential linear regression models. Although model parameters $\boldsymbol{\beta}$ are consistent even when the reference survey weights are ignored, the relevant finite population sizes such as N, $n_1, \overline{\mathbf{X}}^0_{s^0}$ and $\overline{\mathbf{X}}^1_{s^1}$ cannot be estimated without the use of correct reference survey weights. I give a numerical demonstration of this in a simulation study described in section 4.3. Furthermore, if probabilities π_{Rk} contain information on the design or post selection adjustment covariates not recorded directly into the dataset s, ignorability will not hold and neither m or π-type estimation give consistent results. This idea will also be demonstrated in the second part of the simulation study in section 4.3.

4.2 Adjusting the Estimation Algorithms

In the following sections I detail the estimation procedure over the combined data of panel and reference survey samples for three specific examples of a π, m and πm estimators. In all three I continue with the idea of IPW estimation of the estimation mechanism under a fully observed finite population set.

4.2.1 The case of π-type Estimator

The process aims to compute the estimator

$$\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_3 s^3} = N^{-1} \sum_{s^3} Y_k \hat{\underline{\pi}}_{3k}^{-1} \ \ ; \text{ where } \hat{\underline{\pi}}_{3k} = \hat{\pi}_{1k} \cdot \hat{\pi}_{2k} \cdot \hat{\pi}_{3k}$$

under a three phase sequential framework.

Our first step is to estimate $\pi_{1k} = Pr(S_k^1 = 1|s^0)$ the internet selection probability - defined for some populations as the probability of being a frequent Web user. If over the entire population of interest s^0 the pairs $(\mathbf{X}^0, S^1)$ are observed, one approach is to assume that unit probability is independent and follows logit model $p(s_k^1|\mathbf{x}^0, s^0) = exp(\mathbf{x}_k^{0'}; \boldsymbol{\alpha}^0)/[1 + exp(\mathbf{x}_k^{0'}; \boldsymbol{\alpha}^0)]$ which I denote $\pi_{1k}(\mathbf{x}^0; \boldsymbol{\alpha}^0)$ and also that $0 < \pi_{1k}(\mathbf{x}^0; \boldsymbol{\alpha}^0) \leq 1$ for all $k \in s^0$. A consistent estimator $\hat{\boldsymbol{\alpha}}_\pi^0$ would be the solution to the estimation equations

$$U_{s^1}(s^0) = \sum_{s^0} (S_k^1 - \pi_{1k}(\mathbf{x}^0; \boldsymbol{\alpha}^0))\mathbf{x}_k^0 = 0.$$

In practice, $s^0 = s^1 \cup \overline{s}^1$ is not observed but rather only a random sample $s_R = s_R \cap (s^1 \cup \overline{s}^1)$ and we can replace the estimating equations $U_{s^1}(s^0)$ by

$$\begin{aligned} \hat{U}_{\pi s^1}(s_R) &= \sum_{s^0} S_{Rk}[S_k^1 - \pi_1(\mathbf{x}_k^0; \boldsymbol{\alpha}^0)]\mathbf{x}_k^0 \pi_{Rk}^{-1} = 0 \\ &= \sum_{s_R} [S_k^1 - \pi_1(\mathbf{x}_k^0; \boldsymbol{\alpha}^0)]\mathbf{x}_k^0 \pi_{Rk}^{-1} = 0 \qquad (4.9) \end{aligned}$$

where S_{Rk} and π_{Rk} are the unit selection indicator and selection probability associated with the sampling design $p(s_R)$. That $\hat{U}_{\pi s^1}(s_R)$ is unbiased for $U_{s^1}(s^0)$ is immediate from

$$E[\hat{U}_{\pi s^1}(s_R) - U_{s^1}(s^0)] = EE\left\{ \sum_{s^0} (S_{Rk}\pi_{Rk}^{-1} - 1)[S_k^1 - \pi_1(\mathbf{x}_k^0; \boldsymbol{\alpha}^0)]\mathbf{x}_k^0 | s^1, \mathbf{x}^{\mathbf{0}} \right\}$$

which is equal to zero as we assume that $p(s_R)$ is by design independent of Internet connectivity. Thus when π_{Rk} are correctly calculated and $\pi_1(\mathbf{x}_k^0; \boldsymbol{\alpha}^0)$ is correctly specified, the estimated probabilities $\hat{\pi}_{1k} = \pi_{1k}(\mathbf{x}^0; \hat{\boldsymbol{\alpha}}^0)$ are consistent estimators.

Similarly, to estimate the conditional unit probability of joining the Web panel we can assume that selection into the panel is independent and follows a logit model $p(S_k^2 = 1|\underline{\mathbf{x}}^1, s^1) = exp(\underline{\mathbf{x}}_k^{1'}, s^1; \boldsymbol{\alpha}^1)/[1 + exp(\underline{\mathbf{x}}_k^{1'}, s^1; \boldsymbol{\alpha}^1)]$, denoted by $\pi_{2k}(\underline{\mathbf{x}}^1; \boldsymbol{\alpha}^1)$ and that $0 < \pi_{2k}(\underline{\mathbf{x}}^1; \boldsymbol{\alpha}^1) \leq 1$ for all $k \in s^1$. If pairs $(\underline{\mathbf{x}}^1, s^2)$ are observed over the set of Internet population $s^1 = s^2 \cup (\overline{s}^2 \cap s^1)$ consistent estimators $\hat{\boldsymbol{\alpha}}^1$ would be the solution to the estimation equations

$$U_{s^2}(\boldsymbol{\alpha}^1, s^1) = \sum_{s^1}[S_k^2 - \pi_2(\underline{\mathbf{x}}_k^1; \boldsymbol{\alpha}^1)]\underline{\mathbf{x}}_k^1 = 0.$$

However, we do not observe the entire Internet connected population but rather a sample of it extracted from the artificially combined dataset $s = (s_R \cup s^2)$ of reference survey data and Web panel recruitment data. As in section 4.1.3 above we define indicator S_k over s that is equal to one when $k \in s$ and zero otherwise and attach to each unit k in s the artificial unit selection probabilities

$$\pi_k = \begin{cases} 1 & k \in s^2 \\ \pi_{Rk} & k \in s_R \end{cases}$$

which is a partial definition as it ignores the possibility of a unit k being a member of both s^2 and s_R. However, throughout I assume $Pr(S_{Rk} \cdot S_k^2 = 1|s^0) \approx 0$ which can be explained by the separate processes of panel and reference survey selection implying that $Pr(S_{Rk} \cdot S_k^2 = 1|s^0) = Pr(S_{Rk} = 1|s^0)Pr(S_k^2 = 1|s^0)$ as well as noting that $p(s^2|s^0) \approx 0$ for a large finite population.

Given weights π_k^{-1}, we calculate consistent estimators $\hat{\pi}_{2k}$ of π_{2k} by

$$\hat{\mathbf{U}}_\pi = \sum_{k \in s} \pi_k^{-1}[S_k^2 - \pi_2(\mathbf{x}_k^0; \boldsymbol{\alpha})]\mathbf{x}_k^0 = \mathbf{0}$$

which in 4.1.3 is shown to be approximately unbiased of the unavailable finite population estimation equation $U_{s^2}(\boldsymbol{\alpha}^1, s^1)$ defined over the entire Internet connected population s^1.

Finally, we observe $(\underline{\mathbf{X}}_k^2, S_k^3)$ for all $k \in s^2$ the panel population. Assuming the sample selection design and unavoidable unit non response can be accurately modeled jointly by

$$E(S_k^3|\underline{\mathbf{x}}^2, s^2; \boldsymbol{\alpha}^2) = (1 + e^{-\underline{\mathbf{x}}_k^{2'}\boldsymbol{\alpha}^2})^{-1}$$

denoted by $\pi_{3k}(\underline{\mathbf{x}}^2; \boldsymbol{\alpha}^2)$ and that $0 < \pi_{3k}(\underline{\mathbf{x}}^2; \boldsymbol{\alpha}^2) \leq 1$ for all $k \in s^2$ then consistent estimates $\hat{\boldsymbol{\alpha}}^2$ of parameters $\boldsymbol{\alpha}^2$ solve the estimation equations

$$U_{s^3}(s^2) = \sum_{s^2}(S_k^3 - \pi_{3k}(\underline{\mathbf{x}}^2; \boldsymbol{\alpha}^2))\underline{\mathbf{x}}_k^2 = 0.$$

Let $\hat{\pi}_{3k} = \pi_{3k}(\underline{\mathbf{x}}^2; \hat{\boldsymbol{\alpha}}^2)$ and calculate over the entire panel s^2. As noted above, the selection process governing s^3 is in reality a combination of a designed random sample and an unknown non response mechanism. Still it is reasonable to assume that for the panel management team the latter is well understood given the unique control, observational power and data recording abilities an Internet panel has. The sampling model is known by design and so together a correct model specification of this selection process is expected.

The alternative One Phase approach

To isolate the benefit of including an Internet selection phase I introduce briefly a one phase estimator which ignores the Internet usage phase and model $Pr(S^2_k = 1|s^0)$ directly. We model over $s = s^2 \cup s_R$ and the indicator

$$S_k = \begin{cases} 1 & k \in s^2 \\ 0 & k \notin s^2 \end{cases}$$

with associated probabilities $\pi_k = 1$ for all members in s^2 the panel sample and $\pi_k = \pi_{Rk}$ for all units of s_R the reference survey sample. I hypothesize the model $Pr(S^2_k = 1|\mathbf{x}^0, s^0) = [1 + e^{-\mathbf{x}^{0'}\boldsymbol{\alpha}}]^{-1}$ denoted $\pi_2(\mathbf{x}^0_k; \boldsymbol{\alpha})$ then parameters $\boldsymbol{\alpha}$ can be estimated by the likelihood equations

$$U_{s^2}(s^0) = \sum_{s^0}[S^2_k - \pi_2(\mathbf{x}^0_k; \boldsymbol{\alpha})]\mathbf{x}^0_k = 0$$

but as only s is observed the pseudo likelihood equations are used

$$\begin{aligned} \hat{U}_{\pi s^2}(s) &= \sum_{s_R \cup s^2}[S^2_k - \pi_2(\mathbf{x}^0_k; \boldsymbol{\alpha})]\mathbf{x}^0_k\pi^{-1}_k = 0 \\ &= \sum_{s^0} S_k[S^2_k - \pi_2(\mathbf{x}^0_k; \boldsymbol{\alpha})]\mathbf{x}^0_k\pi^{-1}_k = 0. \end{aligned} \tag{4.10}$$

The main distinction is that $Pr(S^2_k = 1|s^0)$ is modeled without the valuable information available in $\mathbf{X}^1$ such as Internet behavioral information.

4.2.2 The case of m-type Estimator

Take the case where Y_k over the population of interest can be modeled by normal linear regression. The general idea which I reviewed in chapter 3 is that under a three phase sequential framework we can describe the estimand $E(\overline{\mathbf{Y}}_{s^0})$ by

$$\begin{aligned} E(\overline{\mathbf{Y}}_{s^0}) &= N^{-1}\mathbf{1}'_{s^0}E\left\{E\left[EE(\mathbf{Y}_{s^0}|\underline{\mathbf{x}}^2_{s^0})|\underline{\mathbf{x}}^1_{s^0}\right]|\mathbf{x}^0_{s^0}\right\} \\ &= N^{-1}\mathbf{1}'_{s^0}E\left\{E\left[EE(\mathbf{Y}_{s^3}|\underline{\mathbf{x}}^2_{s^2}, s^3)|\underline{\mathbf{x}}^1_{s^1}, s^2\right]|\mathbf{x}^0_{s^0}, s^1\right\} \end{aligned} \tag{4.11}$$

where the second line is due to the sequential conditional independence property.

This suggested the following estimation approach: assume the distribution of Y over the sub population s^2 follows a linear regression model and so process $p(s^3|s^2)$ is ignorable for this regression model over s^2. Thus fitting the regression over s^3 will give consistent estimates of $E(Y_k|\underline{\mathbf{x}}^2)$ over s^2. Then assume over s^1 that $E\left(E(Y_k|\underline{\mathbf{x}}^2)|\underline{\mathbf{x}}^1\right)$, a function of $\underline{\mathbf{X}}^1$ can be modeled by a regression model as well. Again, by ignorability, this model can be fitted only over s^2 giving consistent estimators. The remaining unobserved values over s^1 can be predicted using $\underline{\mathbf{X}}^1_{s^1-s^2}$. The process continues in same fashion over s^0 using $\mathbf{X}^0$. The average of these last predictions is a consistent estimator of the population expected value.

More technically, estimation proceeds as follows. First, define $m_{2s^3} = E(\mathbf{Y}_{s^3}|\underline{\mathbf{x}}^2_{s^2}, s^3)$, $m_{1s^2} = E(m_{2s^2}|\underline{\mathbf{x}}^1_{s^1}, s^2)$ and $m_{0s^1} = E(m_{1s^1}|\mathbf{x}^0_{s^0}, s^1)$ and from 4.11 this gives the equation

$$\begin{aligned} E(\overline{\mathbf{Y}}_{s^0}) &= N^{-1}\mathbf{1}'_{s^0}E\left\{E\left[E(m_{2s^2}|\underline{\mathbf{x}}^1_{s^1}, s^2)\right]|\mathbf{x}^0_{s^0}, s^1\right\} \\ &= N^{-1}\mathbf{1}'_{s^0}EE(m_{1s^1}|\mathbf{x}^0_{s^0}, s^1) \\ &= N^{-1}\mathbf{1}'_{s^0}E(m_{0s^0}) \end{aligned}$$

thus, we start by denoting for $k \in s^3$ the observed panel response sample $y_k = m_{3k}$, and specifying a parametric regression model $E(Y_k|\underline{\mathbf{X}}^2_{s^2}; \boldsymbol{\beta}_2) = \underline{\mathbf{x}}^{2'}_k\boldsymbol{\beta}_2$ with independent unit variance $\sigma^2_{2k} = W^{-1}_{2k}$ over the subpopulation s^2. Let $m_{2k} = \underline{\mathbf{X}}^{2'}_k\boldsymbol{\beta}_2$ and note that given ignorability the model can be fitted over the observed set s^3. The consistent estimators $\hat{\boldsymbol{\beta}}_2$ of the model parameters solve the estimating equations

$$U_{m_2}(s^3) = \sum_{s^3}(Y_k - \underline{\mathbf{x}}^{2'}_k\boldsymbol{\beta}_2)\underline{\mathbf{x}}^2_k w_{2k} = 0 \tag{4.12}$$

which are observed in their entirety. Let $\hat{m}_{2k} = \underline{\mathbf{X}}^{2'}_k\hat{\boldsymbol{\beta}}_2$ be a consistent estimator of m_{2k}. Compute $\hat{m}_{2k}$ over s^2, giving $\hat{\mathbf{m}}_{2s^2}$.

Next, specify the parametric regression model $E(m_{2k}|\underline{\mathbf{x}}^1_{s^1}; \boldsymbol{\beta}_1) = \underline{\mathbf{X}}^{1'}_k\boldsymbol{\beta}_1$ with independent unit variance $\sigma^2_{1k} = W^{-1}_{1k}$ over the subpopulation s^1. Let $m_{1k} = \underline{\mathbf{X}}^{1'}_k\boldsymbol{\beta}_1$ and note that given ignorability the model can be fitted over the observed subset s^2. The consistent estimators $\hat{\boldsymbol{\beta}}_1$ solve the estimation equations

$$U_{m_1}(s^2) = \sum_{s^2}(\hat{m}_{2k} - \underline{\mathbf{x}}^{1'}_k\boldsymbol{\beta}_1)\underline{\mathbf{x}}^1_k w_{1k} = 0. \tag{4.13}$$

Let $\hat{m}_{1k} = \underline{\mathbf{X}}^{1'}_k\hat{\boldsymbol{\beta}}_1$ be then a consistent estimator of m_{1k}. If possible we would wish to calculate $\hat{m}_{1k}$ over s^1, however instead, we predict the $\hat{m}_{1k}$'s over the random subset of the Internet connected population $s^1_R \cup s^2$ using the available $\underline{\mathbf{X}}^1_k$'s.

In the last step, specify a parametric linear regression model $E(m_{1k}|\mathbf{x}^0_{s^0};\boldsymbol{\beta}_0) = \mathbf{x}^{0'}_k\boldsymbol{\beta}_0$ with independent unit variance $\sigma^2_{0k} = w^{-1}_{0k}$ over the general population s^0. Let $m_{0k} = \mathbf{x}^{0'}_k\boldsymbol{\beta}_0$ and again note that under ignorability $E(m_{1k}|\mathbf{x}^0_{s^0};\boldsymbol{\beta}_0) = E(m_{1k}|\mathbf{x}^0_{s^0}, s^1;\boldsymbol{\beta}_0)$. If the entire Internet population was observed, consistent estimators of parameters $\boldsymbol{\beta}_0$ would be found by the estimation equations

$$U_{m_1}(s^1) = \sum_{s^1}(\hat{m}_{1k} - \mathbf{x}^{0'}_k\boldsymbol{\beta}_0)\mathbf{x}^0_k w_{0k} = 0. \tag{4.14}$$

However, instead we fit the $\hat{m}_{1k}$'s over the available subset of the Internet connected population $s^1_R \cup s^2$ using the available $\underline{\mathbf{x}}^1_k$'s and so we can estimate the estimating equations (4.14) by

$$\begin{aligned}\hat{U}_{\pi m_1}(s) &= \sum_{s^1} S_k(\hat{m}_{1k} - \mathbf{x}^{0'}_k\boldsymbol{\beta}_0)\mathbf{x}^0_k\pi^{-1}_k w_{0k} = 0\\ &= \sum_{s^1_R \cup s^2}(\hat{m}_{1k} - \mathbf{x}^{0'}_k\boldsymbol{\beta}_0)\mathbf{x}^0_k\pi^{-1}_k\sigma_{0k} = 0\end{aligned}$$

which under the assumption of independence between the panel and reference survey selection process is an unbiased estimator of (4.14).

To calculate the $m-$estimator, ideally we would proceed by predicting the values $\hat{m}_{0k} = \mathbf{X}^{0'}_k\hat{\boldsymbol{\beta}}_0$ over the entire target population s^0 using the available $\mathbf{X}^0_{s^0}$ and estimate the population average by

$$\hat{\bar{\mathbf{Y}}}_{m_3} = N^{-1}\sum_{s^0}\hat{m}_{0k}$$

however, when the entire range of relevant covariates $\mathbf{X}^0_{s^0}$ or $\bar{\mathbf{X}}^0_{s^0}$ are not known- a likely scenario- we replace $\hat{\bar{\mathbf{Y}}}_{m_3}$ with

$$\begin{aligned}\hat{\bar{\mathbf{Y}}}_{m_3} &= N^{-1}\sum_{s}\pi^{-1}_k\hat{m}_{0k}\\ &= N^{-1}\sum_{s}\pi^{-1}_k\mathbf{X}^{0'}_k\hat{\boldsymbol{\beta}}_0 = \hat{\bar{\mathbf{X}}}^{0'}_\pi\hat{\boldsymbol{\beta}}_0 \text{ or alternatively by}\\ &= \hat{N}^{-1}_\pi\sum_{s}\pi^{-1}_k\mathbf{X}^{0'}_k\hat{\boldsymbol{\beta}}_0 = \tilde{\bar{\mathbf{X}}}^{0'}_\pi\hat{\boldsymbol{\beta}}_0\end{aligned}$$

where $\hat{N}^{-1}_\pi = \sum_s \pi^{-1}$ and so $\tilde{\bar{\mathbf{X}}}^{0'}_\pi$ is understood to be a ratio estimator of the population mean of covariates $\mathbf{x}^0$ over the target population.

Finally, as in chapter 3 I describe $\hat{\bar{\mathbf{Y}}}_{m_3}$ in a sequential 'Hat' format, which will be useful in explaining our $\pi-$balanced sampling strategy in section 4.4. Recall that under full observed dataset $\underline{d}^3_{s^3}$

$$\hat{\bar{\mathbf{Y}}}_{m_3} = N^{-1}\sum_{s^3}\underline{h}^2_k\mathbf{y}_k$$

where $\underline{h}^2_k = \mathbf{1}'_{s^0}\underline{\mathbf{H}}^1_{s^1 s^2}\underline{\mathbf{X}}^2_{s^2}\mathbf{A}^2_{s^3}\underline{\mathbf{X}}^2_k\mathbf{W}^2_k$ and $\underline{\mathbf{H}}^1_{s^1 s^2} = \mathbf{H}^0_{s^0,s^1}\mathbf{H}^1_{s^1,s^2}$, and for any $t = 1,...,T$ let $\mathbf{H}^{t-1}_{s^{t-1}s^t} = \underline{\mathbf{X}}^{t-1}_{s^{t-1}}\mathbf{A}^{t-1}_{s^t}\underline{\mathbf{X}}^{t-1'}_{s^t}\mathbf{W}^{t-1}_{s^t}$ with $\mathbf{A}^{t-1}_{s^t} = (\underline{\mathbf{x}}^{t-1'}_{s^t}\mathbf{W}^{t-1}_{s^t}\underline{\mathbf{X}}^{t-1}_{s^t})^{-1}$ and $\mathbf{W}^{t-1}_{s^t}{=}diag(\sigma^2_{t-1k})^{-1}$ a $n_t \times n_t$ diagonal matrix of model weights used in the estimation equations of model m_{t-1}.

However, as s^0 and s^1 based quantities are estimated over s_R the m-estimator takes the following shape

$$\hat{\bar{\mathbf{Y}}}_{m_3} = \hat{N}_{\pi}^{-1}\mathbf{1}'_n\hat{\mathbf{H}}^0_{\pi s,s\cap s^1}\hat{\mathbf{H}}^1_{\pi s\cap s^1,s^2}\mathbf{H}^2_{s^2,s^3}\mathbf{y}_{s^3}$$

where $\mathbf{H}^2_{s^2,s^3} = \underline{\mathbf{X}}^2_{s^2}(\underline{\mathbf{X}}^{2'}_{s^3}\mathbf{W}^2_{s^3}\underline{\mathbf{X}}^2_{s^3})^{-1}\underline{\mathbf{X}}^{2'}_{s^3}\mathbf{W}^2_{s^3}$ is the normal hat matrix of a linear regression fit over $s^3$and projected to s^2, while

$$\begin{aligned}\hat{\mathbf{H}}^0_{\pi s,s\cap s^1} &= \underline{\mathbf{X}}^1_{s\cap s^1}\boldsymbol{\pi}^{-1}_{s\cap s^1}(\underline{\mathbf{X}}^{1'}_{s^2}\mathbf{W}^1_{s^2}\underline{\mathbf{X}}^1_{s^2})^{-1}\underline{\mathbf{X}}^{1'}_{s^2}\mathbf{W}^1_{s^2}\\ \hat{\mathbf{H}}^1_{\pi s\cap s^1,s^2} &= \mathbf{X}^0_s\boldsymbol{\pi}^{-1}_s(\mathbf{X}^{0'}_{s\cap s^1}\boldsymbol{\pi}^{-1}_{s\cap s^1}\mathbf{W}^0_{s\cap s^1}\mathbf{X}^0_{s\cap s^1})^{-1}\mathbf{X}^{0'}_{s\cap s^1}\boldsymbol{\pi}^{-1}_{s\cap s^1}\mathbf{W}^0_{s\cap s^1}\end{aligned}$$

with $\mathbf{W}^2_{s^3} = diag(\sigma^2_{2k})^{-1},\ k \in s^3$, $\mathbf{W}^1_{s^2} = diag(\sigma^2_{1k})^{-1},\ k \in s^2$ and $\mathbf{W}^0_{s\cap s^1} = diag(\sigma^2_{0k})^{-1},\ k \in s \cap s^1$. Note that from $\hat{\mathbf{H}}^1_{\pi s\cap s^1,s^2}$ it is clear that the model is fit over s^2 which is observed entirely but projected to the unobserved Internet population s^1 by the selection probabilities $\boldsymbol{\pi}^{-1}_{s\cap s^1}$ of the combined data set s.

4.2.3 The case of πm-type Estimator

The estimation procedure intertwines the the m-estimator procedure with the π-estimator selection probabilities.

Over the panel population s^2

1. Assume that panel survey selection design and the unavoidable unit non response can be accurately modeled together by

   $$E(S^3_k|\underline{\mathbf{x}}^2, s^2; \boldsymbol{\alpha}^2) = (1 + e^{-\underline{\mathbf{x}}^{2'}_k\boldsymbol{\alpha}^2})^{-1}$$

   which I denote by $\pi_{3k}(\underline{\mathbf{x}}^2; \boldsymbol{\alpha}^2)$ and that $0 < \pi_{3k}(\underline{\mathbf{x}}^2; \boldsymbol{\alpha}^2) \leq 1$ for all $k \in s^2$. Consistent estimates $\hat{\boldsymbol{\alpha}}^2$ of parameters $\boldsymbol{\alpha}^2$ solve the estimation equations

   $$U_{s^3}(s^2) = \sum_{s^2}(S^3_k - \pi_{3k}(\underline{\mathbf{x}}^2; \boldsymbol{\alpha}^2))\underline{\mathbf{x}}^2_k = 0.$$

   Let $\hat{\pi}_{3k} = \pi_{3k}(\underline{\mathbf{x}}^2; \hat{\boldsymbol{\alpha}}^2)$ and calculate for all units over the entire observed panel s^2.

2. Denote $y_k = m_{3k}$ for $k \in s^3$ and specify the parametric regression model over the subpopulation s^2

   $$E(m_{3k}|\underline{\mathbf{x}}^2_{s^2};\boldsymbol{\beta}_2) = \underline{\mathbf{x}}^{2'}_k\boldsymbol{\beta}_2$$

   with independent unit variance $V(m_{3k}|\underline{\mathbf{x}}^2_{s^2}) = \sigma^2_{2k}$. Let $m_{2k} = \underline{\mathbf{X}}^{2'}_k\boldsymbol{\beta}_2$ and note that given ignorability the model can be fitted over the observed set s^3. The consistent estimators $\hat{\boldsymbol{\beta}}_{2\pi}$ of the model parameters solve the estimating equations

   $$U_{m_3}(s^3) = \sum_{s^3}(Y_k - \underline{\mathbf{x}}^{2'}_k\boldsymbol{\beta}_2)\underline{\mathbf{x}}^2_k\sigma^{-2}_{2k}\hat{\pi}^{-1}_{3k} = 0 \qquad (4.15)$$

   which are observed in their entirety. Denote $\hat{m}_{2\pi k} = \underline{\mathbf{X}}^{2'}_k\hat{\boldsymbol{\beta}}_{2\pi}$ and compute over s^2, giving the vector $\hat{\mathbf{m}}_{2\pi s^2}$.

Over the available Internet sub-population $s^1_R \cup s^2$

1. Assume that the panel self selection process is independent, individual and follows

   $$E(S^2_k|\underline{\mathbf{x}}^1, s^1;\boldsymbol{\alpha}^1) = (1 + e^{-\underline{\mathbf{X}}^{1'}_k\boldsymbol{\alpha}^1})^{-1}$$

   over the Internet subpopulation. Denote the unit probabilities by $\pi_{2k}(\underline{\mathbf{x}}^1;\boldsymbol{\alpha}^1)$ and assume that $0 < \pi_{2k}(\underline{\mathbf{x}}^1;\boldsymbol{\alpha}^1) \leq 1$ for all $k \in s^1$. As discussed earlier, the estimator $\hat{\boldsymbol{\alpha}}^1_\pi$ is the solution to the equations

   $$\begin{aligned}\hat{U}_{\pi s^2}(s) &= \textstyle\sum_{s^1} S_k(S^2_k - \pi_{2k}(\underline{\mathbf{x}}^1;\boldsymbol{\alpha}^1))\underline{\mathbf{x}}^1_k\pi^{-1}_k &= 0\\ &= \textstyle\sum_{s^1_R\cup s^2}(S^2_k - \pi_{2k}(\underline{\mathbf{x}}^1;\boldsymbol{\alpha}^1))\underline{\mathbf{x}}^1_k\pi^{-1}_k &= 0\end{aligned}$$

   where $s^1_R = s^1 \cap s_R$ is the set of reference survey respondents who are defined as Internet population members. Let $\hat{\pi}_{2k} = \pi_{2k}(\underline{\mathbf{x}}^1;\hat{\boldsymbol{\alpha}}^1_\pi)$ and calculate these estimated unit probabilities for all k's over $s^1_R \cup s^2$.

2. Specify the parametric regression model over the subpopulation s^1

   $$E(m_{2k}|\underline{\mathbf{x}}^1_{s^1};\boldsymbol{\beta}_1) = \underline{\mathbf{X}}^{1'}_k\boldsymbol{\beta}_1$$

   with independent unit variance $V(m_{2k}|\underline{\mathbf{x}}^1_{s^1}) = \sigma^2_{1k}$. Let $m_{1k} = \underline{\mathbf{X}}^{1'}_k\boldsymbol{\beta}_1$ and note that given ignorability $E(m_{2k}|\underline{\mathbf{x}}^1_{s^1};\boldsymbol{\beta}_1) = E(m_{2k}|\underline{\mathbf{x}}^1_{s^1}, s^2;\boldsymbol{\beta}_1)$ and so the model can be fitted over the observed subset s^2. The model consistent estimators $\hat{\boldsymbol{\beta}}_{1\pi}$ solve the estimation equations

   $$U_{m_2}(s^2) = \sum_{s^2}(\hat{m}_{2\pi k} - \underline{\mathbf{x}}^{1'}_k\boldsymbol{\beta}_1)\underline{\mathbf{x}}^1_k\sigma^{-2}_{1k}\hat{\pi}^{-1}_{2k} = 0. \qquad (4.16)$$

   Denote $\hat{m}_{1\pi k} = \underline{\mathbf{X}}^{1'}_k\hat{\boldsymbol{\beta}}_{1\pi}$ and calculate the predictions $\hat{m}_{1\pi k}$'s over the available random subset of the Internet connected population $s^1_R \cup s^2$ using the observed $\underline{\mathbf{X}}^1_k$'s.

Over the available population data $s_R \cup s^2$

1. Estimate the probability of being a frequent Web user, under the assumption
$$E(S_k^1 | \mathbf{x}^0, s^0; \boldsymbol{\alpha}^0) = (1 + e^{-\mathbf{X}_k^{0'} \boldsymbol{\alpha}^0})^{-1}$$
which I denote $\pi_{1k}(\mathbf{x}^0; \boldsymbol{\alpha}^0)$ and that $0 < \pi_{1k}(\mathbf{x}^0; \boldsymbol{\alpha}^0) \leq 1$ for all $k \in s^0$. A consistent estimator $\hat{\boldsymbol{\alpha}}_\pi^0$ can be found as the solution of the following estimation equations
$$\begin{aligned} \hat{U}_{\pi s^1}(s) &= \textstyle\sum_{s^0} S_k (S_k^1 - \pi_{1k}(\mathbf{x}^0; \boldsymbol{\alpha}^0)) \mathbf{x}_k^0 \pi_k^{-1} &&= 0 \\ &= \textstyle\sum_{s_R \cup s^2} (S_k^1 - \pi_{1k}(\mathbf{x}^0; \boldsymbol{\alpha}^0)) \mathbf{x}_k^0 \pi_k^{-1} &&= 0. \end{aligned}$$
calculated over the available random subset of s^0. Let $\hat{\pi}_{1k} = \pi_{1k}(\mathbf{x}^0; \hat{\boldsymbol{\alpha}}_\pi^0)$ and calculate over the combined set $s_R \cup s^2$.

2. Specify a parametric linear regression model
$$E(m_{1k} | \mathbf{x}_{s^0}^0; \boldsymbol{\beta}_0) = \mathbf{X}_k^{0'} \boldsymbol{\beta}_0$$
with independent unit variance $V(m_{1k} | \mathbf{x}_{s^0}^0) = \sigma_{0k}^2$ over the general population s^0. Let $m_{0k} = \mathbf{X}_k^{0'} \boldsymbol{\beta}_0$ and again note that under ignorability $E(m_{1k} | \mathbf{x}_{s^0}^0; \boldsymbol{\beta}_0) = E(m_{1k} | \mathbf{x}_{s^0}^0, s^1; \boldsymbol{\beta}_0)$ which implies that model fitting can be over s^1. We fit the $\hat{m}_{1\pi k}$'s over the available subset of the Internet connected population $s_R^1 \cup s^2$ using the available $\underline{\mathbf{X}}_k^1$'s by consistently estimating $\boldsymbol{\beta}_0$ using the following equations
$$\begin{aligned} \hat{U}_{\pi m_1}(s) &= \sum_{s^1} s_k (\hat{m}_{1\pi k} - \mathbf{x}_k^{0'} \boldsymbol{\beta}_0) \mathbf{x}_k^0 \pi_k^{-1} \sigma_{0k}^{-2} \hat{\pi}_{1k}^{-1} = 0 \\ &= \sum_{s_R^1 \cup s^2} (\hat{m}_{1\pi k} - \mathbf{x}_k^{0'} \boldsymbol{\beta}_0) \mathbf{x}_k^0 \pi_k^{-1} \sigma_{0k}^{-2} \hat{\pi}_{1k}^{-1} = 0 \end{aligned}$$
denote $\hat{m}_{0\pi k} = \mathbf{X}_k^{0'} \hat{\boldsymbol{\beta}}_{0\pi}$. It is interesting to note that π_k reshapes the estimating equations to resemble the distribution of s^1 while $\hat{\pi}_{1k}$ introduces selection consistency towards the model distribution over s^0 from s^1.

The $m-$estimator is calculated over the combined reference and panel survey sample
$$\begin{aligned} \hat{\bar{\mathbf{Y}}}_{\pi m_3} &= N^{-1} \sum_s \pi_k^{-1} \hat{m}_{0\pi k} \\ &= N^{-1} \sum_s \pi_k^{-1} \mathbf{X}_k^{0'} \hat{\boldsymbol{\beta}}_{0\pi} \\ &= \hat{\bar{\mathbf{X}}}_\pi^{0'} \hat{\boldsymbol{\beta}}_{0\pi} \end{aligned}$$
or when the size of the population N is not known, or for statistical stability reasoning we replace the denominator with $\hat{N}_\pi = \sum_s \pi_k^{-1}$.

In parallel to the m-estimator I describe $\hat{\overline{\mathbf{Y}}}_{\pi m_3}$ in a sequential 'Hat' format. In the case where $\underline{d}^3_{s^3}$ is available in full

$$\hat{\overline{\mathbf{Y}}}_{\pi m_3} = N^{-1} \sum_{s^3} \underline{h}^2_{\pi k} \mathbf{Y}_k$$

Where $\underline{h}^2_{\pi k} = \mathbf{1}'_{s^0} \underline{\mathbf{H}}^1_{\pi s^1 s^2} \underline{\mathbf{X}}^2_{s^2} \mathbf{A}^2_{\pi s^3} \underline{\mathbf{X}}^2_k \mathbf{W}^2_k \hat{\pi}^{-1}_{3k}$ and $\underline{\mathbf{H}}^1_{\pi s^1 s^2} = \mathbf{H}^0_{\pi s^0, s^1} \mathbf{H}^1_{\pi s^1, s^2}$, and for any $t = 1, ..., T$ let $\mathbf{H}^{t-1}_{\pi s^{t-1} s^t} = \underline{\mathbf{X}}^{t-1}_{s^{t-1}} \mathbf{A}^{t-1}_{\pi s^t} \underline{\mathbf{X}}^{t-1'}_{s^t} \mathbf{W}^{t-1}_{s^t} \underline{\hat{\boldsymbol{\pi}}}^{-1}_{ts^t}$ with $\mathbf{A}^{t-1}_{\pi s^t} = (\underline{\mathbf{X}}^{t-1'}_{s^t} \mathbf{W}^{t-1}_{s^t} \underline{\hat{\boldsymbol{\pi}}}^{-1}_{ts^t} \underline{\mathbf{X}}^{t-1}_{s^t})^{-1}$ and $\underline{\hat{\boldsymbol{\pi}}}^{-1}_{ts^t} = diag(\hat{\pi}^{-1}_{tk})^{-1}$ a $n_t \times n_t$ diagonal matrix of estimated selection probabilities calculated separately.

Replacing s^0 and s^1 quantities by reference survey estimators the πm-estimator takes the following shape

$$\hat{\overline{\mathbf{Y}}}_{\pi m_3} = \hat{N}^{-1}_{\pi} \mathbf{1}'_n \hat{\mathbf{H}}^0_{\pi s, s \cap s^1} \hat{\mathbf{H}}^1_{\pi s \cap s^1, s^2} \mathbf{H}^2_{\pi s^2, s^3} \mathbf{Y}_{s^3}$$

where $\mathbf{H}^2_{s^2, s^3} = \underline{\mathbf{X}}^2_{s^2} (\underline{\mathbf{X}}^{2'}_{s^3} \mathbf{W}^2_{s^3} \hat{\boldsymbol{\pi}}^{-1}_{3s^3} \underline{\mathbf{X}}^2_{s^3})^{-1} \underline{\mathbf{X}}^{2'}_{s^3} \mathbf{W}^2_{s^3} \hat{\boldsymbol{\pi}}^{-1}_{3s^3}$ is the weighted hat-type matrix of a linear regression fit over $s^3$and projected to s^2, while

$$\begin{aligned}
\hat{\mathbf{H}}^0_{\pi s, s \cap s^1} &= \mathbf{X}^0_s \boldsymbol{\pi}^{-1}_s (\mathbf{X}^{0'}_{s \cap s^1} \boldsymbol{\pi}^{-1}_{s \cap s^1} \mathbf{W}^0_{s \cap s^1} \hat{\boldsymbol{\pi}}^{-1}_{1 s \cap s^1} \mathbf{X}^0_{s \cap s^1})^{-1} \mathbf{X}^{0'}_{s \cap s^1} \boldsymbol{\pi}^{-1}_{s \cap s^1} \mathbf{W}^0_{s \cap s^1} \hat{\boldsymbol{\pi}}^{-1}_{1 s \cap s^1} \\
\hat{\mathbf{H}}^1_{\pi s \cap s^1, s^2} &= \underline{\mathbf{X}}^1_{s \cap s^1} \boldsymbol{\pi}^{-1}_{s \cap s^1} (\underline{\mathbf{X}}^{1'}_{s^2} \mathbf{W}^1_{s^2} \hat{\boldsymbol{\pi}}^{-1}_{2s^2} \underline{\mathbf{X}}^1_{s^2})^{-1} \underline{\mathbf{X}}^{1'}_{s^2} \mathbf{W}^1_{s^2} \hat{\boldsymbol{\pi}}^{-1}_{2s^2}
\end{aligned}$$

where $s \cap s^1 = s^1_R \cup s^2$ and with $\mathbf{W}^2_{s^3} = diag(\sigma^2_{2k})^{-1},\ k \in s^3$, $\mathbf{W}^1_{s^2} = diag(\sigma^2_{1k})^{-1},\ k \in s^2$ and $\mathbf{W}^0_{s \cap s^1} = diag(\sigma^2_{0k})^{-1} k,\ \in s \cap s^1$. Note that from $\hat{\mathbf{H}}^1_{\pi s \cap s^1, s^2}$ it is clear that the model is fit over s^2 which is observed entirely but projected to the unobserved Internet population s^1 by the selection probabilities $\boldsymbol{\pi}^{-1}_{s \cap s^1}$ of the combined data set s .

4.3 Simulation Studies to Demonstrate Basic Properties of Procedure

4.3.1 Testing the Basic Algorithm and the Use of Survey Weights

In the following section I describe a simple simulation study demonstrating the basic properties of the three sequential estimators $\hat{\overline{\mathbf{Y}}}_{\pi m}$, $\hat{\overline{\mathbf{Y}}}_{\pi}$ and $\hat{\overline{\mathbf{Y}}}_m$ under a setting more close to practice, that is when the estimators are calculated with an independent reference survey s_R. I use the available set up to examine two additional points. First is to demonstrate that modelling sequentially, while adding modeling burden by definition, may still be more efficient than a single model when each modeled sequence is more parsimonious. The second point I make is to demonstrate the behaviour of the estimators when the sampling weights are ignored. As before the estimand of interest is the population average $\overline{\mathbf{Y}}_{s^0}$, the simulation is repeated $2,000$.

The simulated population is constructed as follows. The finite population s^0 is of fixed size $N = 50,000$ where the set of $\mathbf{X}^0$ type of covariates have the following distribution

$$(q^0, \mathbf{w}^p, x^0) \sim N\left(1, \begin{array}{ccccc} 1 & 0.3 & 0.3 & 0.3 & .. \\ & 1 & .. & .. & .. \\ & & 1 & 0.3 & \\ & 0.3 & & 1 & \\ & 0.3 & .. & .. & 1 \end{array} \right)$$

where $p = 1, ..., 5$. That is a multivariate normal distribution of seven variables. I assume throughout that only covariate $\mathbf{X}^0$ is observed to the analyst while q^0 and the set of variables $W^1, ...W^5$ cannot be direcetly used.

Our interest is restricted to a single measurment of interst which has the following distribution

$$Y_k = 1 + x_k^0 + 0.2w_{1k} - .25w_2 + 0.5w_3 + 0.9w_4 - 0.1w_5 + \varepsilon_k$$

with independent standard normal errors. The three phase selection process starts with s^1 randomly selected by

$$p(s^1 = 1|x^0) = (1 + e^{2.5+0.8x_k^0})^{-1}$$

which results in an 'Internet connected population' of approximately $E(n_1) \approx 27,000$. The (self) selection of Web panel s^2 is simulated by the process

$$p(s^2 = 1|\underline{\mathbf{x}}^1, s^1) = (1 + e^{-1+0.08x_k^0+0.025x_k^1})^{-1}$$

where

$$X_k^1 = \begin{cases} e^{\boldsymbol{\alpha}'\mathbf{w}_k} & \forall k \in s^1 \\ - & \forall k \in \overline{s}^1 \end{cases} \quad \text{with } \boldsymbol{\alpha}'\mathbf{w}_k = 0.5w_1 + 0.5w_2 - 0.3w_3 + 0.3w_4 + 0.05w_5 + \varepsilon_k$$

which results in an average response rate of 0.25 from s^1 (the Internet population) and a panel size $E(n_2) \approx 7,000$. Note that the analyst observes only $\mathbf{X}^1$ while the set $\mathbf{W}^0$ of covariates are unobserved. Finally, assume a survey sample s^3 is randomly selected by known design on panel behavior statistic X^2, with a standard normal distribution defined over the s^2. The sampling design takes the form

$$p(s^3 = 1|x^2, s^2) = (1 + e^{\alpha_3 - 0.5x_k^2})^{-1}$$

where $\alpha_3 = -4.5$ resulting in an average survey sample size $E(n_3) \approx 1,000$.

To assist with estimation a reference survey sample is collected separately and independently from the panel selection process. The sample s_R is collected directly from s^0 by design

$$p(s_R = 1|x^0, q^0, s^1) = (1 + e^{\alpha_R - 0.05x_k^0 - 0.05q_k^0})^{-1}$$

where $\alpha_R = -5.68$ resulting in an average survey sample size $E(n_R) \approx 1,000$. It is informative to summarize the different selection and model assumptions in a unified table

Model	Covariates			
s^1	x^0			
s^2		$x^1(\mathbf{w^0})$		
s^3			x^2	
s_R	x^0			q^0
Y_k	x^0	$\mathbf{w^0}$		q^0

from which it is evident that X^2, the panel behavior statistic is redundant to estimation of $\overline{\mathbf{Y}}_{s^0}$, that is X^2 is not included in the minimum adjustment set (Pearl, 2000), and so including the covariate in an estimator (either m or π) will decrease efficiency. Also, note that I assume neither q^0 nor w^0 are observed. While the information in $\mathbf{W}^0$ is 'reduced' by X^1 which is observed, the information in q^0 is available only in the survey weights for s_R. Thus any estimation procedure ignoring s_R weights should introduce bias. The basic statistics of the simulated population and its subsets are described in table 4.1.

s^0	s^1	s^2	s^3	s_R
$\overline{\mathbf{Y}}_{s^0} = 1.009$	$\overline{\mathbf{Y}}_{s^1} = 1.108$	$\overline{\mathbf{Y}}_{s^2} = 1.223$	$\overline{\mathbf{Y}}_{s^3} = 1.219$	$\overline{\mathbf{Y}}_{s_R} = 0.910$
$\overline{\mathbf{X}}^0_{s^0} = 0.006$	$\overline{\mathbf{X}}^1_{s^1} = 2.497$	$\overline{\mathbf{X}}^2_{s^2} = 0.000$	$\overline{\mathbf{X}}^0_{s^3} = 0.095$	$\overline{\mathbf{X}}^0_{s_R} = -0.076$
$N = 50,000$	$n_1 = 27,161$	$n_2 = 7,697$	$n_3 = 95$	$n_R = 102$

Table 4.1: A list of selected statistics over the different populations of interest. Note that but for $N = 50,000$ the size of the finite population all figures are averages over the 2,000 simulations.

We calculate $\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_3}$,$\hat{\overline{\mathbf{Y}}}_{\underline{m}_3}$ and $\hat{\overline{\mathbf{Y}}}_{\underline{\pi m}_3}$ following the estimation steps outlined in the previous sections. All models are correctly specified using X^1, the (redundant) covariate X^2 , and the weights π_R and π_3. In addition I calculate a two phase π-estimator $\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_2}$ which ignores the selection $p(s^1|s^2)$ and directly models $p(s^2|s^0)$ by using $\mathbf{W}$ covariates. Under my description, this is a 'theoretical' estimator built over unobserved covaraite set. The first line in table 4.2 gives the estimation bias and standard deviation calculated over the 2,000 simulations.

	$\hat{\bar{\mathbf{Y}}}_{\underline{m}_3}$		$\hat{\bar{\mathbf{Y}}}_{\underline{\pi m}_3}$		$\hat{\bar{\mathbf{Y}}}_{\underline{\pi}_3}$		$\hat{\bar{\mathbf{Y}}}_{\underline{\pi}_2}$	
Estimation	B	$V^{1/2}$	B	$V^{1/2}$	B	$V^{1/2}$	B	$V^{1/2}$
π_3 and π_R	0.003	0.210	0.005	0.212	-0.002	0.266	-0.007	0.299
π_R only	-0.003	0.204	-0.003	0.202	-0.011	0.248	-0.020	0.289
π_3 nor π_R	-0.068	0.212	-0.069	0.215	-0.063	0.355	-0.075	0.389

Table 4.2: Estimation bias and standard deviation for the four estimators tested over the three estimation strategies. The three estimation strategies include from top to bottom: (i) the case where estimates were calculated taking into account both sets of weights π_3 and π_R , (ii) estimates ignored the within panel survey sample weights π_3, and (iii) the case where neither π_R nor π_3 were applied.

Inspecting the first line of results it is evident first that as expected the variance sizes over the competing three phase estimators has the following order $V(\hat{\bar{\mathbf{Y}}}_{\pi}) > V(\hat{\bar{\mathbf{Y}}}_{\pi m}) \geq V(\hat{\bar{\mathbf{Y}}}_{m})$. However, the difference between the $\pi-$estimator and m or $\pi m-$estimators is smaller than when calculated over the entire population. This is because both types of estimators are calculated over a limited set of observations whereas in the discussion in the previous chapter the m and πm-estimators where calculated recursivley over the entire population. A second point is that we notice that $V(\hat{\bar{\mathbf{Y}}}_{\underline{\pi}_2}) > V(\hat{\bar{\mathbf{Y}}}_{\underline{\pi}_3})$ that is the three phase π-estimator has smaller variance than a two phase estimator. Both estimators are correctly specified and valid, and differ on (i) the number of selection models estimated, and (ii) the number of covariates included in the selection models. Under the setting of this simulation the use of three parsimonious models is more efficient than two larger models.

With the same artificial population I demonstrate as well the effect of omitting the reference or the within panel survey selection weights. The bottom three lines in table 4.2 describes these different scenarios: (i) a normal application using both π_3 and π_R, (ii) a partial application which ignores the unequal unit selection probability π_3 and replaces it with a constant probability $\pi = n_2/n_3$, and (iii) the case where the reference survey weights are replaced by the constant probability $\pi = n_R/N$ as well as treating the selection of s^3 again as following a design of equal unit probability equal to $\pi = n_2/n_3$. Scenarios such as these, under a two phase example, were touched upon in our discussion in section 4.1.3.

Our variable of interest $\mathbf{Y}$ is independent of $\mathbf{X}^2$ and so treating the final sample mistakenly as a simple random sample as we do in the second scenario introduces no error to the two m-based estimators and only small distortion to the two π-based estimators. There is also a small improvement in the efficiency of the estimators. Underlying the third scenario is the fact we are ignoring the available information on covariate q held in the reference survey design weights π_R. Covariate q is not independent of Y and X^0 and as expected all four estimators examined are biased. Interestingly, and less obvious is that the size of the bias across all estimators is quire stable. For the m and $\pi m-$based estimators given the correct specification the bias can be attributed mainly to the invalid estimate of $\overline{\mathbf{X}}^0_{s^0}$- in our setting underestimating the expected population average. For π-estimators the bias is indirect and is a function of the selection probabilities estimated over incorrectly unadjusted estimation equations. In the following section we bring more detail into the mechanism s_R may introduce bias into our estimators.

4.3.2 The Question of s_R and s^3 Sample Sizes

In planning a Web panel based survey study, the practitioner may have the choice of determining how much resource should be invested in the panel survey and how much in the reference survey. Clearly, large samples from both platforms is desirable, but as the reference survey is more expensive companies (such as Harris Interactive or Ipsos) tend to base their estimation on a relatively small reference sample size compared to the panel survey sample. It is interesting then to evaluate the effect different allocations of resources, resulting in different sizes of n_3 and n_R the sample sizes of s^3 and s_R respectively, have on estimator properties. A specific question is whether there is a symmetry in the estimators' behaviour when considering the ratio of n_3 and n_R.

To test this I propose a simulation study using the same population described in the previous section. To create varying achieved sample sizes for the two survey sample sets I change the intercept coefficient in the sampling designs

$$p(s^3 = 1|x^2, s^2) = (1 + e^{\alpha_3 - 0.5x^2_k})^{-1}$$

which samples s^3 by known design on panel behavior statistic $\mathbf{X}^2$ from the panel s^2, and

$$p(s_R = 1|x^0, q^0, s^1) = (1 + e^{\alpha_R - 0.05x^0_k - 0.05q^0_k})^{-1}$$

which samples s_R directly from s^0. The different intercept coefficients and achieved sample sizes are displayed in table 4.3.

Overall we examine the same four estimators $\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_3}\hat{\overline{\mathbf{Y}}}_{\underline{m}_3}\hat{\overline{\mathbf{Y}}}_{\underline{\pi m}_3}$ and $\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_2}$ of the previous section over the 11 scenarios defined by the ratio of the two survey sample sizes

α_3	$E(n_3)$	α_R	$E(n_R)$
-4.95	50	-6.3	50
-4.65	75	-5.85	75
-4.5	100	-5.68	100
-4.0	150	-5.3	150
-3.5	250	-4.8	250
-2.75	500	-4.05	500

Table 4.3: The different intercept coefficients and the achieved average sample sizes for the the within panel survey sample s^3 and the independent random survey sample s_R. The two survey designs follow $p(s^3 = 1|x^2, s^2) = (1 + e^{\alpha_3 - 0.5x_k^2})^{-1}$ and $p(s_R = 1|x^0, q^0, s^1) = (1 + e^{\alpha_R - 0.05x_k^0 - 0.05q_k^0})^{-1}$ respectively.

- $n_R = 50$ and $n_3/n_R = 1, 1.5, 2, 3, 5, 10$
- $n_3 = 50$ and $n_R/n_3 = 1, 1.5, 2, 3, 5, 10$.

To make a stronger distinction between the different allocation of resources I change the estimation algorithms outlined in the previous sections so that for the two $m-$estimators $\hat{\bar{\mathbf{Y}}}_{\underline{m}_3}$ and $\hat{\bar{\mathbf{Y}}}_{\underline{\pi m}_3}$ we calculate the regression coefficient parameters over the panel survey sample s^3 only, rather than the fused sample $s = s^3 \cup s_R$. The population average is estimated as before over the fused sample set. A summary of the main results are summarized in the table 4.4 for 11 scenarios n_3/n_R. The bias B and the $MSE = B^2 + V(\hat{\bar{y}})$ where $V(\cdot)$ is the variance are estimated over 2,000 simulations. A useful graphical summary which supplements table 4.4 is given in figure 4.4 which displays the simulation distributions of the same estimators across the 11 scenarios by a simple Boxplot graph.

Several points can be drawn. First is that the estimators' behaviour across the varying ratio settings of n_R/n_3 are not symmetrical. For example it is obvious that for each of the five pairs of ratio settings with an overall same sample size, that is $n_3 + n_R = n$, for $n = 550, 300, 200, 150$ and 125, the estimator $\hat{\bar{\mathbf{Y}}}_{\underline{\pi}_3}$ is significantly more efficient (or has a lower MSE) for cases where $n_R/n_3 > 1$. For the cases where $n_R \approx 50$ the three phase estimator is still unbiased, but in practice estimates will likely to be considerably 'off the mark' - a classic critique of $\pi-$estimation strategy. This finding gives another demonstration of the effect the size of the reference survey has on Web panel inference. Thus for a small sample size study based on a $\hat{\bar{\mathbf{Y}}}_{\underline{\pi}_3}$ estimator any additional resource should go into adding more reference survey respondents.

The same asymmetry is displayed by $\hat{\bar{\mathbf{Y}}}_{\underline{\pi}_2}$ estimator, but the disparity is substantially narrower. Looking at the two $\pi-$estimators it is worth noting as well, that the small efficiency gain of $\hat{\bar{\mathbf{Y}}}_{\underline{\pi}_3}$ against $\hat{\bar{\mathbf{Y}}}_{\underline{\pi}_2}$ we observed in section 4.3.1 is still displayed here for the cases where $n_R/n_3 > 1$, but when the ratio tilts towards

n_R/n_3		$\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_3}$	$\hat{\overline{\mathbf{Y}}}_{\underline{m}_3}$	$\hat{\overline{\mathbf{Y}}}_{\underline{\pi m}_3}$	$\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_2}$
500/50	B	-0.2%	0.2%	0.3%	0.1%
	MSE	0.089	0.049	0.052	0.091
250/50	B	0.1%	0.5%	0.4%	0.3%
	MSE	0.101	0.056	0.058	0.104
150/50	B	-0.4%	0.1%	0.1%	-0.1%
	MSE	0.107	0.058	0.059	0.108
100/50	B	0.1%	0.4%	0.7%	1.1%
	MSE	0.121	0.062	0.064	0.124
75/50	B	0.0%	-0.3%	-0.3%	1.3%
	MSE	0.123	0.065	0.068	0.133
50/50	B	-0.1%	0.1%	0.1%	2.3%
	MSE	0.695	0.075	0.076	0.205
50/75	B	-4.7%	-0.3%	-0.3%	2.2%
	MSE	0.684	0.060	0.081	0.183
50/100	B	-0.6%	-0.1%	-0.3%	2.3%
	MSE	0.668	0.056	0.062	0.158
50/150	B	-2.9%	0.0%	1.1%	1.3%
	MSE	0.674	0.044	0.050	0.142
50/250	B	-2.1%	0.1%	-0.1%	3.8%
	MSE	0.663	0.038	0.041	0.112
50/500	B	0.0%	0.7%	0.7%	4.9%
	MSE	0.607	0.033	0.033	0.126

Table 4.4: The bias and mean square error of the four estimators $\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_3}\hat{\overline{\mathbf{Y}}}_{\underline{m}_3}\hat{\overline{\mathbf{Y}}}_{\underline{\pi m}_3}$ and $\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_2}$estimated over 2,000 simulation runs for each of the 11 scenarios examined. The 11 scenarios represent the cases where $n_R = 50$ with $n_3/n_R = 1, 1.5, 2, 3, 5, 10$ and $n_3 = 50$ with $n_R/n_3 = 1, 1.5, 2, 3, 5, 10$.

a larger panel sample survey size the $\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_2}$ is much more efficient. It seems that for small samples sizes of s_R the additional selection phase brings much more instability that can be compensated by the more separate parsimonious models underlining $\hat{\overline{\mathbf{Y}}}_{\underline{\pi}_3}$.

The two m-estimators $\hat{\overline{\mathbf{Y}}}_{\underline{m}_3}$ and $\hat{\overline{\mathbf{Y}}}_{\underline{\pi m}_3}$ also are asymmetrical in the sense discussed above, but here the trend reverses. That is to say that both estimators are (slightly) more efficient for cases where $n_3/n_R > 1$. To understand this result it is useful to note that for a single phase πm-estimator, when there exists a p-vector of constants $\boldsymbol{\delta}$ so that $\mathbf{X}^0_{s^0}\boldsymbol{\delta} = \mathbf{1}_{s^0}$ where $\mathbf{1}_{s^0}$ is the N-vector of constant 1, then $\overline{\mathbf{Y}}_{s^0} = \overline{\mathbf{X}}_{s^0}\mathbf{B}_0$ where $\mathbf{B}_0$ is the finite population regression coefficient[4] the

[4] when weighted least square regression is appropriate the assumption changes to state that

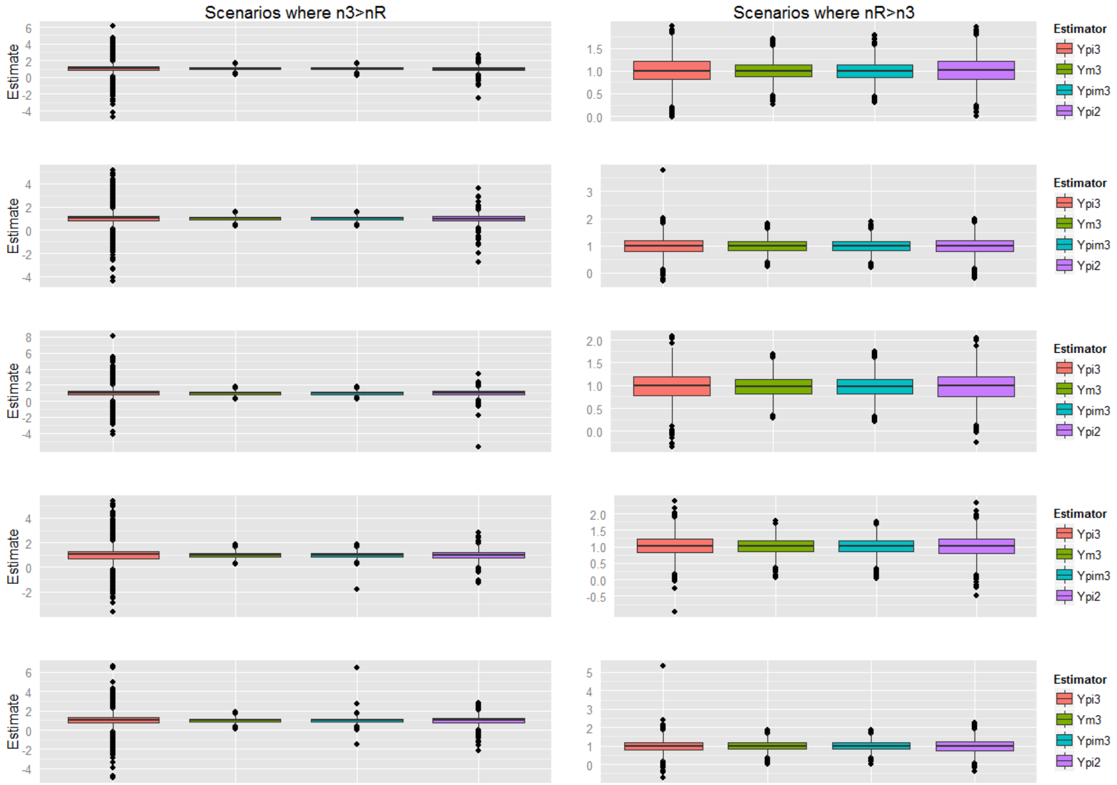


Figure 4.4: A box plot summary of the distribution of $\hat{\overline{\mathbf{Y}}}_{\pi_3}\hat{\overline{\mathbf{Y}}}_{\underline{m_3}}\hat{\overline{\mathbf{Y}}}_{\underline{\pi m_3}}$ and $\hat{\overline{\mathbf{Y}}}_{\underline{\pi_2}}$over 2,000 simulation runs for each of the 11 scenarios examined. The graph shows from top to bottom- on the left hand side the ratios $n_3/n_R = 10, 5, 3, 2, 1.5$ and on the right hand side the ratios $n_3/n_R = 0.1, 0.2, 0.33, 0.5, 0.67$.

πm-estimation error can be written as

$$\hat{\overline{\mathbf{Y}}}_{\pi m} - \overline{\mathbf{Y}}_{s^0} = \left(\hat{\overline{\mathbf{X}}}^0_{\pi s} - \overline{\mathbf{X}}^0_{s^0}\right)' + \overline{\mathbf{X}}^{0\prime}_{s^0}\left(\hat{\boldsymbol{\beta}}_0 - \mathbf{B}_0\right) \tag{4.17}$$

a deconstruction of the error into that of estimating $\overline{\mathbf{X}}^0_{s^0}$ and the difference between the sample and population estimates of the regression parameters $\boldsymbol{\beta}$. Under correct specification of π and m-models the first part of (4.17) will be large when s_R is skewed with respect to model covariates' population averages; The second part of (4.17) on the other hand will generally be small under correct specification even when the sample is skewed. Still, under our simulation setting, where the regression parameters $\boldsymbol{\beta}$ are estimated over s^3 the loss in efficiency in estimator $\hat{\overline{\mathbf{X}}}^0_{\pi s}$ of $\overline{\mathbf{X}}^0_{s^0}$ is smaller than the gains in estimating $\boldsymbol{\beta}$ which is somewhat surprising given the correct specification. Clearly, when s_R is used in estimating both parts of (4.17) then more resources allocated to the random reference survey would probably be advisable.

there exists $\boldsymbol{\delta}$ so that for each $k \in s^0$ the identity $\mathbf{X}^{0\prime}_k\boldsymbol{\delta} = \sigma^2_k$ holds where σ^2_k is the model unit variane.

4.4 Balanced Sampling, Calibration and Bias Robust Strategies

After discussing the mechanics of utilizing a reference survey as a surrogate of the general population, in this section I turn to the sampling design (of s^3 and s_R) part of our estimation strategy. Note that until now our discussion of the sequential framework has been in essence a post sampling estimation one and so here the question is whether there are certain sampling strategies which are beneficial in terms of precision and accuracy. This will bring our sequential framework discussion back to classic survey sampling where an *estimation strategy* includes a sampling design and an accompanying estimator optimal in terms such as unbiasedness and efficiency.

In the first section I return to the question of balance robust sampling for $m-$estimators and expand it to the sequential framework. Following that, I introduce first the idea of balance *random* sampling, or as I denote it π-balanced sampling, and show how it can be used to create robustness against m-model misspecification for πm-estimators, similar to the robustness in the case of m-estimators. In the final section, I detail a sequential $\pi-$balanced estimation strategy which can be implemented in the Web panel survey sampling problem. To achieve the required balance using a reference survey I will take advantage of additional tools such as the Calibration adjustment methods I briefly discussed in section 2.4.

4.4.1 Balanced Sampling Strategy for Bias Robust Under m-Sequential Estimation

In the following section I return to the question of balanced robust sampling for m-estimation I reviewed in section 2.8.4. Recall that the idea of balanced sampling is to create an additional layer of robustness to the $m-$estimators in the sense that for a large family of models, balanced sampling allows valid inference even under model misspecification of the $m-$model. Given our support of a sequential framework in the Web panel context, I will expand the balance strategy to cover the case when more than one phase of sample selection is assumed.

I start with a two phase example where a panel s^2 is self selected directly from the population s^0 and from it a survey sample s^3 is drawn. I then will give a description for the general $T-$phase case. To make the idea of balance explicit consider the case where two single-covariate m-models are assumed

$$m_2(\underline{x}^2) : \begin{cases} E(Y_k|\underline{x}^2, s^2) & = x_k^2\beta_2 \\ V(Y_k) & = x_k^2\sigma_2^2 \end{cases} \text{ and } m_0(x^0) : \begin{cases} E(m_{2k}|x^0, s^0) & = x_k^0\beta_0 \\ V(m_{2k}) & = x_k^0\sigma_0^2 \end{cases} \quad (4.18)$$

where $m_{2k} = x_k^2\beta_2$, which separately, at each phase, underline a classic ratio estimator so that the m-estimator results in

$$\begin{aligned}\hat{\overline{\mathbf{Y}}}_m &= \overline{\mathbf{X}}^0_{s^0}\frac{\overline{\hat{\mathbf{m}}}_{2s^2}}{\overline{\mathbf{X}}^0_{s^2}} \\ &= \frac{\overline{\mathbf{X}}^0_{s^0}}{\overline{\mathbf{X}}^0_{s^2}}\frac{\overline{\mathbf{X}}^2_{s^2}}{\overline{\mathbf{X}}^2_{s^3}}\overline{\mathbf{Y}}_{s^3}.\end{aligned}$$

If, however, the true models are in fact $E(Y_k|\underline{x}^2, s^2) = \alpha + x_k^2\beta_2 + x_k^0\beta_0$ denoted m_{2k} and $E(m_{2k}|x^0, s^0) = \tilde{\alpha} + x_k^0\tilde{\beta}_0$, the expectation of $\hat{\overline{\mathbf{Y}}}_m$ is

$$\begin{aligned}E(\hat{\overline{\mathbf{Y}}}_m) &= EEE((\frac{\overline{\mathbf{X}}^0_{s^0}}{\overline{\mathbf{X}}^0_{s^2}}\frac{\overline{\mathbf{X}}^2_{s^2}}{\overline{\mathbf{X}}^2_{s^3}}(\overline{\mathbf{Y}}_{s^3}|\underline{\mathbf{x}}^2_{s^2}, s^3)|\mathbf{x}^0_{s^0}, s^2)) \\ &= EEE((\frac{\overline{\mathbf{X}}^0_{s^0}}{\overline{\mathbf{X}}^0_{s^2}}\frac{\overline{\mathbf{X}}^2_{s^2}}{\overline{\mathbf{X}}^2_{s^3}}(\alpha + \overline{\mathbf{X}}^2_{s^3}\beta_2 + \overline{\mathbf{X}}^0_{s^3}\beta_0|\underline{\mathbf{x}}^2_{s^2}, s^3)|\mathbf{x}^0_{s^0}, s^2))\end{aligned}$$

which may be biased of $\overline{\mathbf{Y}}_{s^0}$. When within-panel balanced sampling is achieved- That is if $\overline{\mathbf{X}}^0_{s^3} = \overline{\mathbf{X}}^0_{s^2}$ and $\overline{\mathbf{X}}^2_{s^3} = \overline{\mathbf{X}}^2_{s^2}$ then

$$\begin{aligned}E(\hat{\overline{\mathbf{Y}}}_m) &= EE(\frac{\overline{\mathbf{X}}^0_{s^0}}{\overline{\mathbf{X}}^0_{s^2}}(\alpha + \overline{\mathbf{X}}^2_{s^2}\beta_2 + \overline{\mathbf{X}}^0_{s^2}\beta_0|\mathbf{x}^0_{s^0}, s^2)) \\ &= EE(\frac{\overline{\mathbf{X}}^0_{s^0}}{\overline{\mathbf{X}}^0_{s^2}}\overline{\mathbf{m}}_{2s^2}|\mathbf{x}^0_{s^0}, s^2) = E(\frac{\overline{\mathbf{X}}^0_{s^0}}{\overline{\mathbf{X}}^0_{s^2}}(\tilde{\alpha} + \tilde{\beta}\overline{\mathbf{X}}^0_{s^2}))\end{aligned}$$

which can be shown to be equal to $E(\overline{\mathbf{y}}_{s^0})$ if $\overline{\mathbf{X}}^0_{s^0} = \overline{\mathbf{X}}^0_{s^2}$.

As in the one phase case, the exact form of misspecification dictates the specific balance necessary to counter potential bias. If for example the panel level regression model does not include an intercept, that is $E(Y_k|\underline{x}^2, s^2) = x_k^2\beta_2 + x_k^0\beta_0$ then the balance required is only $\overline{\mathbf{X}}^0_{s^0} = \overline{\mathbf{X}}^0_{s^2}$. Similarly, if the overall population model would omit the interaction that is $E(Y_k|\underline{x}^2, s^2) = x_k^0\beta_0$ then no balance at all is required.

Under a T phase selection framework a sequential $m-$estimation bias robust strategy can be stated by the following. Assuming a general linear structure

$$m_{t-1}(\underline{\mathbf{x}}^{t-1}, s^{t-1}) : \begin{cases} E(m_{tk}|\underline{\mathbf{x}}^{t-1}, s^{t-1}) & = \underline{\mathbf{x}}^{t-1\prime}\boldsymbol{\beta}_{t-1} \\ V(m_{tk}|\underline{\mathbf{x}}^{t-1}, s^{t-1}) & = \sigma^2_{t-1k} \end{cases}$$

with independent errors for $t = 1, .., T$. If there exists in each phase a p_{t-1}column vector $\boldsymbol{\lambda}_{t-1}$ and selection design such that

$$\begin{cases} \sigma^2_{t-1k} & = \boldsymbol{\lambda}'_{t-1}\underline{\mathbf{x}}^{t-1}_k \\ \overline{\mathbf{x}}^{t-1}_{js^t} & = \overline{\mathbf{x}}^{t-1}_{js^{t-1}} \text{ for all } j = 1, ..., p_{t-1} \end{cases} \tag{4.19}$$

then $\hat{\overline{\mathbf{Y}}}_m = \overline{\mathbf{Y}}_{s^T}$. The idea then is that under a model where the variance function is a linear combination of covariates that are included in the regression equation the $m-$estimator reduces when the samples are balanced on the true model regression covariates as defined in 4.19 to the simple final phase sample average. As I have discussed above as long as we remain within this relatively wide ranging class of models, this property allows us to misspecify the model but still estimate the finite population without bias.

It is possible then, and clearly desirable, to outline a balanced sampling bias-robust strategy. However, in the context of our problem - the commercial Web panel survey sampling problem - is there any use of such a strategy? The problem is of course, that but the final stage, we have no control of the selection design at all. For example, in the $T = 2$ case that means that the balance $\overline{\mathbf{X}}^0_{s^0} = \overline{\mathbf{X}}^0_{s^2}$ cannot be actively met. Still, as I show in the next section, similar ideas described in the m-model context can be applied in the πm- type estimation approach where some degree of balance can be reached.

4.4.2 Sequential Robust π-Balanced Sampling Given Full Information

I move now to the idea of bias-robust balanced sampling for πm-estimation. First It is useful to distinguish between this idea and the double robust characteristic inherit to πm-estimators. The balanced sampling strategy identifies specific sample charectaristics which if met, for any such s, the estimator is unbiased with regards to the m- distribution even when the estimation model is misspecified. On the other hand by modeling over both m and π-models, DR estimators are unbiased over the joint πm- distribution even when the m-model is misspecified as long as the $\pi-$model is correctly specified.

$\pi m-$estimation does not condition inference on the sampled set s and so a balanced sample for πm-estimators must consider the selection process. Such samples have been studied in the survey sampling context and are called $\pi-$balanced samples. Also, an efficient sampling procedure, the Cube method, to achieve such balance in practice has been proposed. In the following I introduce first the idea of $\pi-$balanced sampling and the Cube method which can be used in implementing the within-Panel sampling design of s^3.

Deville and Tillé (2004); Tillé (2006) give the following definition for a π-balanced sample in a one phase framework. Let $p(s)$ denote the selection process (or when appropriate sampling design) which is assumed to generate the sample s. A *selection process $p(s)$ is said to be $\pi-$balanced* on the auxiliary variables

$X_1^0, ..., X_{p^0}^0$ if and only if it satisfies the balancing equations given by $\hat{\overline{\mathbf{X}}}_{\pi s}^0 = \overline{\mathbf{X}}_{s^0}^0$ which can also be written

$$\sum_s X_{jk}^0 \pi_k^{-1} = \sum_{s^0} X_{jk} \;\; j = 1, ..., p^0 \tag{4.20}$$

assuming here known π_k, $k \in s^0$, for all $s \in \mathcal{S}$ such that $p(s) > 0$.

By this definition, depending on the constraints 4.20, $\pi-$balanced sampling includes any random survey sampling design. For example any fixed size sample design (e.g. simple random sample of size n) can be defined as a random selection process balanced on the covariate $X_k = \pi_k$ because from 4.20 each s satisfies $\sum_s x_k \pi_k^{-1} = n$ the size of the sample. Another example is stratified sampling designs: first define indicator variable δ_{kh} equal to 1 if unit k is a member of strata s_h^0 where $s^0 = \sum s_h^0$ and $N = \sum_h N_h$ for $h = 1, ..., H$. Then, any $p(s)$ balanced on equations $\sum_s \delta_{kh} \pi_k^{-1} = \sum_{s_h} N_h / n_h = N_h$ is a stratified sample of n_h units from population strata s_h^0 of size N_h. A more familiar notation of the balancing equation would be $\hat{N}_{h\pi} = N_h$, $h = 1, .., H$ where $\hat{N}_{h\pi} = \sum_{s \cap s_h^0} \pi_k^{-1}$. The definition of balanced sampling design can even accommodate stratification on overlapping strata, something the clasic theory of stratified sampling designs does not allow since the stratification must be a partition of the population.

Balanced sampling is desirable, but is not trivial because of the combinatory explosion of the number of samples for large populations. To overcome this Deville and Till (2004); Ardilly (2006) propose the cube method, a class of sampling algorithms that selects a balanced sample and exactly satisfies a set of given inclusion probabilities. The cube method is a shortcut that avoids the enumeration of the samples, and is based on a random transformation of the vector of inclusion probabilities $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)$ until a sample is obtained such that:

1. the inclusion probabilities are exactly satisfied,
2. the balancing equations are satisfied to the furthest extent possible.

Its name comes from the geometric representation of the sample indicators S_k $k = 1, ..., N$ as a vertex of an N-cube as as showed in the left panel of figure 4.5. The balancing equations $\sum_{s^0} S_k \frac{X_{jk}^0}{\pi_k} = \sum_{s^0} X_{jk}$, $j = 1, .., P$ with unknowns values S_k define an affine subspace in $\mathcal{R}^N$ of dimension $N - P$ denoted by Q, where

$$Q = \{\mathbf{s} \in \mathcal{R}^N | \sum_{s^0} S_k \frac{X_{jk}^0}{\pi_k} = \sum_{s^0} X_{jk}\}$$

A balanced sampling design thus consists of choosing a vertex of the N-cube (a sample) that remains on the linear sub-space Q as shown in the right panel of

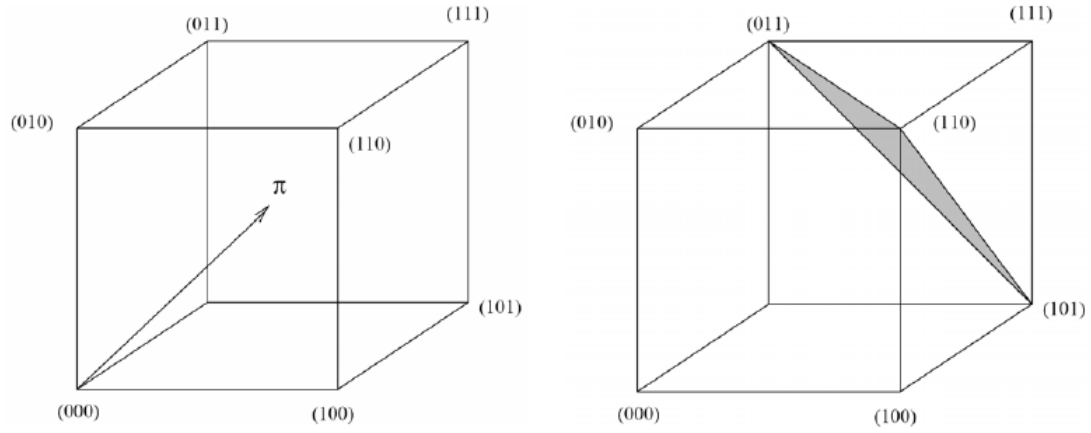


Figure 4.5: On the left hand side a geometric representation (Deville and Tillé, 2004) of all potential samples in a population of size $N = 3$. On the right hand side the case where a constraint of $n = 2$ is applied. There are thre vertices of the cube which remain on the linear sub space defined by the balance constraint.

4.5 where the balancing equations can be met exactly.

The cube method has been implemented in R package 'sampling'. The application has no limit as far as population size is concerned, and can accomodate up to $P = 40$ balancing variables. However, computation time increases with $N \times P^2$. See Till and Matei (2007) for discussion of the method and R package application including approximations when exact balancing is not possible.

We are ready to discuss a bias robust strategy for πm-estimator. I start with a simple example under a single covariate one phase framework where the analyst assumes working model $m_0 : E(Y_k|x^0) = x^0_k\beta_0$ and $V(Y_k|x^0) = x^0_k\sigma^2$ with independent errors, the population s^0 is observed entirely and unit selection probabilities are known. Under m_0 and assuming $\overline{\mathbf{X}}^0_{s^0}$ and $\boldsymbol{\pi}_{s^0} = (\pi_1, ..., \pi_N)$ are known, the $\pi m-$estimator is $\hat{\overline{\mathbf{Y}}}_{\pi m} = \overline{\mathbf{X}}^0_{s^0}\hat{\overline{\mathbf{Y}}}_{\pi}/\hat{\overline{\mathbf{X}}}^0_{\pi}$ the weighted ratio estimator. If, however, the true population model is $\tilde{m}_0 : E(Y_k|x^0) = \alpha + x^0_k\beta_0$ then the $m-$bias of $\hat{\overline{\mathbf{Y}}}_{\pi m}$ is

$$B_{\tilde{m}_0}(\hat{\overline{\mathbf{Y}}}_{\pi m}) = \alpha(\overline{\mathbf{X}}^0_{s^0} - \hat{\overline{\mathbf{X}}}^0_{\pi})/\hat{\overline{\mathbf{X}}}^0_{\pi}.$$

Still, when s is $\pi-$balanced such that $\hat{\overline{\mathbf{X}}}^0_{\pi} = \overline{\mathbf{X}}^0_{s^0}$ then $B_{\tilde{m}_0}(\cdot) = 0$. Further note that as the cube method randomly selects the sample so that all $\pi_k \ ; k \in s^0$ are satisfied exactly the double robust property still holds and so balanced sampling design may offer an additional layer of robustness against m- misspecification.

Now, we can quickly establish a general definition of a bias-robust strategy under a single phase πm-estimation framework. Consider as before $m-$estimators under a general linear model

$$m(\mathbf{x}^0):\begin{cases}E(Y_k) & =\mathbf{x}_k^{0\prime}\boldsymbol{\beta}_0\\ V(Y_k) & =\sigma_{0k}^2.\end{cases}\tag{4.21}$$

where $\mathbf{x}_k^0$ and $\boldsymbol{\beta}_0$ are p-vectors and assuming $\mathbf{x}_{s^0}^0$ is observed. If for all samples s over the sampling space there exists a constant column vector $\boldsymbol{\lambda}_0$ of dimension p_o not depending on k such that for all $k \in s^0$

$$\begin{cases}\sigma_{0k}^2 & =\boldsymbol{\lambda}_0^{\prime}\mathbf{x}_k^0\\ \hat{\overline{\mathbf{x}}}_{j\pi}^0 & =\overline{\mathbf{x}}_{js^0}^0 \text{ for all } j=1,...,p_0\end{cases}\tag{4.22}$$

where $\hat{\overline{\mathbf{x}}}_{j\pi}^0$ is the normal π-estimator with denominator N or $\hat{N}_\pi=\sum_s \hat{\pi}_k^{-1}$ then $\hat{\overline{\mathbf{Y}}}_{\pi m}=\hat{\overline{\mathbf{Y}}}_{\pi}$. The proof follows the same steps as in the m-estimation approach.

The intuition behind this result can be understood best under a single categorical covariate model. When X_k^0 is univariate and categorical $\hat{\overline{\mathbf{Y}}}_{\pi m}$ is simply a post stratification estimator. If $p(s)$ represents a stratified sampling design then the post stratification procedure $\hat{\overline{\mathbf{Y}}}_{\pi m}$ is redundant as we are post stratifying an already stratified sample. The bias robust strategy generalizes this idea as πm-estimators are a general form of post stratification (and a member of the general class of calibration estimators, ass disscussed in chapter 1), while balanced sampling designs generalize the idea of stratification by allowing the use of continuous covariates and overlapping strata.

The implication then is that under the assumptions 4.21-4.22 estimator $\hat{\overline{\mathbf{Y}}}_{\pi m}$ is $m-$unbiased even under misspecification for a wide class of models which all converge to the same $\pi-$estimator. Furthermore, if $\hat{\overline{\mathbf{Y}}}_{\pi m}$ is computed under a correct $\pi-$model then it is $(\pi-)$ unbiased even when assumptions 4.21-4.22 are not met.

Finally, we can extend these ideas to the sequential πm-estimation framework. Consider a T-phase ignorable framework where

$$m_{t-1}(\underline{\mathbf{x}}^{t-1},s^{t-1}):\begin{cases}E(m_{tk}|\underline{\mathbf{x}}^{t-1},s^{t-1}) & =\underline{\mathbf{x}}^{t-1\prime}\boldsymbol{\beta}_{t-1}\\ V(m_{tk}|\underline{\mathbf{x}}^{t-1},s^{t-1}) & =\sigma_{t-1k}^2\end{cases}$$

$$\pi_t(\underline{\mathbf{x}}^{t-1})\ :\begin{cases}p(s_k^t=1|\underline{\mathbf{x}}^{t-1},\underline{s}^{t-1}) & =\pi_t(\underline{\mathbf{x}}_k^{t-1};\boldsymbol{\alpha}^{t-1})\\ p(s_k^t=1|\underline{\mathbf{x}}^{t-1},\underline{s}^{t-1}) & >0\end{cases}$$

where here $\pi_t(\underline{\mathbf{x}}_k^{t-1};\boldsymbol{\alpha}^{t-1})=exp(\underline{\mathbf{x}}_k^{t-1\prime}\boldsymbol{\alpha}^{t-1})/1+exp(\underline{\mathbf{x}}_k^{t-1\prime}\boldsymbol{\alpha}^{t-1})$ follow a logistic distribution over the sub population s^{t-1} and with $m_{t-1}(\underline{\mathbf{x}}^{t-1},s^{t-1})$ having independent errors for $t=1,..,T$. If there exists in each phase a p_{t-1}column vector

$\boldsymbol{\lambda}_{t-1}$ and selection design such that

$$\begin{cases} \sigma^2_{t-1k} & = \boldsymbol{\lambda}^{'}_{t-1}\underline{\mathbf{x}}^{t-1}_k \\ \hat{\bar{\underline{\mathbf{x}}}}^{t-1}_{j\pi s^t} & = \hat{\bar{\underline{\mathbf{x}}}}^{t-1}_{j\pi s^{t-1}} \text{ for all } j = 1, ..., p_{t-1} \end{cases} \tag{4.23}$$

where $\hat{\bar{\underline{\mathbf{x}}}}^{t-1}_{j\pi s^t} = \hat{\bar{\underline{\mathbf{x}}}}^{t-1}_{j\pi s^{t-1}}$ means that $\frac{\sum_{s^t} \underline{\mathbf{X}}^t_k \underline{\hat{\pi}}^{-1}_{tk}}{\sum_{s^t} \underline{\hat{\pi}}^{-1}_{tk}} = \frac{\sum_{s^{t-1}} \underline{\mathbf{X}}^t_k \underline{\hat{\pi}}^{-1}_{t-1k}}{\sum_{s^{t-1}} \underline{\hat{\pi}}^{-1}_{t-1k}}$ then $\hat{\bar{\mathbf{Y}}}_{\pi m} = \hat{\bar{\mathbf{Y}}}_{\pi}$. In words the strategy states that under a sequence of linear $m-$models with variance structures that can be represented as a combination of the regression covariates we can achieve additional robustness to $m-$misspecification by balancing the $m-$model covariates in the sense that for each phase $t = 1, .., T$ the $\pi-$estimator of the $m-$model covariates s^0 population average $\hat{\bar{\underline{\mathbf{X}}}}^{t-1}_{j\pi s^t}$ is equal to $\hat{\bar{\underline{\mathbf{X}}}}^{t-1}_{j\pi s^{t-1}}$ the same covariate average calculated over the higher subpopulation.

As a simple example take the case of $T = 2$. The balancing criteria require that the set s^1 is such that $\hat{\bar{\mathbf{X}}}^0_{\pi_1 s^1} = \bar{\mathbf{X}}^0_{s^0}$ which is similar to the classic one phase case. In addition we require the set s^2 to have the property

$$\hat{\bar{\underline{\mathbf{X}}}}^1_{\pi_2 s^2} = \hat{\bar{\mathbf{X}}}^1_{\pi_1 s^1} \tag{4.24}$$

which can be given two different interpretations. First it states directly that s^2 is such that the two phase π-estimator of the general population averages calculated over s^2 for covariates $\underline{\mathbf{X}}^1$ is exactly equal to the one phase π-estimator calculated over s^1 of the same population quantities. Secondly, (4.24) implies that our first phase balance still holds, that is for the general population covariates $\mathbf{X}^0$, a subset of $\underline{\mathbf{X}}^1$, the $\pi-$estimator is equal exactly to the population averages.

Finally note that the $\pi-$models do not necessary need to be specified correctly or estimated consistently although for $m-$bias robustness to hold. Of course for the double robust property to hold the $\pi-$models do need to be correctly specified and consistently estimated. Also note that by including $\mathbf{X}^0_k = 1$ the selection process is of fixed size in the sense that each phase t projects by $\pi-$estimation the subset population size of s^{t-1} including that of the general population, a desirable property especially when dealing with population statistics.

4.4.3 The Application of π-Balanced Strategy when Reference Survey is Used

Let us now understand how a $\pi-$balanced bias robust strategy can be applied to our main question of estimation over a Web panel survey sample, assisted by a random reference survey. Under a three phase sequential framework, the main issues in practice are

(i) we control only the within-panel survey sampling selection process $p(s^3|s^2)$ while the first two phases are entirely self driven, and

(ii) for estimation, we observe only a subset of $(\mathbf{x}^{t-1}_{s^{t-1}}, s^t) \quad t = 1, .., 3$. At the most the subset s is a union of s^2 and a random reference survey s^3.

Recall that had $\underline{d}^3_{s^3} = (\underline{s}^3, \underline{\mathbf{x}}^2_{s^2}, \mathbf{y}_{s^3})$ been observed fully then a πm-estimator is

$$\hat{\bar{\mathbf{Y}}}_{\pi m_3} \quad = \quad N^{-1}\mathbf{1}^{'}_{s^0}\mathbf{H}^0_{\pi s^0, s^1}\mathbf{H}^1_{\pi s^1, s^2}\mathbf{H}^2_{\pi s^2, s^3}\mathbf{y}_{s^3} \tag{4.25}$$

where $\mathbf{H}^{t-1}_{\pi s^{t-1} s^t} = \underline{\mathbf{x}}^{t-1}_{s^{t-1}}\mathbf{A}^{t-1}_{\pi s^t}\underline{\mathbf{x}}^{t-1'}_{s^t}\mathbf{w}^{t-1}_{s^t}\underline{\hat{\boldsymbol{\pi}}}^{-1}_{ts^t}$ for $t = 1, ..., 3$ with $\mathbf{A}^{t-1}_{\pi s^t} = (\underline{\mathbf{x}}^{t-1'}_{s^t}\mathbf{w}^{t-1}_{s^t}\underline{\hat{\boldsymbol{\pi}}}^{-1}_{ts^t}\underline{\mathbf{x}}^{t-1}_{s^t})^{-1}$ and $\underline{\hat{\boldsymbol{\pi}}}^{-1}_{ts^t}{=}diag(\hat{\pi}^{-1}_{tk})^{-1}$ a $n_t \times n_t$ diagonal matrix of estimated selection probabilities calculated separately.

However, as discussed in detail earlier in this chapter, only a disjointed subset of $\underline{d}^3_{s^3}$ is observed, so we estimate (4.25) by

$$\hat{\bar{\mathbf{Y}}}_{\pi m_3} = \hat{N}^{-1}_{\pi}\mathbf{1}^{'}_{n}\hat{\mathbf{H}}^0_{\pi s, s\cap s^1}\hat{\mathbf{H}}^1_{\pi s\cap s^1, s^2}\mathbf{H}^2_{\pi s^2, s^3}\mathbf{y}_{s^3} \tag{4.26}$$

where $\mathbf{H}^2_{s^2, s^3}$ is defined as for (4.25) while

$$\begin{aligned} \hat{\mathbf{H}}^0_{\pi s, s\cap s^1} \quad &= \quad \mathbf{x}^0_s\boldsymbol{\pi}^{-1}_s(\mathbf{x}^{0'}_{s\cap s^1}\boldsymbol{\pi}^{-1}_{s\cap s^1}\mathbf{w}^0_{s\cap s^1}\hat{\boldsymbol{\pi}}^{-1}_{1s\cap s^1}\mathbf{x}^0_{s\cap s^1})^{-1}\mathbf{x}^{0'}_{s\cap s^1}\boldsymbol{\pi}^{-1}_{s\cap s^1}\mathbf{w}^0_{s\cap s^1}\hat{\boldsymbol{\pi}}^{-1}_{1s\cap s^1} \\ \hat{\mathbf{H}}^1_{\pi s\cap s^1, s^2} \quad &= \quad \underline{\mathbf{x}}^1_{s\cap s^1}\boldsymbol{\pi}^{-1}_{s\cap s^1}(\underline{\mathbf{x}}^{1'}_{s^2}\mathbf{w}^1_{s^2}\hat{\boldsymbol{\pi}}^{-1}_{2s^2}\underline{\mathbf{x}}^1_{s^2})^{-1}\underline{\mathbf{x}}^{1'}_{s^2}\mathbf{w}^1_{s^2}\hat{\boldsymbol{\pi}}^{-1}_{2s^2} \end{aligned}$$

where $s \cap s^1 = s^1_R \cup s^2$ and with $\mathbf{w}^2_{s^3} = diag(\sigma^2_{2k})^{-1} \ k \in s^3$, $\mathbf{w}^1_{s^2} = diag(\sigma^2_{1k})^{-1}k \in s^2$ and $\mathbf{w}^0_{s\cap s^1} = diag(\sigma^2_{0k})^{-1}k \in s \cap s^1$.

Given the density of the notation I restate our objective again in simple terms: After outlining the πm-estimator (4.26) calculated over $s_R \cup s^2$ I discuss now the additional procedures we take to shape the available sample data. The idea behind these procedures is to create (under certain assumption on the m-models) an equilibrium between $\hat{\bar{\mathbf{Y}}}_{\pi m_3}$ and $\hat{\bar{\mathbf{Y}}}_{\pi_3}$ the three phase $\pi-$estimator. As I have shown above, this will give us an additional layer of robustness to m-model misspecification, on top of the double robust property which requires an exactly correct m or π model specification. In the following it is demonstrated that in the ideal case of (4.25) the application is relatively straightforward, while for (4.26) the procedure is substantially more involved.

The estimation strategy starts with phase $t = 1$. Under full information $\mathbf{H}^0_{\pi s^0, s^1}$ is calculated over s^0, and for balance robustness we would require (refer back to 4.23) that m_0 is such that $\sigma^2_{0k} = \boldsymbol{\lambda}_0^{'}\mathbf{x}^0$ and s^1 is such that $\overline{\mathbf{X}}^0_{s^0} = \hat{\overline{\mathbf{X}}}^0_{s^1\pi_1}$ where $\hat{\overline{\mathbf{X}}}^0_{s^1\pi_1} = \sum_{s^1}\mathbf{X}^0_k\hat{\pi}^{-1}_{1k}/\sum_{s^1}\hat{\pi}^{-1}_{1k}$. When the set of covariates $\mathbf{X}^0$ includes the constant 1 the denominator is redundant as $\sum_{s^1}\hat{\pi}^{-1}_{1k} = N$ by definition. In practice, however, we calculate $\hat{\mathbf{H}}^0_{\pi s, s\cap s^1}$ over $s^1_R \cup s^2$ and so for the bias robust strategy to hold we require at this stage that $\hat{\overline{\mathbf{X}}}^0_{s\pi} = \hat{\overline{\mathbf{X}}}^0_{s^1_R\cup s^2\pi}$ which in more detail means the balancing equations are

$$\frac{\sum_s \mathbf{X}^0_k\pi^{-1}_k}{\sum_s \pi^{-1}_k} = \frac{\sum_{s^1_R\cup s^2}\mathbf{X}^0_k\hat{\pi}^{-1}_{1k}\pi^{-1}_k}{\sum_{s^1_R\cup s^2}\hat{\pi}^{-1}_{1k}\pi^{-1}_k} \tag{4.27}$$

where as before $s = s^2 \cup s_R$ and $\pi_k = \pi_R$ for members of s_R and $\pi_k = 1$ for panel member units[5]. It is important to stress here that for m- unbiasedness to hold we require as well that we are balancing towards the correct finite population quantity, that is $\hat{\overline{\mathbf{X}}}^0_{s\pi}$ is in fact equal to $\overline{\mathbf{X}}^0_{s^0}$.

Regardless of whether we balance s^1 or $s^1_R \cup s^2$, fundamentally, the selection process generating s^1 is not controlled and we must resort to calibration. As discussed in chapter 1, the term calibration refers in the classic survey sampling context to a family of post sample adjustments to the selection weights so that a calibrated π-weighted sample estimate is equal to known population quantities. In the specific case of *linear calibration* the application can be shown explicitly. In our case the linear calibration weights are the scalars

$$g_{1k} = (\sum_s \mathbf{x}^0_k\pi^{-1}_k)'(\sum_{s^1_R\cup s^2}\mathbf{x}^0_k\hat{\pi}^{-1}_{1k}\pi^{-1}_k\mathbf{x}^{0'}_k)^{-1}\mathbf{x}^0_k$$

which achieve balance requirement of (4.27) in the sense that

$$\frac{\sum_s \mathbf{X}^0_k\pi^{-1}_k}{\sum_s \pi^{-1}_k} = \frac{\sum_{s^1_R\cup s^2} g_{1k}\mathbf{X}^0_k\hat{\pi}^{-1}_{1k}\pi^{-1}_k}{\sum_{s^1_R\cup s^2} g_{1k}\hat{\pi}^{-1}_{1k}\pi^{-1}_k}$$

or when the sample size is also calibrated, the equations simplify to $\sum_s \mathbf{X}^0_k\pi^{-1}_k = \sum_{s^1_R\cup s^2} g_{1k}\mathbf{X}^0_k\hat{\pi}^{-1}_{1k}\pi^{-1}_k$. In words, we calibrate the π-estimator over $s^1_R \cup s^2$ to the finite population estimator calculated over $s = s_R \cup s^2$, which in turn we assume estimates exactly the population average $\overline{\mathbf{X}}^0_{s^0}$. With regards to the latter point an important recommendation would be then to request from the reference survey sampling statistician to design the selection of s_R on correlated s^0 population quantities, including the population size. If this is not possible directly, an alternative is to include these covariates in the calibration model on the refernce selection weights π^{-1}_{Rk}. In practical terms, a simple application of the process can

[5] If $m_0(\mathbf{x}^0)$ meets the variance structure constraint as stated in (4.23) this balance will lead to the canceling out the components of $\hat{N}^{-1}_\pi\mathbf{1}_n^{'}\hat{\mathbf{H}}^0_{\pi s, s\cap s^1}$.

be done with the use of the *calib* function included in the *survey* package[6] in R. This procedure results in an adjusted component $\hat{\mathbf{H}}^0_{\pi s, s\cap s^1}$ of $\hat{\overline{\mathbf{Y}}}_{\pi m_3}$

$$\hat{\mathbf{H}}^0_{\pi s, s\cap s^1} = \mathbf{x}^0_s \boldsymbol{\pi}^{-1}_s (\mathbf{x}^{0'}_{s\cap s^1} \mathbf{g}_{1s\cap s^1} \boldsymbol{\pi}^{-1}_{s\cap s^1} \mathbf{w}^0_{s\cap s^1} \hat{\boldsymbol{\pi}}^{-1}_{1s\cap s^1} \mathbf{x}^0_{s\cap s^1})^{-1} \mathbf{x}^{0'}_{s\cap s^1} \mathbf{g}_{1s\cap s^1} \boldsymbol{\pi}^{-1}_{s\cap s^1} \mathbf{w}^0_{s\cap s^1} \hat{\boldsymbol{\pi}}^{-1}_{1s\cap s^1}$$

where $\mathbf{g}_{1s\cap s^1}$ is the diagonal matrix with typical value g_{1k} for $k \in s^1_R \cup s^2$. For completeness note that the linear calibration weights here can be viewed as estimators of calibration weights $g_{1k} = (\sum_{s^0} \mathbf{x}^0_k)' (\sum_{s^1} \mathbf{x}^0_k \hat{\pi}^{-1}_{1k} \mathbf{x}^{0'}_k)^{-1} \mathbf{x}^0_k$ we would calculate under full information.

For phase $t = 2$ we take a similar approach but as s^2 is observed in full this leads to a slightly less complicated procedure in practice. Consider first the case where s^1 and s^2 and $\underline{\mathbf{X}}^1_{s^1}$ are fully observed. Under a bias robust strategy and given an appropriate regression and variance function for balance robustness we require that $\hat{\overline{\mathbf{X}}}^1_{\pi s^2} = \hat{\overline{\mathbf{X}}}^1_{\pi s^1}$ for all s^2 given known s^1 which explicitly can be written as

$$\frac{\sum_{s^2} \underline{\mathbf{X}}^1_k \hat{\pi}^{-1}_{2k}}{\hat{N}_{s^2 \pi_2}} = \frac{\sum_{s^1} \underline{\mathbf{X}}^1_k \hat{\pi}^{-1}_{1k}}{\hat{N}_{s^1 \pi_1}}$$

where $\hat{N}_{s^2\pi_2} = \sum_{s^2} \hat{\underline{\pi}}^{-1}_{2k}$ and given s^1 is sampled by balanced design implies $\hat{N}_{s^1\pi_1} = N$. An easier way to consider this (and a technical solution to sequential balancing using a one phase software function) is to define $\underline{\mathbf{Z}}^1_k = \underline{\mathbf{X}}^1_k \hat{\pi}^{-1}_{1k}$. This means that the balanced design is constrained so that

$$\frac{\sum_{s^2} \underline{\mathbf{Z}}^1_k \hat{\pi}^{-1}_{2k}}{\hat{n}_{1s^2}} = \sum_{s^1} \underline{\mathbf{Z}}^1_k / n_1 \tag{4.28}$$

where $\hat{n}_{1s^2} = \sum_{s^2} \hat{\pi}^{-1}_{2k}$ a second phase $\pi-$estimator of the size of s^1. This turns the balancing problem into a simple balancing of the selection process of s^2 from the (sub)population s^1. As before I note that for the subset $\mathbf{X}^0$ of covariates (4.28) implies balance to $\overline{\mathbf{X}}^0_{s^0}$. Now, as s^2 is self selected, we turn again to calibrate the available set and here this would take the form

$$\frac{\sum_{s^2} \underline{\mathbf{Z}}^1_k g_{2k} \hat{\pi}^{-1}_{2k}}{\hat{n}_{1s^2}} = \sum_{s^1} \underline{\mathbf{Z}}^1_k / n_1$$

where now $\hat{n}_{1s^2} = \sum_{s^2} g_{2k} \hat{\pi}^{-1}_{2k}$ and if linear calibration is used

$$g_{2k} = (\sum_{s^1} \underline{\mathbf{Z}}^1_k)' (\sum_{s^2} \underline{\mathbf{Z}}^1_k \hat{\pi}^{-1}_{2k} \underline{\mathbf{Z}}^{1'}_k)^{-1} \underline{\mathbf{Z}}^1_k.$$

In practice only a subset of the Web population s^1 is observed, that is $s^1_R \cup s^2$ along with the weights π_k's. Note as well that $p(s^1)$ is not balanced but rather

[6] T. Lumley (2012) 'survey: analysis of complex survey samples'. R package version 3.28-2. and T. Lumley (2004) Analysis of complex survey samples. Journal of Statistical Software 9(1): 1-19.

calibrated, so we start first with redefining $\underline{\mathbf{Z}}_k^1 = \underline{\mathbf{X}}_k^1 \hat{\pi}_{1k}^{-1} \pi_k^{-1} g_{1k}$ and then setting the calibration equations on the observed subsets leading to

$$\sum_{s^2} \underline{\mathbf{Z}}_k^1 g_{2k} \hat{\pi}_{2k}^{-1} = \sum_{s_R^1 \cup s^2} \underline{\mathbf{Z}}_k^1$$

where in the case a linear calibration is used the second phase calibration weights are defined as before $g_{2k} = (\sum_{s^1} \underline{\mathbf{Z}}_k^1)' (\sum_{s^2} \underline{\mathbf{Z}}_k^1 \hat{\pi}_{2k}^{-1} \underline{\mathbf{Z}}_k^{1'})^{-1} \underline{\mathbf{Z}}_k^1$ but this time when $\underline{\mathbf{Z}}_k^1 = \underline{\mathbf{X}}_k^1 \hat{\pi}_{1k}^{-1} \pi_k^{-1} g_{1k}$. Here, as in the first phase we require as well that the reference based sample equals the full information quantity, that is $\sum_{s_R^1 \cup s^2} \underline{\mathbf{Z}}_k^1 = \sum_{s^1} \underline{\mathbf{Z}}_k^1 / n_1$ for robustness to hold.

Finally for the last phase $t = 3$ in the Web panel context this is a within panel survey sampling exercise where balance can be implemented directly through sampling design by the cube method. In simple terms and assuming full information in the first two phases we define the strategy as sampling s^3 by design $p(s^3|s^2)$ with unit inclusion probabilities π_{3k} so that

$$\sum_{s^3} \underline{\mathbf{X}}_k^2 \underline{\pi}_{3k}^{-1} = \sum_{s^2} \underline{\mathbf{X}}_k^2 \underline{\hat{\pi}}_{2k}^{-1} \tag{4.29}$$

for all s^3 where $\underline{\pi}_{3k}^{-1} = (\hat{\pi}_{1k} \cdot \hat{\pi}_{2k} \cdot \pi_{3k})^{-1}$. If we let $\underline{\mathbf{Z}}_k^2 = \underline{\mathbf{X}}_k^2 \underline{\hat{\pi}}_{2k}^{-1}$ then the balancing equations (4.29) are simply

$$\sum_{s^3} \underline{\mathbf{Z}}_k^2 \pi_{3k}^{-1} = \sum_{s^2} \underline{\mathbf{Z}}_k^2$$

and simple one phase application such as in the *cube* R function can be applied directly. Given the use of the reference survey in the first two phases we need to redefine the balancing covariates to include calibration weights so that $\underline{\mathbf{Z}}_k^2 = \underline{\mathbf{X}}_k^2 \underline{\hat{\pi}}_{2k}^{-1} \pi_k^{-1} \underline{g}_{2k}$ where $\underline{g}_{2k}$ and $\underline{\hat{\pi}}_{2k}^{-1}$ are the product of the first two phase calibration weights and estimated selection probabilities respectively. Given the re-defined $\underline{\mathbf{Z}}_k^2$ the balancing constraints for the sampling design are as above, that is $\sum_{s^3} \underline{\mathbf{Z}}_k^2 \pi_{3k}^{-1} = \sum_{s^2} \underline{\mathbf{Z}}_k^2$. As s^2 is fully observed the selection can be technically applied by the cube method.

A final comment on the $\pi-$balanced strategy in practice is that assuming a large reference survey, for example when s_R is an establishment survey, the estimator variance will be dominated by the final within-panel survey sampling phase. Such cases emphasize the benifit of applying the balancing constraints. By its definition a balanced sampling design is a restricted sampling procedure and reduces the sampling variance of the estimator.

4.5 Summary and Final Words

In an ongoing march, for over a decade now, Web based surveys and in particular Web access panels have transformed the landscape of survey based research and have become a key tool for the study of general populations. With the deepening of the Internet and associated information and communication technologies into our day to day life, the importance of Web surveys will continue to grow.

While being a practitioner of Web panel surveys, my aim in this work has not been to defend the scientific validity of the methodology nor to be a proponent of the platform, but rather the initial goal I set out had been to describe the myriad ad hoc estimation procedures practiced in commercial settings through a detailed mathematical and methodological presentation. By this I aimed to potentially offer ways of improving existing approaches.

This is not the first attempt to put Web access panels on a firmer statistical footing, and this work has certainly built on available published research. Previous investigation into the Web panel question focused on demonstrating the limitations of traditional survey sampling techniques in compensating for the errors innate to this survey method. Such studies (for exmple Isaksson et al., 2004; Lee, 2006; Dever et al., 2008; Rivers 2009; Valliant and Dever 2011) went on to introduce to the survey sampling community novel statistical adjustments such as propensity scores weighting or sampling by statistical matching and demonstrated their superior applicability to the Web panel question compared to classical methods such as post stratification.

Through a careful and structured investigation I have expanded on these ideas, and offer in this work several contributions:

First, I intended to explain more clearly common approaches practiced in Web survey sampling estimation procedures. For example, purposive sampling from the Web panel, based on population statistics, can be understood as a robust model based approach strategy under a general class of linear models. Another commonly used procedure is the calibration of propensity score adjustment weights to available (or estimated) population statistics. Previous research (such as Lee, 2004) described this as a combination of coverage bias correction (calibration) and selection bias correction (propensity score). Others offered it as a general selection bias adjustment (propensity score) with a second correction layer (calibration) aimed at variance reduction. Here I have shown that when viewing calibration as an $m-$model estimator, the combination of it with propensity score weighting (a $\pi-$estimator) puts the procedure within the large class of $\pi m-$estimators. This class of estimators has been shown to be both more efficient then $\pi-$estimators and offer a double robust property that can be argued to have a better potential for bias adjustment (Bang and Robins 2005; Robins et al., 2007).

The second contribution of this work is the introduction of the sequential framework to Web panel survey sampling, and a suit of associated sequential based estimators. I have argued that the framework has several benefits. (i) The sequential framework represents much better the real recruitment and selection process of a member of the general population into the final survey set. This way, when taking a $\pi-$estimation approach, the researcher modeling the selection process can use the strong operational understanding of the different phases of the process, as well as the entirely known within-panel sampling design to inform better the overall $\pi-$ model. (ii) An important result of adapting a sequential framework is the availability of Internet and panel associated variables in the modeling and estimation stages. When assuming a single phase model, variables such as Web usage, Online consumer characteristics or even panel response behavior cannot be utilized as they are not independent of the selection process. However, when the underlying (π or m type) models are deconstructed into a sequence of conditional models these dependencies are broken and we can build potentially better preforming models. (iii) I have also outlined in a clear mathematical presentation a sequential πm-estimation strategy which combines a procedure of sampling from the panel a $\pi-$balanced sample on a set of population statistics. The combination of a $\pi m-$estimator with a the balanced sampling procedure adds an additional layer of robustness, on top of the DR property of the estimator. Furthermore, this πm-balanced strategy is general enough to include most of the common practiced estimation approaches such as propensity score weighting, purposive sampling or propensity score weighting and calibration and more. (iv) I describe a detailed algorithm that implements these estimation strategies to the more practical scenario where a random reference survey sample replaces the unknown population distribution.

To test the performance of the estimation procedures and discuss questions facing researchers in practice I conducted several simulation studies. These included an investigation of the effect of the sample size ratio between the Web panel and reference survey set. Another study was to examine the common practice of coding zero ,the values of Internet related variables for non Internet members of the population. I also tested a simple bootstrap estimator for the variance of different estimators under a sequential framework.

The separate modeling of the sequential population and sample sets, defined by the selection phases, potentially captures more accurately the true underlying distributions. However it can be argued, that the introduction of these additional models bring with it a higher risk of bias due to model misspecification. It is also clear that in the case of relying on a sequential $\pi-$model, there is an inevitable addition of estimation variance. I expect that both bias from model misspecification and higher variance from the additional layers of modeling may be an issue when applying these estimation approaches to empirical data. These observations are hoped to guide future research on the topic.

Chapter 5

Appendix

5.1 The Influence Function

The influence function (2.32) of the population average is

$$\varphi(d_k, \overline{\mathbf{y}}_{s^0}) = X_k\hat{\beta} - \overline{\mathbf{Y}}_{s^0} + S_k(Y_k - X_k\hat{\beta}) \tag{5.1}$$

and the variance of the estimator is

$$\begin{aligned} V(\hat{\overline{\mathbf{y}}}_{s^0}) &= E(\varphi(d_k, \overline{\mathbf{y}}_{s^0})^2)/N \\ &= V(\sum_{s^0} \varphi(d_k, \overline{\mathbf{y}}_{s^0})/N^2 \\ &= V(\sum_{\overline{s}} \hat{y}_k - \sum y_k)/N^2 \\ &= \sum_{s} V(Y_k)(\frac{\overline{\mathbf{x}}_{\overline{s}}}{\overline{\mathbf{x}}_s})^2 + \sum_{\overline{s}} V(Y_k) \end{aligned}$$

where the first line is true asymptotically by definition of an influence function under regularity conditions and the second is true as $\sum_{s^0} E(\varphi_k^2) = \sum_{s^0} V(\varphi_k)$ and while $E(\varphi_k^2) = V(\hat{\overline{\mathbf{y}}}_{s^0})$ then $V(\hat{\overline{\mathbf{y}}}_{s^0}) = \sum_{s^0} V(\varphi_k)/N$. More generally, when the linear model is $Y_k = \boldsymbol{\beta}'\mathbf{x}_k + \varepsilon_k$, then by similar derivation

$$\begin{aligned} V(\hat{\overline{\mathbf{y}}}_{s^0}) &= N^{-2}[V(\mathbf{a}_s'\mathbf{Y}_s) + V(\mathbf{1}_{\overline{s}}'\mathbf{Y}_{\overline{s}})] \\ &= N^{-2}[\sum_{k\in s} a_k^2 V(Y_k) + \sum_{k\in\overline{s}} V(Y_k)] \end{aligned} \tag{5.2}$$

where $a_k = \mathbf{1}_{\overline{s}}'\mathbf{x}_{\overline{s}}\mathbf{A}_s^{-1}\mathbf{x}_k v_k^{-1}$ with $\mathbf{A}_s = \mathbf{x}_s'\mathbf{V}_{ss}^{-1}\mathbf{x}_s$.

(Valliant et al., 2000, chapter 5) derive this identical variance as part of their BLUP theory by developing directly the prediction error variance rather than using the influence function. Dorfman (1991) shows that normally (5.2) is dominated by the first component, that associated with the sampled set, for regular case where the sample set is negligible in size to the population. More specifically, Dorfman (1991) and (Valliant et al., 2000, chapter 5) show that

$$\begin{aligned} V(\hat{\bar{\mathbf{y}}}_{s^0}) &= N^{-2}[\sum_{k \in s} a_k^2 V(y_k) + \sum_{k \in \bar{s}} V(y_k)] \\ &\approx O\left(\frac{(N-n)^2}{n}\right) + O(N-n) \end{aligned}$$

The structure of the approximate variance in (5.2) suggests a sandwich estimator which is found by replacing the unknown $V(y_k)$'s by the observed square errors r_k^2 , thus we can suggest

$$\begin{aligned} \hat{V}_m &= N^{-2} \sum_{k \in s} a_k^2 \hat{V}(y_k) \\ &= N^{-2} \sum_{k \in s} a_k^2 (\hat{y}_{mk} - y_k)^2 \qquad (5.3) \end{aligned}$$

which is consistent of $V_{\mathbf{y}}(\hat{\bar{\mathbf{y}}}_m)$, the $\mathbf{y}$ variance of the estimator. (Valliant et al., 2000, Lemma 5.3.1, p.136) offer several alternative estimators which attempt to estimate the smaller component of the approximate variance, see (Valliant et al., 2000, , p.145) for further details.

5.2 GREG Coefficient Sequential Properties

For any $t = 1, .., T$, the estimator $\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}$ is approximately πm-unbiased of $\boldsymbol{\beta}^{t-1}$ if selection models up to t and model $E(Y_k) = \underline{\mathbf{x}}_k^{t-1'} \boldsymbol{\beta}_{t-1}$ are both correctly specified and consistently estimated.

That is

$$E\left(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}\right) \approx \boldsymbol{\beta}^{t-1}$$

if models $\pi_T(\underline{\mathbf{x}}^{T-1}), ..., \pi_{t+1}(\underline{\mathbf{x}}^t)$ as well as $E(Y_k) = \underline{\mathbf{x}}_k^{t-1'} \boldsymbol{\beta}_{t-1}$ are correctly specified and consistently estimated.

Let

$$\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)} = \left(\sum_{s^0} \underline{S}_k^T \underline{\mathbf{X}}_k^{t-1'} \underline{\mathbf{X}}_k^{t-1} \underline{\hat{\pi}}_{Tk}^{-1} \right)^{-1} \left(\sum_{s^0} \underline{S}_k^T \underline{\mathbf{X}}_k^{t-1} Y_k \underline{\hat{\pi}}_{Tk}^{-1} \right)$$

and note that

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}\right) &= EE\left\{\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}|\underline{\mathbf{x}}^{t-1},\underline{s}^{t-1}\right\} \\ &= EE\left\{E\left[\cdots E\left(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}|y,\underline{\mathbf{x}}^{T-1},\underline{s}^{T-1}\right)\cdots|y,\underline{\mathbf{x}}^{t},\underline{s}^{t}\right]|\underline{\mathbf{x}}^{t-1},\underline{s}^{t-1}\right\}. \end{aligned}$$

First assume that $\pi_T(\underline{\mathbf{x}}^{T-1},\underline{s}^{T-1})$ holds and is estimated consistently so that $\hat{\pi}_{Tk}\underset{n_T\to\infty}{\longrightarrow}\pi_{Tk}$ then

$$E\left(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^T)}|y,\underline{\mathbf{x}}^{T-1},\underline{s}^{T-1}\right)\approx\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{T-1})}$$

where

$$\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{T-1})}=\left(\sum_{s^0}\underline{S}_k^{T-1}\underline{\mathbf{X}}_k^{t-1'}\underline{\mathbf{X}}_k^{t-1}\underline{\hat{\pi}}_{T-1k}^{-1}\right)^{-1}\left(\sum_{s^0}\underline{S}_k^{T-1}\underline{\mathbf{X}}_k^{t-1}Y_k\underline{\hat{\pi}}_{T-1k}^{-1}\right).$$

Similarly if $\pi_{T-1}(\underline{\mathbf{x}}^{T-2},\underline{s}^{T-2})$ holds and $\hat{\pi}_{T-1k}\underset{n_{T-1}\to\infty}{\longrightarrow}\pi_{T-1k}$ then $E\left(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{T-1})}|y,\underline{\mathbf{x}}^{T-2},\underline{s}^{T-2}\right)\approx$ $\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{T-2})}$ where

$$\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{T-2})}=\left(\sum_{s^0}\underline{S}_k^{T-2}\underline{\mathbf{X}}_k^{t-1'}\underline{\mathbf{X}}_k^{t-1}\underline{\hat{\pi}}_{T-2k}^{-1}\right)^{-1}\left(\sum_{s^0}\underline{S}_k^{T-2}\underline{\mathbf{X}}_k^{t-1}Y_k\underline{\hat{\pi}}_{T-2k}^{-1}\right)$$

and continue this sequentially under $\pi_{T-2}(\underline{\mathbf{x}}^{T-3},\underline{s}^{T-3})$.... to $\pi_{t+1}(\underline{\mathbf{x}}^{t},\underline{s}^{t})$ so that we reach

$$\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{t})}=\left(\sum_{s^0}\underline{S}_k^{t}\underline{\mathbf{X}}_k^{t-1'}\underline{\mathbf{X}}_k^{t-1}\underline{\hat{\pi}}_{tk}^{-1}\right)^{-1}\left(\sum_{s^0}\underline{S}_k^{t}\underline{\mathbf{X}}_k^{t-1}Y_k\underline{\hat{\pi}}_{tk}^{-1}\right).$$

Now, if $E(Y_k)=\underline{\mathbf{x}}_k^{t-1'}\boldsymbol{\beta}_{t-1}$ holds then

$$E\left(\hat{\boldsymbol{\beta}}^{t-1}_{\pi(s^{t})}\right)=EE\left\{\left(\sum_{s^0}\underline{S}_k^{t}\underline{\mathbf{X}}_k^{t-1'}\underline{\mathbf{X}}_k^{t-1}\hat{\pi}_{tk}^{-1}\right)^{-1}\left(\sum_{s^0}\underline{S}_k^{t}\underline{\mathbf{X}}_k^{t-1'}Y_k\hat{\pi}_{tk}^{-1}\right)|\underline{\mathbf{x}}^{t-1},\underline{s}^{t}\right\}=\boldsymbol{\beta}^{t-1}$$

as $E(Y_k|\underline{\mathbf{x}}^{t-1},\underline{s}^{t})=\underline{\mathbf{x}}^{t-1'}\boldsymbol{\beta}^{t-1}$.

5.3 Simulation Summary and Distributions

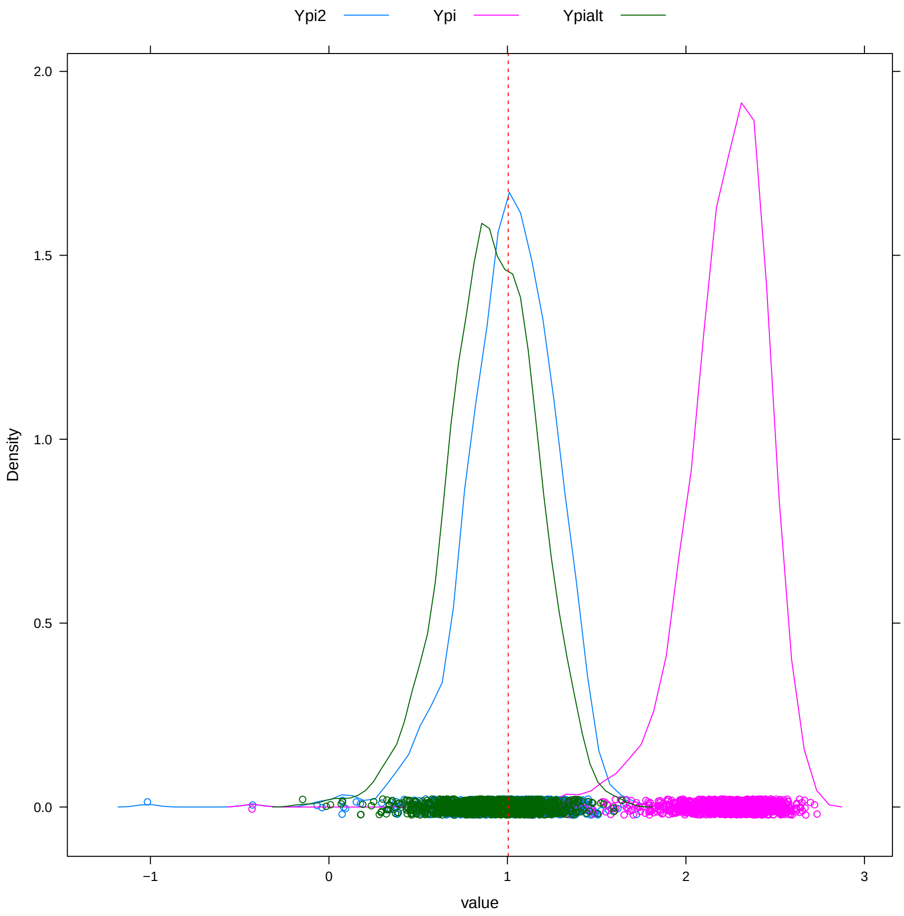


Figure 5.1: π-estimators

$\pi-$Estimators	$\hat{\bar{\mathbf{y}}}_{\underline{\pi 2}}$	$\hat{\bar{\mathbf{y}}}_{\underline{\pi x^0}}$	$\hat{\bar{\mathbf{y}}}_{\underline{\pi \underline{x}^1}}$
	1.01	2.22	0.90
$V(\cdot)$	0.07	0.06	0.06

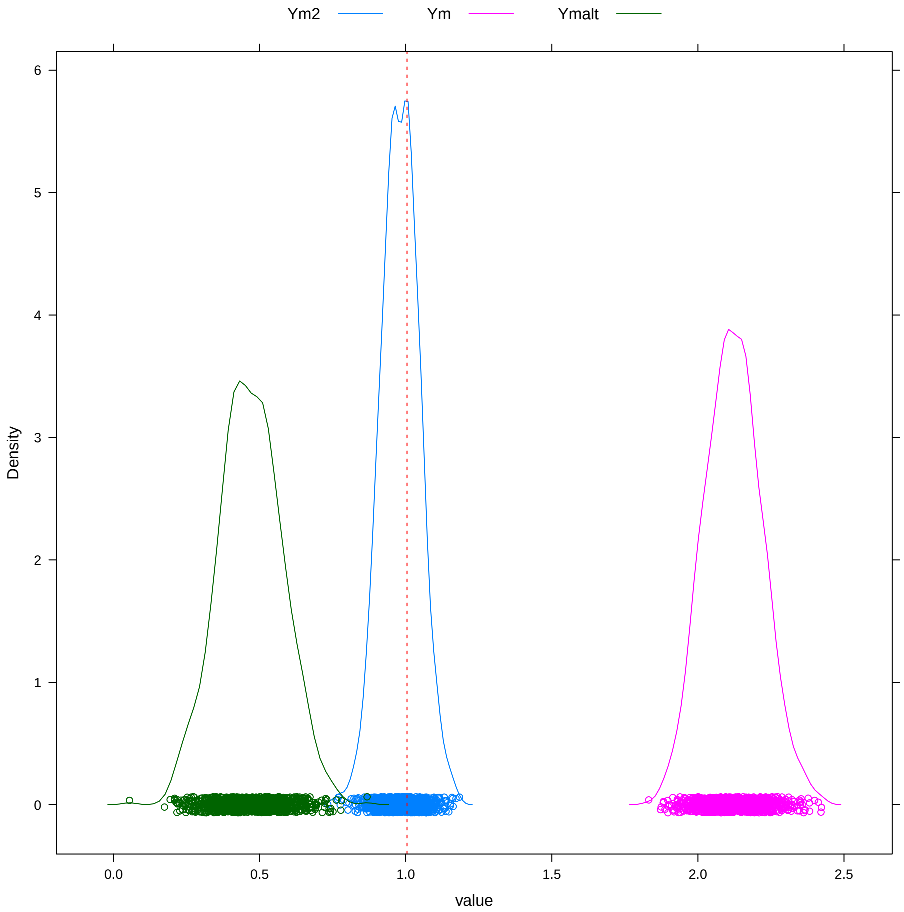


Figure 5.2: m-estimators

$m-$Estimators	$\underline{\hat{\bar{\mathbf{y}}}_{m2}}$	$\underline{\hat{\bar{\mathbf{y}}}_{mx^0}}$	$\underline{\hat{\bar{\mathbf{y}}}_{mx^1}}$
	0.99	2.12	0.47
$V(\cdot)$	0.00	0.01	0.01

5.4 GREG Estimator DR Property

Given the previous distributional assumptions, the GREG estimator $\hat{\bar{\mathbf{Y}}}_{\pi y}$ under an ignorable T phase selection model is unbiased of the population mean $\bar{\mathbf{Y}}_{s^0}$ if

(i) Either the selection model $\pi_1(\mathbf{x}^0)$ or the measurment model $m_1(\mathbf{x}^0)$ holds, as well as that

(ii) selection models $\pi_t(\underline{\mathbf{x}}^{t-1}, \underline{s}^{t-1})$ for $t = 2, ..., T$ are correctly specified and consistently estimated.

For a T phase selection process deconstruct the bias of the estimator $\hat{\bar{\mathbf{Y}}}_{\pi y}$ as

$$\begin{aligned}
B(\hat{\overline{\mathbf{Y}}}_{\pi y}) &= N^{-1}\sum_{s^0} E\left\{\sum_{t=1}^{T}\underline{S}_k^{t-1}(\hat{\pi}_{tk}-S_k^t)(\hat{Y}_{t\pi k}-Y_k)\hat{\underline{\pi}}_{tk}^{-1}\right\}\\
&= E\left\{(\hat{\pi}_{1k}-S_k^1)(\hat{Y}_{1\pi k}-Y_k)\hat{\pi}_{1k}^{-1}+\sum_{t=2}^{T}\underline{S}_k^{t-1}(\hat{\pi}_{tk}-S_k^t)(\hat{Y}_{t\pi k}-Y_k)\hat{\underline{\pi}}_{tk}^{-1}\right\}\\
&= \underline{\overline{I}}+\underline{\overline{II}}
\end{aligned}$$

Consider first quantity $\underline{\overline{II}}$.

As selection models $\pi_t(\underline{\mathbf{x}}^{t-1},\underline{s}^{t-1})$ for $t=2,...,T$ are assumed to be correctly specified and consistently estimated then $\hat{\pi}_t(\cdot)\xrightarrow[n_t\to\infty]{}\pi_t(\cdot)$ and so each $\hat{\mathbf{B}}^{t-1}_{\pi(s^T)}$ is selection consistent and approximately unbiased estimator of the coressponding finite population quantities $\mathbf{B}^{t-1}_{s^{t-1}}$. Thus quantity $\underline{\overline{II}}$:

$$\sum_{t=2}^{T} E\left\{\underline{S}_k^{t-1}(\hat{\pi}_{tk}-S_k^t)(\underline{\mathbf{X}}_k^{t-1\prime}\hat{\mathbf{B}}^{t-1}_{\pi(s^T)}-Y_k)\hat{\underline{\pi}}_{tk}^{-1}\right\}$$

can be approximated by

$$\approx\sum_{t=2}^{T} E\left\{\underline{S}_k^{t-1}(\pi_{tk}-S_k^t)(\underline{\mathbf{X}}_k^{t-1\prime}\mathbf{B}^{t-1}_{s^{t-1}}-Y_k)\hat{\pi}_{1k}^{-1}\cdot\pi_{2k}^{-1}\cdots\pi_{tk}^{-1}\right\}$$

which is equal to

$$\sum_{t=2}^{T} EE\left\{\underline{S}_k^{t-1}(\pi_{tk}-S_k^t)(\underline{\mathbf{X}}_k^{t-1\prime}\mathbf{B}^{t-1}_{s^{t-1}}-Y_k)\hat{\pi}_{1k}^{-1}\cdot\pi_{2k}^{-1}\cdots\pi_{tk}^{-1}|\underline{\mathbf{x}}^{t-1},\underline{s}^{t-1},y_k\right\}\quad(5.4)$$

or equivelently $$\sum_{t=2}^{T} EE\left\{\underline{S}_k^{t-1}(\pi_{tk}-S_k^t)(\underline{\mathbf{X}}_k^{t-1\prime}\mathbf{B}^{t-1}_{s^{t-1}}-Y_k)\hat{\pi}_{1k}^{-1}\cdot\pi_{2k}^{-1}\cdots\pi_{tk}^{-1}|\underline{\mathbf{x}}^{t-1},\underline{s}^{t}\right\}\quad(5.5)$$

(a) If $\pi_1(\cdot)$ is correct and consistently estimated then $\hat{\pi}_{1k}\to\pi_{1k}$ and (5.4) is consistent of

$$\sum_{t=2}^{T} EE\left\{\underline{S}_k^{t-1}(\pi_{tk}-S_k^t)(\underline{\mathbf{X}}_k^{t-1\prime}\mathbf{B}^{t-1}_{s^{t-1}}-Y_k)\underline{\pi}_{tk}^{-1}|\underline{\mathbf{x}}^{t-1},\underline{s}^{t-1},y_k\right\}$$

which is equal to zero as $E(S_k^t|\underline{\mathbf{x}}^{t-1},\underline{s}^{t-1},y)=E(S_k^t|\underline{\mathbf{x}}^{t-1},\underline{s}^{t-1})$ for all $t=2,..,T$.

(b) If $m_1(\cdot)$ is correct and consistently estimated then the identity (5.5) is

$$\sum_{t=2}^{T} EE\left\{\underline{S}_k^{t-1}(\pi_{tk} - S_k^t)(\underline{\mathbf{X}}_k^{t-1'}\mathbf{B}_{s^{t-1}}^{t-1} - Y_k)\hat{\pi}_{1k}^{-1} \cdot \pi_{2k}^{-1} \cdots \pi_{tk}^{-1}|\underline{\mathbf{x}}^{t-1}, \underline{s}^t\right\} = 0$$

as $E(Y_k|\underline{\mathbf{x}}^{t-1}, \underline{s}^t) = \underline{\mathbf{x}}_k^{t-1'}\boldsymbol{\beta}^{t-1}$ and $E(\underline{\mathbf{X}}_k^{t-1'}\mathbf{B}_{s^{t-1}}^{t-1}|\underline{\mathbf{x}}^{t-1}, \underline{s}^t) = \underline{\mathbf{x}}_k^{t-1'}\boldsymbol{\beta}^{t-1}$.

As to quantity $\underline{\overline{I}}$:

$$E\left\{(\hat{\pi}_{1k} - S_k^1)(\hat{Y}_{1\pi k} - Y_k)\hat{\pi}_{1k}^{-1}\right\} = E\left\{(\hat{\pi}_{1k} - S_k^1)(\mathbf{X}_k^{0'}\hat{\mathbf{B}}_{\pi(s^T)}^0 - Y_k)\hat{\pi}_{1k}^{-1}\right\} \quad (5.6)$$

(a) If $\pi_1(\cdot)$ is correct and consistently estimated then $\hat{\pi}_{1k} \to \pi_{1k}$ and so (5.6) can be approximated to

$$E\left\{(\pi_{1k} - S_k^1)(\mathbf{X}_k^{0'}\mathbf{B}^0 - Y_k)\pi_{1k}^{-1}\right\} = EE\left\{(\pi_{1k} - S_k^1)(\mathbf{X}_k^{0'}\mathbf{B}^0 - Y_k)\pi_{1k}^{-1}|\mathbf{x}^0, y\right\}$$

which is equal to zero as $E(S_k^1|\mathbf{x}^0, y) = \pi_{1k}$. On the other hand

(b) If $m_1(\cdot)$ is correct and consistently estimated then

$$E\left\{(\hat{\pi}_{1k} - S_k^1)(\mathbf{X}_k^{0'}\hat{\mathbf{B}}_{\pi(s^T)}^0 - Y_k)\hat{\pi}_{1k}^{-1}\right\} \approx EE\left\{(\hat{\pi}_{1k} - S_k^1)(\mathbf{X}_k^{0'}\mathbf{B}^0 - Y_k)\hat{\pi}_{1k}^{-1}|\mathbf{x}^0, s^1\right\}$$

is equal to zero as $E\left(\mathbf{X}_k^{0'}\mathbf{B}^0 - Y_k|\mathbf{x}^0, s^1\right) = 0$.

5.5 Cube Method

Let $\mathbf{x}_k = (x_{1k}, ..., x_{kp})'$ the observed value of random vector $\mathbf{X}_k$ and let $\mathbf{X} = [\mathbf{X}_1, ..., \mathbf{X}_k, ..., \mathbf{X}_N]'$ be the $N \times p$ population matrix. A sampling design $p(s)$ is said to be balanced on the auxiliary variables $\mathbf{x}$, if and only if it satisfies the balancing equations given by

$$\begin{aligned} \sum_U \mathbf{x}_k &= \sum_s \mathbf{x}_k \pi_k^{-1} \text{ and} \\ V\left\{\sum_s \mathbf{X}_k \pi_k^{-1}|\mathbf{X} = \mathbf{x}\right\} &= 0 \end{aligned}$$

where the second equation (the variance constraint) can be interpreted so that balance in the first equation is satisfied for all $s \in \mathcal{S}$ such that $p(s) > 0$.

The cube method is an algorithm which calculates the unit inclusion probabilities π_k for $k = 1, .., N$ defining a design $p(s)$ balanced on $\mathbf{X}$. The algorithm can be explained concisely by viewing the problem geometrically. A design $p(s)$ defined on $\mathcal{S}$ the set of samples s. Now, each sample s is in fact a vector of 0's and 1's which can be considered as a vertex of the $N-$cube. Also, the balancing equations $\sum_U \mathbf{a}_k s_k = \sum_U \mathbf{a}_k \pi_k$ where $\mathbf{a}_k = \mathbf{x}_k \pi_k^{-1}$ define a linear subspace in R^N of dimension $N - p$. So the problem is to choose a vertex of the $N-$cube (a sample) that remains on the linear sub space. The solution is found by a linear program solved by simplex algorithm (see (Tillé, 2006, sec 8.6)) .

5.6 Ignorability in Non Sequential Case

Denote $P(\underline{S}_{tk} = \underline{1}_t | y_k, \mathbf{x}_k^0) = p(\underline{s}_{tk} | y_k, \mathbf{x}_k^0)$. I show that sequential Ignorability implies independence as defined single phase Ignorability:

$$\begin{aligned} p(\underline{s}_{tk} | y_k, \mathbf{x}_k^0) &= p(s_{tk} | \underline{s}_{t-1k}, y_k, \mathbf{x}_k^0) p(\underline{s}_{t-1k} | y_k, \mathbf{x}_k^0) \\ &\overset{4-II}{=} p(s_{tk} | \underline{s}_{t-1k} \mathbf{x}_k^0) p(\underline{s}_{t-1k} | y_k, \mathbf{x}_k^0) \end{aligned}$$

repeating the factorisation of $p(\underline{s}_{t-1k} | y_k, \mathbf{x}_k^0)$ until period 0, I obtain the desired result

$$p(\underline{s}_{tk} | y_k, \mathbf{x}_k^0) = p(\underline{s}_{tk} | \mathbf{x}_k^0).$$

5.7 Example Estimators- Sequential Selection BLUP Estimators

Note that they all are ML estimators for $m_2(\underline{\mathbf{x}}^1)$.

$m_2(\underline{\mathbf{x}}^1)$	$\underline{\mathbf{X}}^1 = \underline{\mathbf{x}}^1$	$\hat{\overline{\mathbf{y}}}_m = \underline{\overline{\mathbf{x}}}^1 \hat{\boldsymbol{\beta}}$
$[X^{p0} X^{p1}]^Y$	$\underline{\mathbf{X}}^1$	$\sum_{p^0} \sum_{P^1} W_{p^0 p^1} \overline{\mathbf{y}}_{s^2 p^0 p^1}{}^{(*)}$
$[X^{p0} X^{p1}]^Y$	$(\mathbf{X}^{p0}, \mathbf{X}_{s^1}^{p1})$	$\frac{\sum_{p^0} \sum_{P^1} w'_{p^0 p^1} n_{2p^0 p^1} \overline{\mathbf{y}}_{s^2 p^0 p^1}}{\sum_{p^0} \sum_{P^1} w'_{p^0 p^1} n_{2p^0 p^1}} (**)$
$[X^{p0} + X^{p1}]^Y$	$(\mathbf{X}^{p0}, \mathbf{X}_{s^1}^{p1})$	$\frac{\sum_{p^0} \sum_{P^1} w'_{p^0 p^1} n_{2p^0 p^1} (\hat{\beta} + \hat{\beta}_{p^0} + \hat{\beta}_{p^1})}{\sum_{p^0} \sum_{P^1} w'_{p^0 p^1} n_{2p^0 p^1}} (***)$
		add $[X^{p0} + \underline{X}^1]^Y$?
$[X^{p1}]^Y$	$(\mathbf{X}^{p0}, \mathbf{X}_{s^1}^{p1})$	$\frac{\sum_{P^1} w^*_{p^1} \overline{\mathbf{y}}_{s^2 \cdot p^1}}{\sum_{P^1} w^*_{p^1}}$ (****)
$[X^{p0}]^Y$	$(\mathbf{X}^{p0}, \mathbf{X}_{s^1}^{p1})$	$\sum_{p^0} \frac{N_{p^0}}{N} \overline{\mathbf{y}}_{s^2 p^0 .}$
$[\phi]^Y$	$(\mathbf{X}^{p0}, \mathbf{X}_{s^1}^{p1})$	$\sum_{p^0} \frac{n_{2p^0 .}}{n_2} \overline{\mathbf{y}}_{s^2 p^0 .}$

comments-

(*) $W_{p^0p^1} = N_{p^0p^1}/N$

(**) $w'_{p^0p^1} = \frac{N_{p^0.}}{n_{1p^0}}\frac{n_2}{N}/\frac{n_{2p^0p^1}}{n_{1p^0p^1}}$. The estimator reduces to $\hat{\bar{\mathbf{y}}}_m = \sum_{p^0}\sum_{P^1}\frac{N_{p^0.}}{N}\frac{n_{1p^0p^1}}{n_{1p^0}}\bar{\mathbf{y}}_{s^2p^0p^1}$

(***) If the model for Y is $[X^0 + X^1]^Y$, then $\hat{y}_{kp^0p^1} = \hat{\beta} + \hat{\beta}_{p^0} + \hat{\beta}_{p^1}$,predicted values from an additive model fitted to the respondent data on s^2.

(****) $w^*_{p^1} = \sum_{p^0} w'_{p^0p^1} n_{2p^0p^1}$

5.8 Balancing score Properties in the Three Phase Sequential Case

Presuming assumptions (1) to (4), the following statements hold

- $Y_k \perp S_{1k} | b_{1k}(\mathbf{x}^0)$ for all $b_1(\mathbf{x}^0)$ such that $E\{p_1(\mathbf{x}^0)|b_1(\mathbf{x}^0)\} = p_1(\mathbf{x}^0)$

Proof: see earlier case.

---

functions that fulfill the balancing score condition is the propensity score $p_1(\mathbf{x}^0)$. This is identical to the one phase case.

- $Y_k \perp S_{2k} | b_{2k}(\underline{\mathbf{x}}^1, s_1)$ for all $b_2(\underline{\mathbf{x}}^1, s_1)$ such that $E\{p_2(\underline{\mathbf{x}}^1, s_1)|b_2(\underline{\mathbf{x}}^1, s_1)\} = p_1(\underline{\mathbf{x}}^1, s_1)$.

Proof:

$$\begin{aligned}
Pr(S_2 = 1|y, b_2(\underline{\mathbf{x}}^1, s_1)) &= E_{\underline{\mathbf{x}}^1, s_1}\{Pr(S_2 = 1|\underline{\mathbf{x}}^1, s_1, y, b_2(\underline{\mathbf{x}}^1, s_1))|y, b_2(\underline{\mathbf{x}}^1, s_1)\} \\
&\overset{4-I}{=} E_{\underline{\mathbf{x}}^1, s_1}\{p_2(\underline{\mathbf{x}}^1, s_1)|y, b_2(\underline{\mathbf{x}}^1, s_1)\} \\
&\overset{b.d}{=} Pr(S_2 = 1|\mathbf{x}^0, s_1).
\end{aligned}$$

---

functions that fulfil the balancing score condition is $p_2(\underline{\mathbf{x}}^1, s_1)$, $\underline{p}_2(\underline{\mathbf{x}}^1)$ as well as $[p_2(\underline{\mathbf{x}}^1, s_1), s_1]$ and $[\underline{p}_2(\underline{\mathbf{x}}^1), s_1]$.

- $Y_k \perp S_{3k} | b_{3k}(\underline{\mathbf{x}}^2, \underline{s}_2)$ for all $b_{3k}(\underline{\mathbf{x}}^2, \underline{s}_2)$ such that $E\{p_3(\underline{\mathbf{x}}^2, \underline{s}_2) | b_{3k}(\underline{\mathbf{x}}^2, \underline{s}_2)\} = p_3(\underline{\mathbf{x}}^2, \underline{s}_2)$.

Proof:

$$\begin{aligned} Pr(S_3 = 1 | y, b_{3k}(\underline{\mathbf{x}}^2, \underline{s}_2)) &= E_{\underline{\mathbf{x}}^2, \underline{s}_2}\{Pr(S_3 = 1 | \underline{\mathbf{x}}^2, \underline{s}_2, y, b_{3k}(\underline{\mathbf{x}}^2, \underline{s}_2)) | y, b_{3k}(\underline{\mathbf{x}}^2, \underline{s}_2)\} \\ &\overset{4-I}{=} E_{\underline{\mathbf{x}}^2, \underline{s}_2}\{p_3(\underline{\mathbf{x}}^2, \underline{s}_2) | y, b_{3k}(\underline{\mathbf{x}}^2, \underline{s}_2)\} \\ &\overset{b.d}{=} Pr(S_3 = 1 | \underline{\mathbf{x}}^2, \underline{s}_2). \end{aligned}$$

---

functions that fulfill the balancing score condition is $p_3(\underline{\mathbf{x}}^2, \underline{s}_2)$, $\underline{p}_3(\underline{\mathbf{x}}^2)$ as well as $[p_3(\underline{\mathbf{x}}^2, \underline{s}_2), \underline{s}_2]$ and $[\underline{p}_3(\underline{\mathbf{x}}^2), \underline{s}_2]$.

5.9 The Three Phased Sequential $\pi-$Estimator is Unbiased

I show that the three phased $\pi-$estimator

$$\hat{\bar{Y}}_\pi = \sum_{k \in s_0} Y_k \underline{S}_{3k} \underline{\hat{\pi}}_3(\underline{\mathbf{x}}_k^2)^{-1} / \sum_{k \in s_0} \underline{S}_{3k} \underline{\hat{\pi}}_3(\underline{\mathbf{x}}_k^2)^{-1}$$

indeed estimates the desired quanity- The population average of measurement $E(Y)$.

As in previous derivations, I assume for simplicity that the probabilities π are estimated consistently and with enough regularity such the the following exposition based on true probabilities holds asymptotically with estimated probabilities as well.

I also replace the scale factor $\sum_{k \in s_0} \underline{S}_{3k} \underline{\hat{\pi}}_3(\underline{\mathbf{x}}_k^2)^{-1}$ by its estimand, the population total N. Note that

$$\begin{aligned} \hat{\bar{Y}}_\pi &= \sum_{k \in s_0} Y_k \underline{S}_{3k} \underline{\hat{\pi}}_3(\underline{\mathbf{x}}_k^2)^{-1} / \sum_{k \in s_0} \underline{S}_{3k} \underline{\hat{\pi}}_3(\underline{\mathbf{x}}_k^2)^{-1} \\ &= N^{-1} \sum_{k \in s_0} Y_k \underline{S}_{3k} \underline{\hat{\pi}}_3(\underline{\mathbf{x}}_k^2)^{-1} + O(n^{-1}). \end{aligned}$$

Thus

$$\begin{aligned}
N \cdot E(\hat{\bar{Y}}_{\pi}) &= \sum_{k \in s_0} E \frac{Y_k \underline{S}_{3k}}{\underline{\pi}_3(\underline{\mathbf{x}}_k^2)} \\
&= \sum_{k \in s_0} E \frac{Y_k S_{1k} S_{2k} S_{3k}}{\pi_1(\mathbf{x}^0; \boldsymbol{\alpha}_1) \pi_2(\underline{\mathbf{x}}^1, s_1; \boldsymbol{\alpha}_2) \pi_3(\underline{\mathbf{x}}^2, \underline{s}_2; \boldsymbol{\alpha}_3)} \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0} \left[E(\frac{Y_k S_{1k} S_{2k} S_{3k}}{\pi_2(\underline{\mathbf{x}}^1, s_1; \boldsymbol{\alpha}_2) \pi_3(\underline{\mathbf{x}}^2, \underline{s}_2; \boldsymbol{\alpha}_3)} | \mathbf{x}^0) \cdot \frac{1}{\pi_1(\mathbf{x}^0; \boldsymbol{\alpha}_1)} \right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0} \left[E(\frac{Y_k S_{2k} S_{3k}}{\pi_2(\underline{\mathbf{x}}^1, s_1; \boldsymbol{\alpha}_2) \pi_3(\underline{\mathbf{x}}^2, \underline{s}_2; \boldsymbol{\alpha}_3)} | \mathbf{x}^0, s_{1k}) p(s_{1k} = 1 | \mathbf{x}^0) \cdot \frac{1}{\pi_1(\mathbf{x}^0; \boldsymbol{\alpha}_1)} \right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0} \left[E(\frac{Y_k S_{2k} S_{3k}}{\pi_2(\underline{\mathbf{x}}^1, s_1; \boldsymbol{\alpha}_2) \pi_3(\underline{\mathbf{x}}^2, \underline{s}_2; \boldsymbol{\alpha}_3)} | \mathbf{x}^0, s_{1k}) \right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0} \left[E_{\mathbf{x}^1} \left\{ E(\frac{Y_k S_{2k} S_{3k}}{\pi_2(\underline{\mathbf{x}}^1, s_1; \boldsymbol{\alpha}_2) \pi_3(\underline{\mathbf{x}}^2, \underline{s}_2; \boldsymbol{\alpha}_3)} | \underline{\mathbf{x}}^1, s_{1k}) \right\} | \mathbf{x}^0, s_{1k} \right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0} \left[E_{\mathbf{x}^1} \left\{ E(\frac{Y_k S_{2k} S_{3k}}{\pi_3(\underline{\mathbf{x}}^2, \underline{s}_2; \boldsymbol{\alpha}_3)} | \underline{\mathbf{x}}^1, s_{1k}) \frac{1}{\pi_2(\underline{\mathbf{x}}^1, s_1; \boldsymbol{\alpha}_2)} \right\} | \mathbf{x}^0, s_{1k} \right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0} \left[E_{\mathbf{x}^1} \left\{ E(\frac{Y_k S_{3k}}{\pi_3(\underline{\mathbf{x}}^2, \underline{s}_2; \boldsymbol{\alpha}_3)} | \underline{\mathbf{x}}^1, \underline{s}_{2k}) p(s_{2k} = 1 | \underline{\mathbf{x}}^1, s_{1k}) \frac{1}{\pi_2(\underline{\mathbf{x}}^1, s_1; \boldsymbol{\alpha}_2)} \right\} | \mathbf{x}^0, s_{1k} \right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0} \left[E_{\mathbf{x}^1} \left\{ E(\frac{Y_k S_{3k}}{\pi_3(\underline{\mathbf{x}}^2, \underline{s}_2; \boldsymbol{\alpha}_3)} | \underline{\mathbf{x}}^1, \underline{s}_{2k}) \right\} | \mathbf{x}^0, s_{1k} \right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0} \left[E_{\mathbf{x}^1} \left\{ E_{\mathbf{x}^2} \left(E(\frac{Y_k S_{3k}}{\pi_3(\underline{\mathbf{x}}^2, \underline{s}_2; \boldsymbol{\alpha}_3)} | \underline{\mathbf{x}}^2, \underline{s}_{2k}) \right) | \underline{\mathbf{x}}^1, \underline{s}_{2k} \right\} | \mathbf{x}^0, s_{1k} \right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0} \left[E_{\mathbf{x}^1} \left\{ E_{\mathbf{x}^2} \left(E(Y_k | \underline{\mathbf{x}}^2, \underline{s}_{3k}) \frac{p(s_{3k} = 1 | \underline{\mathbf{x}}^2, \underline{s}_{2k})}{\pi_3(\underline{\mathbf{x}}^2, \underline{s}_2; \boldsymbol{\alpha}_3)} \right) | \underline{\mathbf{x}}^1, \underline{s}_{2k} \right\} | \mathbf{x}^0, s_{1k} \right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0} \left[E_{\mathbf{x}^1} \left\{ E_{\mathbf{x}^2} \left(E(Y_k | \underline{\mathbf{x}}^2, \underline{s}_{3k}) \right) | \underline{\mathbf{x}}^1, \underline{s}_{2k} \right\} | \mathbf{x}^0, s_{1k} \right]
\end{aligned}$$

Now by sequentially for each phase, using the ignorability assumption, followed by averaging over the relevant auxiliary vector

$$\begin{aligned}
N \cdot E(\hat{\overline{Y}}_{\pi}) &= \sum_{k \in s_0} E_{\mathbf{x}^0}\left[E_{\mathbf{x}^1}\left\{E_{\mathbf{x}^2}\left(E(Y_k|\underline{\mathbf{x}}^2, \underline{s}_{3k})\right)|\underline{\mathbf{x}}^1, \underline{s}_{2k}\right\}|\mathbf{x}^0, s_{1k}\right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0}\left[E_{\mathbf{x}^1}\left\{E_{\mathbf{x}^2}\left(E(Y_k|\underline{\mathbf{x}}^2, \underline{s}_{2k})\right)|\underline{\mathbf{x}}^1, \underline{s}_{2k}\right\}|\mathbf{x}^0, s_{1k}\right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0}\left[E_{\mathbf{x}^1}\left\{E(Y_k|\underline{\mathbf{x}}^1, \underline{s}_{2k})\right\}|\mathbf{x}^0, s_{1k}\right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0}\left[E_{\mathbf{x}^1}\left\{E(Y_k|\underline{\mathbf{x}}^1, \underline{s}_{1k})\right\}|\mathbf{x}^0, s_{1k}\right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0}\left[E(Y_k|\mathbf{x}^0, s_{1k}\right] \\
&= \sum_{k \in s_0} E_{\mathbf{x}^0}\left[E(Y_k|\mathbf{x}^0\right] \\
&= \sum_{k \in s_0} E(Y_k) \\
&= N \cdot E(Y)
\end{aligned}$$

5.10 Example Estimators- Sequential Selection $\pi-$Estimators

As in the one phase case, note that many of estimators coincide with the outcome model ML estimators.

$\pi_2(\underline{\mathbf{x}}^1)$	$\underline{\mathbf{X}}^1 = \underline{\mathbf{x}}^1$	$\hat{\bar{\mathbf{y}}}_\pi = \sum_{k\in s^2} y_k \underline{\hat{\pi}}_2(\underline{\mathbf{x}}^1_k)^{-1} / \sum_{k\in s^2} \underline{\hat{\pi}}_2(\underline{\mathbf{x}}^1_k)^{-1}$
$[\underline{X}^1]^S$	$\underline{\mathbf{X}}^1$	$\sum_{s^2} y_k \hat{\pi}_{2k}^{-1} / \sum_{s^2} \hat{\pi}_{2k}^{-1*}$
$[X^{p0}]^{S1}$,$[\underline{X}^{p1}]^{S2\mid S1}$	$\underline{\mathbf{X}}^1$	$\sum_{p^0} \sum_{P^1} W_{p^0p^1} \bar{\mathbf{y}}_{s^2p^0p^1}$ **
$[X^0]^{S1}$,$[\underline{X}^1]^{S2\mid S1}$	$(\mathbf{X}^0, \mathbf{X}^1_{s^1})$	$\sum_{k\in s^2} y_k \underline{\hat{\pi}}_{2k}{}^{-1} / \sum_{k\in s^2} \underline{\hat{\pi}}_{2k}{}^{-1}$
$[X^{p0}]^{S1}$,$[\underline{X}^{p1}]^{S2\mid S1}$	$(\mathbf{X}^0, \mathbf{X}^1_{s^1})$	$\sum_{p^0} \sum_{P^1} \frac{N_{p^0\cdot}}{N} \frac{n_{1p^0p^1}}{n_{1p^0}} \bar{\mathbf{y}}_{s^2p^0p^1}$ ***
$[X^{p0}]^{S1}$,$[\underline{X}^{p1}, Z^1]^{S2\mid S1}$	$(\mathbf{X}^0, \mathbf{X}^1_{s^1})$	$\sum_{p^0} \hat{\pi}_{p^0}^{-1} \sum_{k\in s^2} \hat{\pi}_{2k}^{-1} y_k / \sum_{p^0} \hat{\pi}_{p^0}^{-1} \sum_{k\in s^2} \hat{\pi}_{2k}^{-1}$ ^{v}
$[X^{p0}]^{S1}$,$[\underline{X}^{p1}, Z^1]^{S2\mid S1}$	$(\mathbf{X}^0, \mathbf{X}^1_{s^1})$	$\sum_{p^0} \hat{\pi}_{p^0}^{-1} \sum_{k\in s^2} \hat{\pi}_{2k}^{-1(w)} y_k / \sum_{p^0} \hat{\pi}_{p^0}^{-1} \sum_{k\in s^2} \hat{\pi}_{2k}^{-1(w)}$ ^{vv}
$[X^{p0}]^{S1}$,$[X^{p1}]^{S2\mid S1}$	$(\mathbf{X}^0, \mathbf{X}^1_{s^1})$	$\sum_{p^0} \sum_{p^1} \underline{\hat{\pi}}_{p^0p^1}^{-1} n_{2p^0p^1} \bar{\mathbf{y}}_{s^2p^0p^1} / \sum_{p^0} \sum_{p^1} \underline{\hat{\pi}}_{p^0p^1}^{-1} n_{2p^0p^1}{}^{vvv}$
$[X^{p0}]^{S1}$,$[X^{p1}]^{S2\mid S1}$	$(\mathbf{X}^0, \mathbf{X}^1_{s^1})$	$\sum_{p^0} \sum_{p^1} \underline{\hat{\pi}}_{p^0p^1}^{-1(w)} n_{2p^0p^1} \bar{\mathbf{y}}_{s^2p^0p^1} / \sum_{p^0} \sum_{p^1} \underline{\hat{\pi}}_{p^0p^1}^{-1(w)} n_{2p^0p^1}{}^{vvvv}$
$[X^{p0}]^{S1}$,$[Z^1]^{S2\mid S1}$	$(\mathbf{X}^0, \mathbf{X}^1_{s^1})$	$\sum_{p^0} \sum_{p^1} \underline{\hat{\pi}}_{p^0p^1}^{-1} n_{2p^0p^1} \bar{\mathbf{y}}_{s^2p^0p^1} / \sum_{p^0} \sum_{p^1} \underline{\hat{\pi}}_{p^0p^1}^{-1} n_{2p^0p^1}{}^{*}$
$[X^{p0}]^{S1}$,$[Z^1]^{S2\mid S1}$	$(\mathbf{X}^0, \mathbf{X}^1_{s^1})$	$\sum_{p^0} \sum_{p^1} \underline{\hat{\pi}}_{p^0p^1}^{-1(w)} n_{2p^0p^1} \bar{\mathbf{y}}_{s^2p^0p^1} / \sum_{p^0} \sum_{p^1} \underline{\hat{\pi}}_{p^0p^1}^{-1(w)} n_{2p^0p^1}{}^{**}$
$[Z^0]^{S1}$,$[X^{p1}]^{S2\mid S1}$	$(\mathbf{X}^0, \mathbf{X}^1_{s^1})$	$\sum_{p^1} (\frac{n_{2\cdot p^1}}{n_{1\cdot p^1}})^{-1} \sum_{k\in s^2_{p1}} \hat{\pi}_{1k} y_k / \sum_{p^1} (\frac{n_{2\cdot p^1}}{n_{1\cdot p^1}})^{-1} \sum_{k\in s^2_{p1}} \hat{\pi}_{1k}{}^{***}$
$[Z^0]^{S1}$,$[X^{p1}]^{S2\mid S1}$	$(\mathbf{X}^0, \mathbf{X}^1_{s^1})$	$\frac{\sum_{p1} \hat{N}_{p^1} \{\sum_{k\in s^2_{p1}} \hat{\pi}_{1k}^{-1} y_k / \sum_{k\in s^2_{p1}} \hat{\pi}_{1k}^{-1}\}}{\sum_{p1} \hat{N}_{p^1}}{}^{****}$

* a one phase estimator using fitted probabilities from unconditional model $\hat{\pi}_{2k} = \pi_2(\underline{\mathbf{x}}^1; \hat{\boldsymbol{\alpha}})$.
** Specifica example of * where we assume $\underline{X}^{p1} = (X^{p0}, X^{p1})$ is known for all resulting in a simple one phase post stratification estimator. Underlying assumption is equal selection probability within groups defined by $\underline{\mathbf{x}}^1$.
*** Identical to the MLE with partial information in model based approach. Found by letting $\pi_1(\mathbf{x}^0; \hat{\boldsymbol{\alpha}}_1) = n_{1p^0\cdot}/N_{p^0\cdot}$ and $\pi_2(\underline{\mathbf{x}}^1, s^1; \hat{\boldsymbol{\alpha}}_2) = n_{2p^0p^1}/n_{1p^0p^1}$ into $\pi-$ estimator.

^{v} where $\hat{\pi}_{p^0} = (n_{1p^0\cdot}/N_{p^0\cdot})$ and $\hat{\pi}_{2k} = \pi_2(\underline{x}^{p1}, z^1, s^1; \hat{\boldsymbol{\alpha}}_2)$. The case where first phase is modelled by categorical while 2nd includes continuous variables as well. Is unbiased as follows the assumptions.

^{vv} Same data available and model assumptions as previous but where π_{2k} are fitted by weighted model, that is $\hat{\pi}_{2k}^{-1(w)} = \pi_2(\underline{x}^{p1}, z^1\hat{\pi}_{p^0}, s^1; \hat{\boldsymbol{\alpha}}_2^{(w)})$, for example weighted logistic regression. As ^{v} is unbiased the additional weighting should be redundant.

^{vvv} The Double Expansion Estimator (DEE). Assumes $\underline{\pi}_2(\underline{\mathbf{x}}^1) = [\pi_2(\mathbf{x}^1, s^1; \boldsymbol{\alpha}_2), \pi_1(\mathbf{x}^0; \boldsymbol{\alpha}_1)]$. This is a special case as the 2^{nd} phase model ignores 1^{st} phase information. It estimates by raw response rates so that $\underline{\hat{\pi}}_{p^0p^1}^{-1} = \hat{\pi}_{p^0} \cdot \hat{\pi}_{p^1}$ where $\hat{\pi}_{p^0} = n_{1p^0\cdot}/N_{p^o}$ and $\hat{\pi}_{p^1} = n_{2\cdot p^1}/n_{1\cdot p^1}$. This is generally a biased estimator as $\hat{\pi}_{p^1} = n_{2\cdot p^1}/n_{1\cdot p^1}$ are not unbiased estimators of the population strata size defined by $\mathbf{X}^{p1} = \mathbf{x}^{p1}$**.** The DEE will be unbiased only when $Y \sim X^1$ *and* $\pi_2 \sim X^{p1}$.

^{vvvv} The Reweighted Expansion Estimator (REE). The same as (DEE), but estimates 2^{nd} phase probabilities by weighted response rates, that is $\underline{\hat{\pi}}_{p^0p^1}^{-1(w)} =$

$\hat{\pi}_{p^0} \cdot \hat{\pi}_{p^1}^{(w)}$ where $\hat{\pi}_{p^1} = n_{2 \cdot p^1}/n_{1 \cdot p^1}$ as before while $\hat{\pi}_{p^1}^{(w)} = \sum_{k \in s^2_{p0.}} \hat{\pi}_{p^0}^{-1} / \sum_{k \in s^1_{p0.}} \hat{\pi}_{p^0}^{-1}$. The (REE) estimator will be unbiased if either Y depends only on X^{p1} or S^2 depends only on X^{p1}. This can be shown by simple conditional expectation $E(\hat{\bar{\mathbf{y}}}_\pi) = E_y E_s(\hat{\bar{\mathbf{y}}}_\pi - \bar{\mathbf{y}}|\mathbf{y}) = E_s E_y(\hat{\bar{\mathbf{y}}}_\pi - \bar{\mathbf{y}}|s^2)$.

* Another case where the 2^{nd} model ignores the previous selection model predictors. Specifically, $\underline{\hat{\pi}}_{p^0 p^1}^{-1} = \hat{\pi}_{p^0} \cdot \hat{\pi}_{p^1}$ where $\hat{\pi}_{p^0} = n_{1p^0.}/N_{p^o}$ and $\hat{\pi}_{p^1} = \pi_2(z^1, s^1; \hat{\boldsymbol{\alpha}}_2)$

**The same as above, but here the propensity score model is weighted by $\hat{\pi}_{p^0} = n_{1p^0.}/N_{p^o}$ so that $\hat{\pi}_{p^1}^{(w)} = \pi_2(z^1, s^1, \hat{\pi}_{p^0}; \hat{\boldsymbol{\alpha}}_2^{(w)})$. The interesting question is whether the weighting will help in reducing bias if Y is influenced by X^{p0}.

*** The classic case of (DEE) and (REE) under the setting of KOTT (2011) . The probability π_{1k} is estimated by $\hat{\pi}_{1k} = \pi_1(\mathbf{z}^0; \hat{\boldsymbol{\alpha}}_1)$ while $\hat{\pi}_{2k} = (n_{2 \cdot p1}/n_{1 \cdot p1})$. This results in a normally biased (DEE) estimator.

**** This is the setting KOTT (2011) discusses. First $\hat{\pi}_{1k} = \pi_1(\mathbf{z}^0; \hat{\boldsymbol{\alpha}}_1)$ as above while $\hat{\pi}_{2k}^{(w)} = \pi_2(\mathbf{x}^{p1}, s^1, \hat{\pi}_{1k}; \hat{\boldsymbol{\alpha}}_2) = \frac{\sum_{k \in s^2_{p1}} \hat{\pi}_{1k}^{-1}}{\sum_{k \in s^1_{p1}} \hat{\pi}_{1k}^{-1}}$. Also note that denominator $\sum_{k \in s^1_{p1}} \hat{\pi}_{1k}^{-1} = \hat{N}_{p^1}$ an unbiased estimator of the count in strata defined by $X^{p1} = x^{p1}$. Thus $\hat{\bar{\mathbf{y}}}_\pi = \frac{\sum_{p1} \{\hat{\pi}_{2k}^{(w)-1} \sum_{k \in s^2_{p1}} \hat{\pi}_{1k}^{-1} y_k\}}{\sum_{p1} \{\hat{\pi}_{2k}^{(w)} \sum_{k \in s^2_{p1}} \hat{\pi}_{1k}^{-1}\}}$ and as the denominator $\sum_{p1} \{\hat{\pi}_{2k}^{(w)} \sum_{k \in s^2_{p1}} \hat{\pi}_{1k}^{-1}\} = \sum_{p1} \frac{\sum_{k \in s^2_{p1}} \hat{\pi}_{1k}^{-1}}{\sum_{k \in s^1_{p1}} \hat{\pi}_{1k}^{-1}} \sum_{k \in s^2_{p1}} \hat{\pi}_{1k}^{-1} = \sum_{p1} \hat{N}_{p^1}$ and so $\hat{\bar{\mathbf{y}}}_\pi = \frac{\sum_{p1} \{\hat{\pi}_{2k}^{(w)-1} \sum_{k \in s^2_{p1}} \hat{\pi}_{1k}^{-1} y_k\}}{\sum_{p1} \hat{N}_{p1}} = \frac{\sum_{p1} \hat{N}_{p1} \{\sum_{k \in s^2_{p1}} \hat{\pi}_{1k}^{-1} y_k / \sum_{k \in s^2_{p1}} \hat{\pi}_{1k}^{-1}\}}{\sum_{p1} \hat{N}_{p1}}$.

Bibliography

Abadie, A. and J. Gardeazabal (2008). Terrorism and the world economy. *European Economic Review 52*(1), 1–27.

Abadie, A. and G. W. Imbens (2009). Matching on the estimated propensity score. Technical report, National Bureau of Economic Research.

Augurzky, B. and C. M. Schmidt (2001). The propensity score: A means to an end.

Baker, R., S. Blumberg, J. Brick, M. Couper, M. Courtright, J. Dennis, D. Dillman, M. Frankel, P. Garland, R. Groves, et al. (2010). Research synthesis aapor report on online panels. *Public Opinion Quarterly 74*(4), 711–781.

Bang, H. and J. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Barnes, W., G. Bright, and C. Hewat (2008). Making sense of labour force survey response rates. *Economic and Labour Market Review 2*(12), 32–41.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part i. In *Foundations of statistical inference: proceedings of the Symposium on the Foundations of Statistical Inference prepared under the auspices of the René Descartes Foundation and held at the Department of Statistics, University of Waterloo, Ont., Canada, from March 31 to April 9, 1970*, pp. 203. Holt McDougal.

Bethlehem, J. (2008). How accurate are self selection web surveys. Discussion Paper 08014, Statistics Netherland.

Betts, P. (2010). The application of alternative modes of data collection on uk government social surveys. Technical report, Office for National Statistics.

Binder, D., M. Kovacevic, and G. Roberts (2004). Design-based methods for survey data: Alternative uses of estimating functions. In *Proceedings of the Section on Survey Research Methods*, pp. 3301–3312.

Boruch, R. and G. Terhanian (1996). So what? the implications of new analytic methods for designing nces surveys. In G. Hoachlander, J. E. Griffith, and J. H. Ralph (Eds.), *From Data to Information: New Directions for the*

National Center for Education Statistics. U.S. Department of Education, National Center for Education Statistics.

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer (2006). Variable selection for propensity score models. *American journal of epidemiology 163*(12), 1149–1156.

Carpenter, J., M. Kenward, and S. Vansteelandt (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 169*(3), 571–584.

Cassel, C., C. Särndal, and J. Wretman (1977). Foundations of inference in survey sampling.

Cassel, C. M., C.-E. Särndal, and J. Wretman (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika 63*, 615–620.

Chang, L. and J. A. Krosnick (2009). National surveys via rdd telephone interviewing versus the internet comparing sample representativeness and response quality. *Public Opinion Quarterly 73*(4), 641–678.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science 22*(4), 341–352.

Clarke, K. A. (2009). Return of the phantom menace omitted variable bias in political research. *Conflict Management and Peace Science 26*(1), 46–66.

Clarke, K. A., B. Kenkel, and M. R. Rueda (2011). Misspecification and the propensity score: the possibility of overadjustment. *Unpublished, July 12.*

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics 24*, 295–313.

Couper, M. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly* (64), 464–494.

Couper, M. P. (2008). *Designing effective web surveys*, Volume 75. Cambridge University Press New York.

Couper, M. P., M. W. Traugott, and M. J. Lamias (2001). Web survey design and administration. *Public opinion quarterly 65*(2), 230–253.

D'Agostino Jr, R. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine 17*(19), 2265–2281.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of The Royal Statistical Society B 41*(1), 1–31.

Dawid, A. P. and J. M. Dickey (1977). Likelihood and bayesian inference from selectively reported data. *Journal of the American Statistical Association 72*, 845–850.

Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statzstical Association 94*, 1053–1062.

Dehejia, R. H. and S. Wahba (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics 84(1)*, 151–161.

Deming, W. A. and F. Stephan (1941). On interpretation of censuses as samples. *Journal American Statistical Association 36*, 45–49.

Dever, J. A., A. Rafferty, and R. Richard Valliant (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods 2*(2), 47–62.

Deville, J. and C. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*(418), 376–382.

Deville, J. and Y. Tillé (2004). Efficient balanced sampling: the cube method. *Biometrika 91*(4), 893.

Dorfman, A. (1991). Sound confidence intervals in the heteroscedastic linear model through releveraging. *Journal of the Royal Statistical Society. Series B 53*(2), 441–452.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics 7*, 1–26.

Engle, R., D. Hendry, and J. Richard (1983). Exogeneity. *Econometrica: Journal of the Econometric Society 51*(2), 277–304.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian data analysis.* Boca Raton, FL: Chapman and Hall/CRC.

Groves, R., R. Cialdini, and M. Couper (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly 56*(4), 475–495.

Groves, R. M. and M. P. Couper (1998). *Nonresponse in Household Interview Surveys.* New York: Wiley and Sons.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey Methodology.* New York: Wiley and Sons.

Hade, E. M. and B. Lu (2014). Bias associated with using the estimated propensity score as a regression covariate. *Statistics in medicine 33*(1), 74–87.

Heckman, J. J., H. Ichimura, J. A. Smith, and P. E. Todd. (1998). Characterizing selection bias using experimental data. *Econometrica 66*, 1017–1098.

Ho, D., K. Imai, G. King, and E. Stuart (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis 15*(3), 199–236.

Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–685.

Imbens, G. W. and D. Rubin (2010, December). Causal inference. Unpublished.

Isaksson, A., S. Danielsson, and G. Forsman (2004). On the variability of estimates based on propensity score weighted data from web panels. In *2004 Proceedings of the American Statistical Association*, pp. 3689–3696.

Isaksson, A. and S. Lee (2005). Simple approaches to estimating the variance of the propensity score weighted estimator applied on volunteer panel web survey data a comparative study. In *ASA Section on Survey Research Methods*. American Statistical Association.

Kang, J. and J. Schafer (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Kellner, P. (2008, August). Down with random samples. *Research World 31*.

Kish, L. (1995). *Survey Sampling*. New York, NY: Wiley.

Kosuke, I., G. King, and E. A. Stuart (2006). Misunderstandings among experimentalists and observationalists: Balance test fallacies in causal inference. Technical report, Harvard University.

Kott, P. and M. Fetter (1999). Using multi-phase sampling to limit respondent burden across agriculture surveys. In *Proceedings of the Survey Methods Section, Statistical Society of Canada*. Citeseer.

Kridel, D., P. Rappoport, and L. Taylor (2002). The demand for high-speed access to the internet. *Forecasting the Internet: Understanding the Explosive Growth of Data Communications 39*, 11–22.

Krosnick, J. A. and L. Chang (2004). National surveys via rdd telephone interviewing vs. the internet: Comparing sample representativeness and response quality. *Center for Survey Research, Ohio State University*.

Lechner, M. (2008). A note on endogenous control variables in causal studies. *Statistics & Probability Letters 78*(2), 190–195.

Lechner, M. (2009). Sequential potential outcome models to analyze the effects of fertility on labor market outcomes. *Causal Analysis in Population Studies 23*, 31–57.

Lechner, M. and R. Miquel (2005). *Identification of the effects of dynamic treatments by sequential conditional independence assumptions.* Department of Economics, University of St. Gallen.

Lechner, M. and R. Miquel (2010). Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics 39*(1), 111–137.

Lee, S. (2004). *Statistical Estimation Methods in Volunteer Panel Web Surveys.* Ph. D. thesis, Joint Program in Survey Methodology, University of Maryland.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics 22*(2), 329–349.

Lee, S. and R. Valliant (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods Research 37*(3), 319–343.

Little, R. (1983). Superpopulation models for nonresporse. *Incomplete data in sample surveys 2*, 341–382.

Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data* (Second ed.). Hoboken, NJ:: John Wiley & Sons.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review 54*, 139–157.

Little, R. J. A. and D. B. Rubin (1987). *Statistical analysis with missing data.* New York: John Wiley.

Lohr, S. (2009). *Sampling: design and analysis.* Thomson.

Lumley, T., P. A. Shaw, and J. Y. Dai (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review 79*(2), 200–220.

Lunceford, J. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine 23*(19), 2937–2960.

Lynn, P., D. Elliot, and G. Britain (1999). *The British Crime Survey: A Review of Methodology.* National Centre for Social Research.

Mach, L., J. Dumais, and A. Robinson (2005). Study of the properties of a bootstrap variance estimator under sampling without replacement. In *Federal Committee on Statistical Methodology (FCSM) Research Conference, Arlington (Va).* Citeseer.

Moodie, E. (2009). A note on the variance of doubly-robust g-estimators. *Biometrika 96*(4), 998–1004.

Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly 70*(5), 737–758.

Pearl, J. (2000). *Causality: models, reasoning and inference.* Cambridge Univ Press.

Pearl, J. (2012). A solution to a class of selection-bias problems. Technical report, Tech. Rep.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review 61*(2), 317–337.

Pickering, K. (2008). British crime survey: options for extending the coverage to children and people living in communal establishments. Technical report, Home Office.

Ramasubramanian, V., R. Singh, and A. Rai (2002). Resampling-based variance estimation under two-phase sampling. *Jour. Ind. Soc. Ag. Statistics 55*(2), 197–208.

Ridgeway, G. and D. McCaffrey (2007). Comment-Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science 22*(4), 540–543.

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings, Salt Lake City, UT.* Stanford University and Polimetrix, Inc.

Rivers, D. (2010). Aapor report on online panels. Pacific Chapter of American Association for Public Opinion Research. Accessed from `http://http://www.papor.org/files/2010/Presentations/rivers.pdf`.

Rivers, D. and D. Bailey (2009). Inference from matched samples in the 2008 u.s. national elections. In *American Association of Public Opinion Research-JSM 2009.*

Robertson, A., D. Soopramanien, and R. Fildes (2007). A segment-based analysis of internet service adoption among uk households. *Technology in Society 29*(3), 339–350.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling 7*(9-12), 1393–1512.

Robins, J. (1987). Addendum to a new approach to causal inference in mortality studies with a sustained exposure period. *Comput Math Appl 14*(9-12), 923–945.

Robins, J. (1999). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, Volume 2000, pp. 6–10.

Robins, J. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*, pp. 189–326. Springer Publishing Co New York, NY.

Robins, J. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*(429), 122–129.

Robins, J., M. Sued, Q. Lei-Gomez, A. Rotnitzky, et al. (2007). Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science 22*(4), 544–559.

Robins, J. M., A. Rotnitzky, and M. van der Laan (2000). On profile likelihood: Comment. *Journal of the American Statistical Association*, 477–482.

Robinson, P. and C. Särndal (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā , A 45*, 240–248.

Rosenbaum, P. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A 5*(147), 656–666.

Rosenbaum, P. R. (2002). *Observational Studies* (Second Edition ed.). New York: Springer.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rubin, D. (1976). Inference and missing data. *Biometrika 63*(3), 581. With Discussion.

Rubin, D. and N. Thomas (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics 52*, 254–268.

Rubin, D. B. (2004)). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics 3*, 343–367.

Rubin, D. B. (2006). Estimating treatment effects from nonrandomized studies using subclassification on propensity scores. In D. A. Hantula (Ed.), *In Advances in Social and Organizational Psychology: A Tribute to Ralph Rosnow*, Volume 41–59, Erlbaum, Mahwah, NJ.

Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine 28*(9), 1420–1423.

Särndal, C. and S. Lundstrom (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics-Stockholm 24*(2), 167.

Särndal, C. E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal Statistical Planning and Inference 7*, 155–70.

Särndal, C.-E., B. Swensson, and J. Wretmann (1992). *Model-Assisted Survey Sampling*. New York: Springer- Verlag.

Scharfstein, D., A. Rotnitzky, and J. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Schonlau, M., R. Fricker, and M. Elliott (2002). Conducting research surveys via e-mail and the web. Monograph 1480, RAND Corporation. Accessed from `http://www.rand.org/pubs/monographreports/MR1480`.

Schonlau, M., A. Van Soest, and A. Kapteyn (2007). Conducting research surveys via e-mail and the web. Monograph 506, RAND Corporation. Accessed from `http://ssrn.com/abstract=1006108 or http://dx.doi.org/10.2139/ssrn.1006108`.

Schonlau, M., K. Zapert, L. Simon, K. Sanstad, S. Marcus, J. Adams, M. Spranca, H. Kan, R. Turner, and S. Berry (2004). A comparison between responses from a propensity-weighted web survey and an identical rdd survey. *Social Science Computer Review 22*(1), 128–138.

Shao, J. and D. Tu (1996). *The jackknife and bootstrap*. New York: Springer.

Smith, T. M. F. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society A 146*, 394–403.

Smith, T. M. F. and R. A. Sugden (1988). Sampling and assignment mechanisms in experiments, surveys and observational studies. *International Statistical Review 56*(2), 165–180.

Steel, M. (2012). The labour force survey data quality issues. Technical report, Office for National Statistics.

Stefanski, L. and D. Boos (2002). The calculus of m-estimation. *The American Statistician 56*(1), 29–38.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association 101*(476), 1619–1637.

Tan, Z. (2007). Comment: Understanding or, ps and dr. *Statistical Science 22*(4), 560–568.

Taylor, H., J. Bremer, C. Overmeyer, J. W. Siegel, and G. Terhanian (2001). The record of internet-based opinion polls in predicting the results of 72 races in the november 2000 us elections. *International Journal of Market Research 43*(2), 127–136.

Terhanian, G., J. Bremer, R. Smith, and R. Thomas (2000). Correcting data from online survey for the effects of nonrandom selection and nonrandom assignment. Research paper, Harris Interactive.

Tillé, Y. (2006). *Sampling algorithms.* Springer Verlag.

Tsiatis, A. and M. Davidian (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics 22*(4), 569.

Valliant, R. (2002). Variance estimation for the general regression estimator. *Survey Methodology 28*(1), 103–108.

Valliant, R. and J. Dever (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research 40*(1), 105.

Valliant, R., A. Dorfman, and R. Royall (2000). *Finite population sampling and inference: a prediction approach*, Volume 504. John Wiley & Sons.

Varedian, M. and G. Forsman (2003). Comparing propensity score weighting with other weighting methods: A case study on web data. In *Proceedings of the Section on Survey Statistics.*

Vartivarian, S. L. and R. Little (2003). Weighting adjustments for unit nonresponse with multiple outcome variables. Working Paper 21, The University of Michigan Department of Biostatistics Working Paper Series.

Weitzen, S., K. L. Lapane, A. Y. Toledano, A. L. Hume, and V. Mor (2004). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and drug safety 13*(12), 841–853.

Williamson, E., R. Morley, A. Lucas, and J. Carpenter (2012). Variance estimation for stratified propensity score estimators. *Statistics in Medicine 31*, 1617–1632.