ABSTRACT

Title of Dissertation: STATISTICAL ESTIMATION METHODS IN VOLUNTEER PANEL WEB SURVEYS

Sunghee Lee, Ph.D., 2004

Dissertation Directed By: Professor, Richard Valliant
Joint Program in Survey Methodology

Data collected through Web surveys, in general, do not adopt traditional probability-based sample designs. Therefore, the inferential techniques used for probability samples may not be guaranteed to be correct for Web surveys without adjustment, and estimates from these surveys are likely to be biased. However, research on the statistical aspect of Web surveys is lacking relative to other aspects of Web surveys.

Propensity score adjustment (PSA) has been suggested as an alternative for statistically surmounting inherent problems, namely nonrandomized sample selection, in volunteer Web surveys. However, there has been a minimal amount of evidence for its applicability and performance, and the implications are not conclusive. Moreover, PSA does not take into account problems occurring from uncertain coverage of sampling frames in volunteer panel Web surveys.

This study attempted to develop alternative statistical estimation methods for volunteer Web surveys and evaluate their effectiveness in adjusting biases arising from nonrandomized selection and unequal coverage in volunteer Web surveys.

Specifically, the proposed adjustment used a two-step approach. First, PSA was utilized as a method to correct for nonrandomized sample selection, and secondly calibration adjustment was used for uncertain coverage of the sampling frames.

The investigation found that the proposed estimation methods showed a potential for reducing the selection and coverage bias in estimates from volunteer panel Web surveys. The combined two-step adjustment not only reduced bias but also mean square errors to a greater degree than each individual adjustment. While the findings from this study may shed some light on Web survey data utilization, there are additional areas to be considered and explored. First, the proposed adjustment decreased bias but did not completely remove it. The adjusted estimates showed a larger variability than the unadjusted ones. The adjusted estimator was no longer in the linear form, but an appropriate variance estimator has not been developed yet. Finally, naively applying the variance estimator for linear statistics highly overestimated the variance, resulting in understating the efficiency of the survey estimates.

STATISTICAL ESTIMATION METHODS IN VOLUTEER PANEL WEB SURVEYS

By

Sunghee Lee

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2004

Advisory Committee:
Research Professor Richard Valliant, Chair
Research Professor J. Michael Brick
Research Associate Professor Michael P. Couper
Professor Partha Lahiri
Professor Robert Mislevy
Professor Trivellore E. Raghunathan

Acknowledgements

Collection of data in the Behavioral Risk Factor Surveillance Survey was funded in part through grant no. UR6/CCU517481-03 from the National Center for Health Statistics to the Michigan Center for Excellence in Health Statistics.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Survey methodology has a relatively short history as an academic field. It was not until the infamous debacle of the 1936 presidential election polling by the *Literary Digest* that the needs for scientific data collection were recognized. Since then, the survey methodology field has evolved dynamically along with the cultural and technological changes in the society.

Among the evolutions the most notable is the telephone interview (Groves and Kahn, 1979; Dillman, 1998; and Dillman, 2002). When the idea of conducting surveys over telephone was first introduced, researchers were not fully convinced about its utility, because the failed *Literary Digest* poll used a telephone list and because the prevailing belief was that surveys should involve face-to-face interactions. Since the *Health Survey Methods Conference* in 1972 where telephone interviewing first received attention as a serious data collection mode (Dillman, 1998), there has been 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 (e.g., Groves, 1989). A well-defined probability sampling procedure by random digit dialing has also been developed for telephone surveys (e.g., Mitofsky, 1970; Waksberg, 1978; Lepkowski, 1988; Casady and Lepkowski, 1993). Practical considerations and societal changes have also 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 most developed countries.

The survey research field is experiencing another challenging breakthrough – Internet surveys. The origin of the Internet dates as early as 1962 when J.C.R. Licklider raised the 'Galactic Network' concept which depicted a set of computers globally interconnected through which everyone could quickly access data and programs from any site (Leiner *et al.*, 2000). This was initiated by the military during the Cold War (Slevin, 2000), which set up the Advanced Research Projects Agency (ARPA) within the US Department of Defense in order to develop technologies for interlinking computer networks and facilitating computer-mediated communication. In 1969, ARPANET, the first packet switching network of four host computers at universities in the southwestern US, was launched and is the origin from which the Internet has grown. The Internet embodies a key underlying technical idea – open architecture networking (Leiner *et al.*, 2000). Under this networking, the choice of individual network technology is not dictated by one particular network architecture which enables coexistence of multiple independent networks of rather arbitrary design.

Widespread development of Local Area Networking (LAN) and personal computers in the 1980's sped up the usage of the Internet by the public. In 1992, CERN (the European Laboratory for Particle Physics) released the World Wide Web (WWW), graphics-based software. At the similar time, HyperText Markup Language (HTML) was invented at CERN. These two components later led to Web browsers, such as Netscape® and Microsoft Explorer® (Gattiker, 2001).

Now, utilization of the Internet is heavily dependent on graphics-based interaction, as more and more sites adopt this technology and graphical browsers are used to access the Internet. According to Leiner *et al.* (2000), the Internet is a world-wide broadcasting capability, a mechanism for information dissemination, and a medium for collaboration and interaction between individuals and their computers regardless of geographic locations.

There are various forms of the Internet – e-mail, newsgroups (Usenet), Multi-User Domains (MUDs), Internet Relay Chat (IRC), File Transferring Program (FTP), electronic mailing lists (listserv) and WWW (Web, hereafter) are some of the examples. Compared to other applications, the Web is user friendly as it does not require a high level of computing knowledge. The contents on the Web are displayed on browsers that enable an intuitive graphic-based interface between the contents and the web users. Sorting, retrieving, and sharing information based on a web of hyperlinks and hypertext are not complicated. Thanks to hypertext and hyperlinks, Web users may move from one webpage to another without a glitch, while deciding which information they wish to have transferred to their browser and which links they want to skip. Moreover, unlike conventional communication media relying on nonhuman channels, the Web carries information expressed in a multi-media format including text, sound, and still and moving graphics. Due to its prominence, the term "Web" will be used interchangeably with "Internet" throughout this study, although it is one device to employ the Internet.

The popularity of personal computers and the convenience of the Web have made it the fastest growing communication medium in developed countries. It is not a radical

idea any more to have a flower shop deliver a bouquet to parents in another country or to pay bills over the Web. Technology changes; so does the society.

'Our survey methods are more a dependent variable of society than an independent variable,' according to Dillman (2002). The ideal survey methodology is likely to reflect the society and its culture. 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 (Taylor and Terhanian, 2003). All these changes in survey modes occur because survey methods inevitably manifest societal trends.

Nevertheless, there are mixed views about Web surveys. While many researchers think that Web surveys have a great potential as an addition to the existing methods and for the measurement improvement (e.g., Taylor, 2000; Couper, 2001a, 2001b; Dillman, 2002), others express pessimistic conjectures towards Web surveys (e.g., Mitofsky, 1999). The negative views seem due to the fact that there does not exist a well-accepted Web survey methodology for selecting probability sample surveys targeting the general population, as Web surveys are new to the field and the rapid increase in their use has far surpassed that of the methodological development. No matter how strongly survey methodologists warn about limitations of Web survey quality, it is unlikely that the field will give up on Web surveys. 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 the methodologists' responsibility to devise ways to improve Web survey statistical methods (e.g., sample selection and estimation) and measurement techniques (e.g., questionnaire design and interface usability).

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 do participate in the surveys. Less attention has been given to statistical inference based on Web surveys. A basic statistical question is whether the data collected from a set of Web survey respondents can be used to make inferences about a desired target population. However, statistical properties of Web survey outcomes deviate from those in traditional surveys. Survey organizations may hope that their Web surveys represent the general population of households or persons. But, it is unrealistic to assume that Web surveys targeting the general population are based on randomization, because the frame coverage is uncertain, which means that drawing a probability sample from the target population is impossible. Moreover, response rates on Web surveys are low. Therefore, it is highly likely that Web surveys inherently carry errors related to coverage, sampling, and nonresponse.

There are post-survey statistical approaches to compensate for these errors in traditional surveys, such as face-to-face and telephone surveys. Their performance on Web survey errors is open to discussion, as the underlying mechanism of these errors may be unique for Web surveys. To explore this possibility, this study will focus on the statistical aspect of Web surveys, more specifically post-survey adjustment. It will examine the existing survey adjustment methods and expand the possibilities by proposing and examining propensity score adjustment and calibration methods specifically devised for Web surveys.

The remainder of this study is comprised of the following eight chapters. The classification of the current Web survey practice and the structure of Web survey errors related to the cyber culture and Web usage will be introduced in Chapter 2. Chapter 3 will state the purposes of this dissertation and summarize of the work in the subsequent chapters. The extent to which traditional post-survey adjustment methods correct for coverage and nonresponse error will be evaluated in Chapter 4. The core of this study is Chapter 5, 6, and 7 where the propensity score adjustment and calibration will be examined as alternatives to more traditional post-stratification adjustment. Chapter 5 will start by documenting the propensity score adjustment as a bias reduction method in observational studies and will review the literature on propensity score adjustment. Chapter 6 will identify how this method along with calibration adjustment can improve estimation using Web survey data by relating to the characteristics of the Web sample discussed in Chapter 2. It will provide mathematical notation for the propensity score adjustment as well as the calibration adjustment. Chapter 7 will consist of two case studies where proposed adjustment methods are applied to the survey data and will appraise the magnitude of error reduction in simulations. Propensity score model building strategies and variance estimation issues will be also examined. This study will conclude with Chapter 8 with a summary of the implications and limitations of this research and suggestions for future research in order to advance this work.

Chapter 2: Web Survey Practice and Its Errors

Surveys can be conducted on the Web at any time in any place with many types of colors and multi-media features literally at no cost. The facts that an increasing number of people use the Internet is an ordinary tool of communication, a channel for information, and a place for various daily activities have attracted an enormous amount of attention from survey researchers. The growth of Web survey practice is rapid, considering that the possibility of conducting surveys on the Web was first discussed less than a decade ago. There is an apparent gap between statistical and measurement features of Web survey practice and methodological research. Despite the facts that Web surveys have not been thoroughly studied and survey professionals express suspicions about their quality, the Internet seems to be somewhat overloaded with these dubious data collections.[1] This, however, should not discourage survey methodologists from seeing the Web as a potential data collection tool. Understanding Web surveys from different disciplinary and methodological perspectives should improve the quality of Web-based surveys.

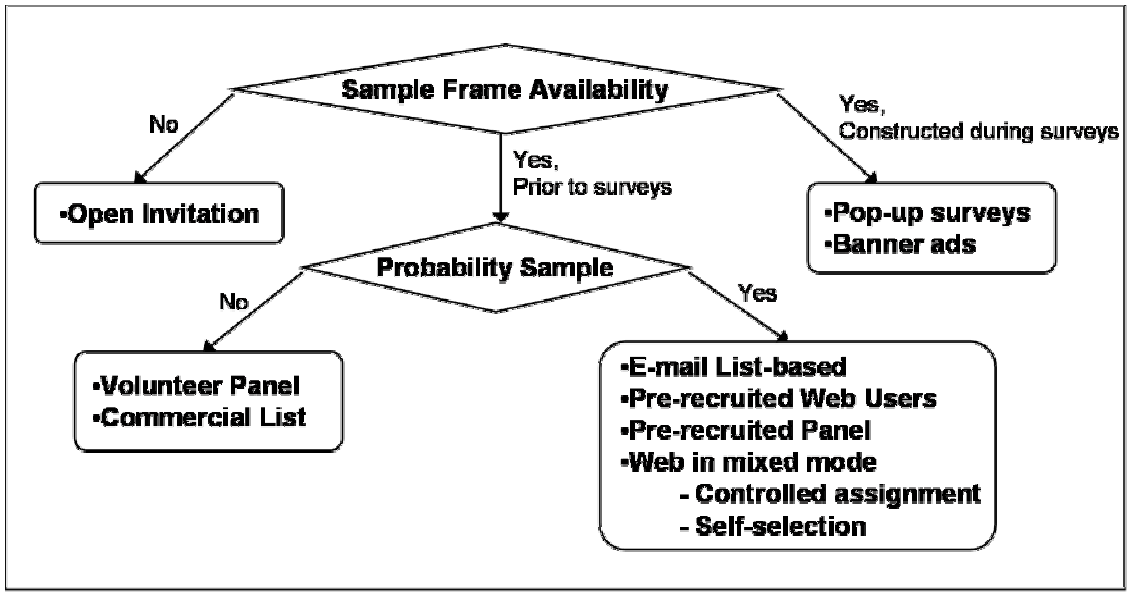
2.1 Types of Web Surveys

Web surveys are not the same as Internet surveys, as Internet surveys include both Web and e-mail surveys, whereas Web surveys include only those presented via WWW

[1] The existence of websites, which claim that Internet users can make money by taking surveys, could be evidence of this concern (e.g. http://www.surveys4money.com)

browsers. Due to limitations with storage and software compatibility, e-mail surveys are less popular than Web surveys; thus, this research mainly focuses on Web surveys.

Web surveys can be first classified into three categories as in Figure 1. This classification is based on the availability and the construction method of a sampling frame (Couper, 2001a; Manfreda, 2001; Couper, 2002; Couper and Tourangeau, 2002). When sampling frames are not available, the open invitation type of Web survey is conducted. Examples of this are entertainment polls, like QUICKVOTE on http://www.cnn.com, and unrestricted self-selection surveys. This survey is virtually open to anyone with Web access, and if they want to take the survey, they can respond as many times as they wish. Open invitation Web surveys are not suitable for scientific research, because researchers do not have any control over the participation mechanism.



Source: Manfreda (2001); Couper (2001a); Couper (2002)

Figure 2.1. Classification of Web Surveys

The second type of Web surveys constructs a list of participants during data collection, and this list may be used as a frame. Survey participants are recruited as they are intercepted to designated survey sites or encounter pop-up surveys or banner ads, when they log onto certain websites for other purposes. Depending on the intercept implementation methods, these surveys may accommodate probability sampling. However, their response rates are typically very low (far less than 10%), making this type of Web surveys unsuitable for scientific research.

The third category of Web surveys has a sampling frame prior to data collection, which allows individual invitation of sample units. Researchers may have full control over respondents' participation by restricting the survey access. The quality of this Web survey method is considered better than the previous ones. This Web survey is further dichotomized depending on the probabilistic nature of the sample. The first uses nonprobability samples drawn from volunteer panels or commercially available e-mail lists. One example for this type would be the method currently used by Harris Interactive. Panel members in volunteer Web surveys self-select to join the panel, and commercial e-mail lists include Internet users who register for some other services on the Web. Such frames may have duplicate listings and there can be problems in identifying multiple listings on the sampling frame as well as in the sample and, thus, in obtaining the probability of inclusion.

The second type of the Web surveys with sample frames constructed prior to data collection uses probability sampling. Under this method, there are currently four different ways to conduct Web surveys: (1) Web surveys using a list of some unique population whose members all have Web access, (2) Web surveys recruiting Internet

users via traditional survey modes with probabilistic mechanism, (3) Web surveys providing Web access to a set of recruited panel members who were probabilistically sampled from the general population, and (4) Web survey option in mixed-mode probability sample surveys[2]. The probability of inclusion is obtainable in these Web surveys and may be used in estimation. Strictly speaking, design-based statistical inferences can be drawn only under these last four Web survey methods.

2.2 Cyber Culture and Web Surveys

One way of gaining fundamental knowledge about Web surveys is to understand cyber culture. This is because the relationship between survey methods and the cultural phenomena is substantial as discussed in Chapter 1. This section will examine the culture in cyberspace in order to provide integrative views on the Web survey, its respondents and its errors.

The Internet is a special medium, for it enables both reciprocal and non-reciprocal communication. On the one hand, the Internet forms some types of solidarity among its users by deconstructing physical and social boundaries (Reid, 1991) and connecting all users who are willing to participate. On the other hand, the concept of 'community' does not appear to exist in the cyber world, because the culture in the cyber community is distinctive from that in the everyday community.

[2] Web options in mixed-mode surveys differ by the control method of participation assignment. While some mixed-mode surveys use a random assignment, enabling researchers to know which units are answering on the Web prior to survey recipients' participation, others make respondents choose a preferred mode.

Cyber culture tends to have been treated negatively, as it is viewed to bring a destructive effect on both personal identity and social culture (Turkle, 1995). Turkle (1995) argues that 'in the real-time communities of cyberspace, we are dwellers on the threshold between the real and the virtual, unsure of our footing, inventing ourselves as we go along.' Cyber world connives at personal identities being de-centered, dispersed and multiplied. This fluctuating identity may be best portrayed by one term – anonymity.

Anonymity, indeed, is one of the highlights in identity formation on the Internet (Slevin, 2000; Burnett and Marshall, 2003). While scarce in real life, anonymity is omnipresent in cyber space. The idea that the physical or lawful being of users is not always verifiable on the Internet seems to have led people to counterfeit their identities or appear under many different identities. Nonetheless, the reality is that our Web activities leave remnants that can be traced and identified. While anonymity or identity invention is an elusive idea, Internet users misperceive that others are not able to obtain their true identity, unless they reveal it. 'Anonymity continues to operate as the boundary that one traverses as a Web user – whether as a lurker in chatgroups or as a multiple personality in usegroups and chatgroups (Burnett and Marshall, 2003)'.

The possibility of locating one's true identity in cyberspace does not stop Internet users from enjoying their anonymity. Ironically, this possibility triggers another issue – threats to the real-life privacy. Internet users are aware that it is easy to obtain personal information with the development of the Internet and that it is possible for some strangers to access and use their identity. Privacy has become a luxury item in the cyber world (Moore, 2002), and this has increased the privacy concerns.

The Internet has been found by some authors to cause a negative effect on interpersonal relationships (Kraut *et al.*, 1998; Nie and Erbring, 2000). Internet usage weakens traditional relationships, lessens total social involvement, increases loneliness and depression. These authors argue that the quality of the Internet social relationships is poorer than those of the face-to-face relationships and that the time spent only to create a weak tie in the cyberspace takes away opportunities to form strong face-to-face ties with real human beings. Heavy Internet usage somehow makes its users lose touch with the social environment. In sum, Internet society does not require as much coherence in interpersonal relationships as real society does.

The 'fluctuating identity' and 'social incoherence' (Burnett and Marshall, 2003) in cyberspace may affect response behavior in Web surveys in three ways. First, people may perceive a lower degree of social obligation, when they are online. E-mail addresses, the common route to sample and contact survey recipients, may not convey as much importance as needed for survey participation and completion. Moreover, the recipients know that their individual identity is not easy to verify through e-mail addresses. This may provide a safe feeling, when they discard the survey invitations or even when they behave as if they are someone else and forge the responses accordingly. The weak interpersonal ties and less-structured culture in the Internet society add more reasons for lowered social obligation. Social exchange theory, once used to explain how to stimulate survey cooperation in other surveys (e.g., Groves and Couper, 1998; Groves, 1989; Dillman, 2000), may not hold in Web surveys.

Second, the heightened privacy concern on the Internet may make online behavior more vigilant, even when there is a slight chance of exposing the true identity. Two

survey errors may arise from the respondent behavior caused by the privacy concern. First, when an Internet user receives survey invitation e-mail from some organization that the user is not familiar with, the person is unlikely to pay attention to the invitation. Second, the user may want to provide desirable responses if some well-known organization, which the user believes to have a capability to track him or her down, conducts the survey. In this case, the respondent may want to depict himself or herself in a socially acceptable way. The second error may be completely opposite of Web survey pioneers' prediction that Web surveys, as a type of self-administered data collection, will obtain information free from the self-presentation pressure.

Third, Web survey respondents' behavior may be affected by their Web usage behavior. Internet users are used to switching from one task to another by clicking and closing windows or moving to other websites, whenever they encounter something other than what they expect or something that they are not necessarily interested in. There are countless distracting features on the Web, from pop-up ads to instant messengers. This environment itself makes it difficult for Web user to focus their attention on one task. Accordingly, survey recipients may not open the invitation e-mail, if it appears uninteresting. Even when survey recipients open the survey, there is a great chance to depart from the survey at any time, if they find the survey is not as interesting or urgent as they first think. Their return to the survey is not guaranteed. There is likely to be more than one stimulus on the recipients' computer monitor, although survey researchers wish that the survey questionnaire is the only feature. In this case, the level of cognitive capacity consumed solely for the Web survey may be low. Computer viruses may be another factor of the Internet environment. Since they are spread widely via the Internet,

one recommendation for computer protection is to delete any suspicious e-mails. Imagine a Web survey fielded unfortunately during a virus epidemic – why would people keep the invitation e-mail in their mailbox?

2.3 Web Usage by Demographic Characteristics and Web Surveys

The demographic characteristics of Web users are another source of understanding Web survey respondents and may reveal information on their behaviors and subsequent survey errors. As in the previous section, we will examine who is on the Web and who Web surveys are likely to attract.

Existing Web survey literature seems to take the possibility of conducting useful surveys on the Web for granted. This can be deceptive, because only a selected portion of the general population is privileged to have Internet access. Futurologist Toffler (1970; 1980; 1991), even before the Internet was introduced to the public, predicted that the technological changes would endanger people by leaving them behind in the postindustrial economy, if they do not heed and act on the changes. As predicted in his book *Powershift* (1991), an unconventional economic power paradigm is emerging – the power is shifting from the people with more material resources to those with more information. The Internet is a critical medium to acquire bountiful and opportune information in a short time. However, the Internet usage is not evenly distributed with respect to the socio-economic status and demographic characteristics, which leads to an unequal chance to obtain the power predicted by Toffler, especially for less-privileged people.

Internet access rates differ considerably among countries, implying that the target population that can be covered by Web surveys will be much different as well. According to the 2003 International Telecommunication Union report (available at http://www.itu.int/ITU-D/ict/statistics/), there are only ten countries where more than half the population uses the Internet.[3] Some countries, like Myanmar, Tajikistan and Democratic Republic of the Congo, less than 10 out of 10,000 people use the Internet. The divergent Internet usage level across countries seems closely related to their economic status and telecommunication infrastructure, which is, in turn, related to education.

Let us assume that there is a survey conducted in the U.S. including U.S. territories and outlying areas via Web. Given the facts that Web users may be different from nonusers and that people from each state, for instance, may be disproportionately represented, results from this survey may not be generalized to any degree. Until there are substantial proportions of Internet users around the world, the possibility of conducting Web surveys free from the physical and geographical boundaries may remain as a daydream.

In the U.S., there is a broad range of information about Web usage by different demographic groups. There is a great concern about *digital divide*, the difference between online and offline population. *A Nation Online* (2002) indicated uneven Internet usage by age, income level, educational attainment, employment status, race/ethnicity, household composition, urbanicity, and health status. Not surprisingly, young people are

[3] These countries are: Iceland: 67.5%, Republic of Korea: 60.3%, Sweden: 57.3%, US: 55.1%, New Zealand: 52.6%, Netherlands: 52.2%, Canada: 51.3%, Finland: 50.9%, Singapore: 50.4%, Norway: 50.3%.

leading the Internet usage, as 75% of youth between the ages of 5 and 17 years old use the Internet. In addition, the following groups of people are less likely to use the Internet than their respective counterparts: people with lower income, without employment, with lower education, or with disabilities; people living in the central city, or in non-family household or family household without children; or Blacks and Hispanics. Although there is evidence that the gaps in those characteristics between online and offline population are decreasing (*US Department of Commerce*, 2002), the uneven levels of Web usage with respect to these background characteristics are likely to remain. Moreover, there will remain certain groups of people who are unable to go online for financial, technical, or health reasons.

This *digital divide* may affect the quality of Web surveys. Unless the people on the Internet are the population of interest, Web surveys are likely to include people with higher socioeconomic status and more socially engaged and younger people at disproportionately higher rates than traditional surveys. Depending on the target population of a survey, this can result in unequal coverage, as Internet nonusers may be systematically under-represented. Internet users may also have distinctive survey response behaviors – for example, higher noncontact or nonresponse rates or lower compliance in completing the survey task. This will also cause different combinations and levels of survey errors than traditional surveys.

2.4 Web Survey Errors

The best way to understand Web surveys is a systematic comparison between Web surveys and traditional surveys, such as telephone and face-to-face surveys, with

respect to total survey errors (Deming, 1944; Groves, 1989). Following the traditional approach illustrated in Groves (1989), this section will examine all components of the total survey errors: coverage error, sampling error, nonresponse error, and measurement error in Web surveys (also refer to Couper, 2002; Couper and Tourangeau, 2002).

2.4.1 Coverage error

Coverage error arises when the survey frame does not cover the population of interest. Although Web surveys can be subject to either undercoverage or overcoverage, the former is the most serious problem in Web surveys. The Internet users in US are estimated by *A Nation Online* (2002) at 143 million, and about two million additional Americans go online annually. It is likely that the world shall see an increase in the number of Internet users and the continuation of this trend. While these numbers and the growth in the numbers are impressive, Internet users account for 54% of the American population. Consequently, even though the Internet population is large and growing, a huge portion of the general population would be omitted from a Web survey. Although some may claim that their large sample sizes would protect their surveys from systematic exclusion of large segment of the population, this is fallacious as sample sizes are not related to coverage error at all – coverage error is a function of coverage rate and differences between covered and omitted units.

It is true that there are certain populations whose members all have Web access, for example, faculty or students at colleges or universities and employees at government agencies or large corporations. In Web surveys targeting these populations, the frame may achieve full coverage, and their coverage errors may not be serious. Once the Web survey target population departs from these special groups, the coverage properties

become jeopardized. A possible solution for this problem may be providing Internet access to the offline population. This idea is currently practiced by Knowledge Networks (Huggins and Eyerman, 2001) – pre-recruited panel Web survey examined in Section 2.1. In order to construct a controlled panel, first eligible telephone numbers are called via random digit dialing, and eligible people who answer the phone are invited to join a Web survey panel. If the call recipients agree to be panel members, they receive a Web TV[4], regardless of their Web usage status prior to the recruitment.[5]

Overcoverage of Web surveys is related to the possibility of multiple Internet identities which Section 2.2 introduced as an attribute of the cyber culture. In effect, any Internet users encounter many chances to set up multiple e-mail addresses, whether they intend to or not. For instance, a college freshman has an e-mail address which he has used since high school and is using it to communicate with his high school friends and his family. His college automatically assigned him another e-mail address, and he mainly uses it for school-related matters. Imagine his part-time job involves some computing and he sets up his third e-mail address for better work delivery within the company. This student already has three e-mail addresses. It is a matter of time for him to get assigned additional e-mail addresses that he may or may not be aware of. This possibility implies existence of overcoverage in volunteer panel Web surveys and commercially available e-mail list-based Web surveys. There is a potential threat that Web survey volunteers may

[4] In principle, this may solve coverage problems, but its operation has shown some limitations: there are areas where the Web TV service is not available. This may be viewed as nonresponse error. However, it is not clear whether people who do not respond to the RDD invitation or who decline to join the panel affect coverage properties systematically.

[5] KN is now allowing panel members who already have a computer and an Internet access to use their own system. For these members, KN provides different monetary incentives.

join the panel multiple times with different identities in order to increase the odds of receiving incentives. For commercial e-mail lists, it is impossible to distinguish to whom each e-mail address belongs. One approach to identify the duplicate units and adjust for them in these frames is to ask a sample person whether he/she has other email addresses and, if so and if possible, what they are. The selection probability for each person could then be adjusted in the same way that a household selection probability is adjusted in a random digit dialing telephone survey where the household has more than one telephone line.

2.4.2 Sampling Error

Sampling error occurs due to the fact that not every unit in the target population is in the survey. The concept is usually considered in the context of probability sampling. In Web survey practice, nonprobability sampling is dominant because of its convenience and inexpensiveness. Researchers should bear in mind that nonprobability sampling can give biased estimates, as in the *Literary Digest* incident, and requires that strong structural assumptions hold in order for inferences to be valid.

There is an effort by Harris Interactive as previously introduced to compensate for the coverage and sampling errors by sophisticated weighting. This technique adopts propensity score adjustment originally proposed by Rubin and Rosenbaum (1983) for causal inferences using observational data. Propensity score adjustment balances out the covariate differences between the treatment and control groups whose assignment mechanism is not random. Harris Interactive collects reference survey data through RDD telephone surveys as if they come from a control group and Web survey data as a treatment group. Through the use of weights, the estimated distribution from the Web

survey is adjusted to match that of the reference survey on certain variables that are collected in both. Although Harris Interactive has been advocating the effectiveness of propensity score adjustment, there have not been well-documented technical procedures for this application. Moreover, the amount of evaluation on the adjustment performance is very limited (e.g., Terhanian *et al.*, 2000a; Taylor *et al.*, 2001; Schonlau *et al.*, 2003; Varedian and Forsman, 2003), which leads to inconclusive implications. This method will be elaborated in Chapter 5 and 6 and examined in Chapter 7.

2.4.3 Nonresponse Error

Nonresponse error arises when not all survey recipients respond. This error is a multiplicative function of two components: the response rate and the difference between respondents and nonrespondents. One substantial problem of Web survey nonresponse is that response rates are not always measurable. For volunteer panel Web surveys or open-invitation Web surveys, it is impossible to measure the number of potential respondents who are actually exposed to the survey invitation. Web surveys using commercial e-mail lists may potentially allow response rates to be measured, but confront difficulties identifying whether the e-mail addresses are still being used. Thus, the nonresponse rate among eligibles is entangled with the rate of ineligibility on the frame.

Web surveys whose response rates are measurable have achieved relatively poor results. Response rates for the intercept or pop-up surveys do not exceed 10%; around 20 to 30% for volunteer panel Web surveys (e.g., Harris Interactive); and around 50% for surveys on panel members who are given Web access (e.g., Knowledge Networks).

When the use of Web surveys started to increase, many researchers noted the problems associated with coverage and sampling errors. Interestingly, few were

concerned about the nonresponse in Web surveys. Some pioneers were even optimistic about the response rates by arguing that respondents could take surveys on the Internet at their convenience and this gives more chances to respond. In reality, response rates in Web surveys are low relative to other survey modes. After adjusting for the cumulative nature of Web panel recruitment and survey participation, the final response rates may dip far below the nominal response rates noted above.

What are the possible causes of Web survey nonresponse? First of all, compared to traditional surveys, it is difficult in a Web survey to provide tangible financial incentives and is impossible to build rapport between the survey conductor and takers. This is because an interviewer who plays a role as a motivator and a mediator is eliminated. It is also related to the laxity of the Internet society – Web survey recipients may not feel obligated to abide by the survey request.

A second source of nonresponse error may be found in limited computer literacy among some groups. While it is true that browsing websites does not require a high level of computer literacy thanks to the adoption of Graphic User Interfaces, there are people, especially older and less educated people, who may still feel uncomfortable with using computers and the Internet. Although the Web survey design quality is most likely to influence the measurement error which will be examined shortly, the lack of computer literacy may not permit them to access or operate Web surveys. When considering the frequency of encountering badly designed Web questionnaires, the cognitive challenges that these people may perceive on top of the burden caused by low computer literacy, may elicit a high level of nonresponse.

The level of system accessibility may be another reason. Depending on the popularity and the age of computer platforms and/or Internet browsers, Web questionnaires may appear in various ways. Some survey recipients with an older platform or a less popular browser, for instance, may not even have a chance to view the questionnaire as implemented. Those with slower modems or processors may experience a lengthy delay in questionnaire loading and give up carrying out survey task. These recipients become nonrespondents or partial respondents, not because they avoid surveys, but because their system restricts them from accessing survey instruments.

The most critical cause for nonresponse in Web surveys seems related to the cyber culture examined in Section 2.2. The guaranteed anonymity and relaxed social ties add more reasons for respondents to neglect the survey requests. Heightened concerns about the personal privacy may weaken the legitimacy of the survey organizations in the minds of potential respondents, while the authority of survey organizations has been found to have positive effect on the completion of other surveys (Presser *et al.*, 1992; Groves and Couper, 1998). Quick and easy navigation from one location to another or one task to another and distracting features on the Web may produce higher levels of nonresponse and break-offs.

2.4.4 Measurement Error

Unlike the previous three types of survey errors, measurement error exists within collected data. Among four survey error components, measurement is the area where Web surveys may have distinctive advantages over other data collection modes. Accordingly, it has been studied more rigorously than other error components.

What are the measurement advantages of conducting surveys on the Web? First, interviewers are eliminated, which can be a key source of response error and variance. Ideally, this nullifies interviewer effect on survey statistics and helps to minimize respondents' fear of exposing sensitive answers. This advantage, however, is common to all self-administered surveys.

Second, Web surveys with a minimal addition in programming make it feasible to automate and customize the questionnaires: skip patterns, item branching, randomization on question and response-option order, answer range checks, and tailoring of question wording may be built into the questionnaire. Feedback or error messages may be pre-programmed so that the survey instrument could point the respondents in the right direction whenever mistakes occur. Note that the automation and customization are not unique only for Web surveys – they are attainable in all computer assisted survey modes.

The greatest advantage of using the Web is its richness of visual presentation. There is an unlimited range of colors and images one can choose for Web surveys, which would cause a substantial cost increase in other modes. Even multi-media features, such as video clips, which are not always possible to implement in other modes, can be freely employed in Web surveys, if the respondents have the appropriate equipment. These unique characteristics of Web surveys may not only make survey instruments look more appealing but also reduce the cognitive and operational burden of respondents.

These advantageous attributes of Web surveys, unfortunately, may turn into disadvantages, because it is easy to overuse or misuse them. If colors, images and multi-media features do not match to the respondents' cognitive map, they may confuse respondents. This is because respondents may try to make inferences from those

features, which are not intended by the survey designers. Question wording customization could backfire with sensitive topics, as personalized questions may trigger respondents' privacy concerns. With feedbacks, help menus and instructions, Web surveys attempt to facilitate respondents' question comprehension and minimize questionnaire operation errors. However, it is uncertain whether respondents use these features and whether they find them informative and useful. Absence of interviewers may result in a greater chance of satisficing response behavior, as respondents may sense a lower degree of motivation.

Unlike other surveys, Web surveys demand a higher degree of cognitive capability and computer knowledge. In addition to the cognitive processes solely for survey tasks, respondents need to allocate their remaining cognitive capability to manage the questionnaire design components and distracting Web features and to understand the operation of the questionnaire. Unequal technological competence among respondents may cause a problem – novice and expert Internet users may encounter different burdens, therefore, produce different measurement errors. If a Web survey targets a population of novice Internet users, the measurement error may be detrimental.

We have examined types of Web surveys and integrated errors in Web surveys with the cyber culture and webographics. To recapitulate, first, it is important not to lump all types of Web surveys into one. Burnett and Marshall (2003) documented that "Unifying the Web into a simple medium is fraught with inconsistencies and exceptions to a degree that is unparalleled in past media. Researchers have been more successful at laying claim to the idea of 'television', where its intrinsic modality was evident." The same argument made by Burnett and Marshal (2003) seems to hold for Web surveys.

There are few variations of telephone surveys one can carryout. The error mechanism for each of these telephone surveys is rather simple and predictable. However, the story changes completely for Web surveys – there are a number of different Web surveys, at least nine types were identified in this chapter based on the method used for sampling. These surveys are all idiosyncratic with respect to survey errors – they differ from one another with respect to the most critical error components, the sources of errors, and the absolute and relative magnitude of each error. This may be clear in a comparison between open invitation and pre-recruited Web user Web surveys. While the latter is capable in covering the target population and drawing probability samples, the former is unlikely to achieve these. In addition, there is a dramatic difference in response rates between the two. The properties of measurement error, however, may be comparable. Therefore, it is necessary to understand and evaluate particular Web surveys at one time, not Web surveys as one unity.

Second, there is a need for systematic investigation of Web survey errors. Studies of Web survey error to date have made a laundry list of errors and are limited in providing a meaningful foundation of mechanisms for those errors. This chapter described a number of sources of Web survey errors in the cyber culture and *digital divide*. It may be necessary to incorporate findings from other fields in order to broaden the understanding of the error mechanism in Web surveys.

Chapter 3: Statement of Purpose and Work

The proposed research is intended to find innovative statistical approaches for adjusting errors caused by unrepresentativeness of Web surveys. Based on the implications in Chapter 2, among various types of Web surveys, this study will focus on one – volunteer panel Web surveys. The foremost problem is that, unlike in traditional surveys, the samples in this Web survey type are not guaranteed to be randomly selected. Units in those samples are comprised of either probabilistically or nonprobabilistically drawn units from a set of nonrandom volunteers. Because of nonresponse, the responding units generally cannot be considered as a probability sample even from the frame of volunteers. They are likely to systematically differ from the scope of survey target populations, reflecting the unequal ownership of a Web access and the impossibility to place a control on the frame population.

The occurrence of nonrandomization in Web surveys inevitably increases biases in survey estimates. Bias reduction becomes crucial to make use of results from these Web surveys. As the biases are difficult to control in the survey preparation phase, some post-survey adjustments may reduce bias more efficiently. There is one approach that has been discussed as a potential method of compensating for the nonrandomness in causal studies – propensity score adjustment. Harris Interactive first introduced propensity score adjustment for their Web survey data, which are collected from volunteer panels (e.g., Taylor, 2000; Terhanian and Bremer, 2000). Propensity score adjustment uses covariates collected in surveys and provides additional layer of weights in order to produce post-survey weights that ideally remedy selection bias in Web

surveys. Harris Interactive claims that the results from their volunteer panel Web surveys are generalizable to the U.S. population, according to their report which can be accessed from http://www.harrisinteractive.com/tech/HI_Methodology_Overview.pdf.

Although there have been a few studies examining the application of PSA for volunteer panel Web surveys (e.g., Schonlau *et al.* 2004, Danielssen, 2002, Varedian and Forsman, 2002, Taylor *et al.*, 2001, Taylor, 2000, Terhanian *et al.*, 2000), more in-depth evaluation is needed for a number of reasons. First, the resemblance between Web surveys and the situations where propensity score adjustment originated needs to be scrutinized, before adopting it for Web survey data. Second, the technical procedure of the propensity score adjustment is not well documented. This makes the adjustment method more a mystery than a well-proved scientific method. The mathematics behind the propensity score adjustment for Web survey data needs to be clearly presented. Third, adjusted Web estimates in those studies have often been compared to estimates from other surveys, typically telephone surveys which were conducted in parallel to the Web surveys. Since both estimates are subject to sampling, coverage, nonresponse, and measurement error, the implication of any observed differences is unclear. Fourth, existing studies have focused only on bias properties of the estimates. The other component of survey errors, variance, has not been examined, although propensity score adjustment is likely to increase variability. Weights, in general, add an extra component to the variability of the estimates and, thus, decrease the precision. Therefore, it is important to examine both aspects of errors in evaluating the performance of the propensity score adjustment. Fifth, some of the existing studies favored Web surveys by comparing the Web polling estimates and the election outcomes. These findings may not

be indicative of the quality of Web surveys on other subjects; these conclusions may be flawed, if Web survey respondents are more likely to vote than others. This fact alone may make Web surveys favorable, because, in this case, the likelihood of voting may determine the election outcomes. The last issue is that propensity score adjustment needs to be used in conjunction with another adjustment that compensates for the coverage errors. As we will show in later chapters, coverage adjustments are needed, because the propensity score adjustment can correct imbalances between the Web sample and some reference sample from the target population. It is worthwhile to examine the performance of the propensity score adjustment when interacting with other adjustments.

This research attempts to overcome the shortcomings in the existing literature of propensity score adjustment described above. It will examine the validity of modifying propensity score adjustments for studies other than causal inferences, exploit the adjustment as a candidate for improving Web survey data, present the mathematical procedure for its application, and evaluate its performance. The evaluation will be extensive, as it includes several study variables measuring different characteristics, the choice of covariates for building propensity score models, the inclusion of additional adjustments for coverage errors and its interaction with the propensity score adjustment, and the effect of adjustment on three aspects of errors: mean square error, bias and variance.

In order to accomplish the stated purposes, this research will carry out the following activities in subsequent chapters:

Chapter 4. Review and apply traditional adjustment methods, which are currently used to correct for nonresponse and coverage errors in Web surveys. Evaluate the performance of these adjustments.

Chapter 5. Introduce propensity score adjustments, and review the ways it can be applied: pair matching, subclassification, and covariance adjustment. Identify pertinence of employing propensity score adjustment for correcting estimates from Web survey data.

Chapter 6. Present the mathematical procedure for deriving weights using propensity score adjustment for the lack of randomness in the Web survey data. Introduce calibration as an additional adjustment method for compensating for coverage problems in Web survey data.

Chapter 7. Apply the identified propensity score adjustment method and calibration adjustment in two case studies. Simulation using the 2002 General Social Survey and 2002 Behavioral Risk Factor Surveillance Survey will be used for the application. The effectiveness of different types of adjustments will be discussed in relation to all error components.

Chapter 8. Conclude the research with its implications and limitations. Suggest directions that future research may take to address the limitations in this research.

Chapter 4: Application of Traditional Adjustments for Web Survey Data

4.1 Introduction

Possible sources of errors in Web surveys are examined in Chapter 2. The good news is that it may be possible to control those errors, especially nonresponse and coverage errors, using traditional post-survey statistical adjustments. This is feasible because Web survey companies create a panel pool whose members provide a range of background information before taking actual surveys. How effectively this can be done depends on the population to which inferences are to be made.

Pre-recruited probability panel Web surveys invented by Knowledge Networks (KN) described in Huggins and Eyerman (2001) use one of the distinctive survey protocols (See Figure 4.1 for the illustration). KN recruits a controlled panel via random digit dialing (RDD) and equips the entire panel with a Web accessing medium regardless of their prior Web usage status. At the first Web survey, the panel members take a profile survey collecting a range of background information. Therefore, it is the idea that for any given subsequent survey, the profile data are available for both respondents and nonrespondents that participate in the initial panel. In addition, reliable population estimates for many of the profile characteristics may be obtained from large-scale government surveys. The abundance of covariates may shed light on how different weighting approaches to Web surveys could improve data quality.

Ideally, the recruited Web panel described above represents the population of households or persons that have telephones as the panel members have a known

probability of selection into the panel and the samples drawn from the panel also have a known probability. This protocol may diminish unequal coverage and nonprobabilistic sampling problems, which are inherent to other Web surveys. It may be viewed as the most scientific method among Web surveys. However, there are significant complications. Partly shown in Figure 4.1 and partly discussed above, potential respondents go through roughly four stages before any survey that they participate: initial RDD panel recruitment, Web device installation, profile survey completion, and post profile panel retention. All these stages as well as actual survey participation are susceptible to some type of loss in the potential respondent pool. The coverage and nonresponse errors are intertwined in this protocol.

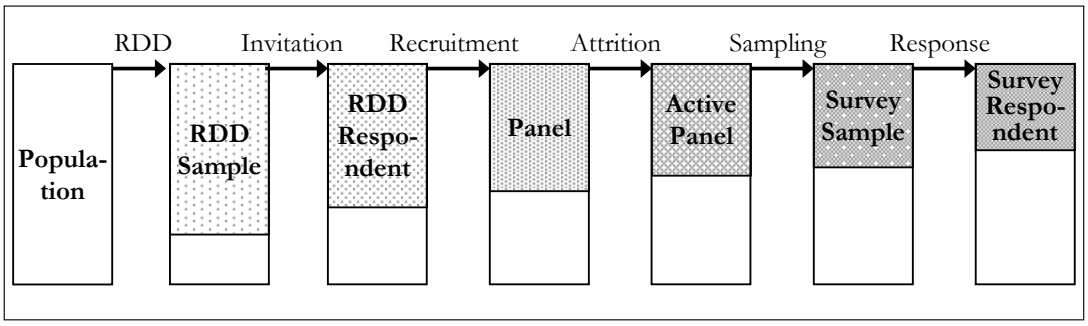


Figure 4.1. Protocol of Pre-recruited Probability Panel Web Surveys

Traditional post-survey adjustments, such as post-stratification, are used as a one-shot remedy for both errors in practice. The application of these adjustments implicitly assumes that the error mechanism is ignorable in the sense of Little and Rubin (1987). Since the Web survey in this chapter employs a multi-step protocol not found in other surveys, it may not be reasonable to assume ignorability. Therefore, traditional

adjustments may not be effective enough to compensate for coverage and nonresponse errors in Web surveys of this type. Moreover, the fact that these two errors are corrected simultaneously makes the respective error evaluation especially difficult to disentangle. One study (Vehovar and Manfreda, 1999) examined the effect of post-stratification for a Web survey, but its findings are somewhat limited. The sample was considered self-selected due to ambiguity of the eligibility of the units in the frame. The standard of comparison came from a telephone survey, which may not be a reliable source for adjustment as it is also subject to coverage and nonresponse errors.

This chapter attempts to evaluate the magnitude of nonresponse and coverage errors in a particular type of Web survey which aims to form and maintain a panel of respondents obtained through probability-based samples. There are statistics known for the Web survey respondents, the Web survey full sample, and the target population. This enables one to carry out a separate examination of the two errors. Section 4.2 will provide a detailed description about the data sources and the variables used in the analysis. Nonresponse properties will be evaluated in Section 4.3. The full sample which includes both respondents and nonrespondents will be assumed to provide the true values. Two adjustment approaches, ratio-raking and multiple imputation, will be applied. Unadjusted and two types of adjusted respondent estimates will be compared to the true values. Section 4.4 will examine the coverage error. Population estimates from a large government survey will be assumed to be true. Ratio-raking will be used to compensate for coverage error. The deviation of unadjusted and adjusted full sample estimates from the true values will be examined. The last section will summarize findings and raise considerations for future research.

4.2 Data Source

The analysis involves a two-stage adjustment and requires three types of data sets, one for the respondents, one for the full sample, and one for the population. The first two data sets will come from a Web survey and the last from the Current Population Survey (CPS).

4.2.1 Web Survey Data

The Web survey data come from the 2002 Survey Practicum class at the Joint Program in Survey Methodology (JPSM). Data collection was funded jointly by the Bureau of Labor Statistics (BLS) and JPSM for the practicum class at JPSM. The data were collected through a Web panel survey conducted by KN from August 23, 2002 to November 4, 2002. KN employs the special protocol introduced in Section I for its Web surveys. Note that the profile data are available for both Web survey respondents and its nonrespondents, as the KN web surveys are conducted solely among the panel members.

KN drew a sample of 2,501 households containing at least one parental figure with at least one child between the ages of 14 and 19 from its enrolled panel. Because later comparisons will be made between the Web survey and the CPS data, households with 18 and 19 year olds are dropped from the analysis to make the two stages of error compensation comparable.[6] This decreases the full sample to 1,700. Among the sampled units, 978 households completed the Web survey. The response rate to the Web survey was 57.4%. In order to qualify as a responding household, both parental figure and teen

[6] The closest possible teen age category identifiable in the CPS was 14 to 17

were expected to complete the survey. This might have played a negative role in the response rate. After incorporating nonresponse from the four pre-survey stages examined previously as well as two additional layers particular for this Web survey due to teen's involvement in the survey, the cumulative response rate became 5.5%. This final response rate is calculated with the nominal response rate within the survey (57.4%) in conjunction with other stages in the overall survey operation: panel recruitment rate (36%), Web TV connectability rate (67%), profile completion rate (98%), post-profile survey retention rate (47%), and parent's consent rate for teen's participation (86%).

Two data sets are created by combining the Web survey data and the profile data. The respondent data ($n = 978$) are constructed by applying the response status in the Web survey to the profile data. The KN full sample data ($n = 1,700$) are the entire profile data for the eligible sample units. The existence of profile data allows one to examine differences between survey responders and nonresponders and to examine various kinds of survey adjustments. The teen profiles are subject to a large amount of item missing data because parental consent was required for the profile survey. Thus, the target population for this analysis focuses only on parents living with at least one teen member between 14 and 17 in the same household.

4.2.2 Current Population Survey Data

The population estimates come from the 2001 September Current Population Survey (CPS).[7] This particular wave of CPS contained the Computer and Internet Use

[7] When considering the temporal equivalency, the 2002 September CPS seems more appealing, since the Web survey was conducted around that time. Nevertheless, this paper will use the data from 2001, as the 2002 data do not include computer and Internet

Supplement which collected information about Internet and computer usage of the eligible members of the sampled households (for methodological documents about this CPS supplement, refer to http://www.bls.census.gov/cps/computer/2001/smethdocz.htm). When restricting the 2001 September CPS sample to the scope of the target population defined above, the eligible sample size decreases from 143,300 to 11,290.

The CPS target population and its samples do include persons living in households that do not have telephones, whereas this type of Web survey starts from the telephone population. This is a source of noncomparibility between the coverage of our data set and the CPS, despite that only 3.5% of persons in the U.S. fall under nontelephone category.[8] However, Web survey organizations often claim that their surveys represent the full population including telephone as well as nontelephone. To evaluate this claim, we have used estimates based on the full CPS for comparison.

4.2.3 Variables of Interest and Covariates

All variables used in the analysis are available from both data sources. There are four dependent variables whose means will be estimated: number of owned computers in the household (none, one or more); prior Web usage experience (no, yes); employment status (unemployed, employed); and household size (number of household members), denoted as y_1, y_2, y_3 and y_4. Estimates based on these variables will be adjusted with respect to the following covariates: age level (20-40, 41-45, 46-50, 51 or older); education level (less than high school, high school, some college, college or above); ethnicity (White Non-Hispanics, Black Non-Hispanics, other Non-Hispanics, Hispanics);

---

usage and the distributions of covariates described in the following section are very close between the 2001 and the 2002 September CPS.

[8] The estimate is based on the 2001 CPS data.

region (Northeast, Midwest, South, West); and gender (male, female), denoted as $x_1,...,x_5$ in ratio-raking adjustment or $x_1,...,x_9$ in multiple imputation.[9] These covariates are selected as they are currently used in KN's existing ratio-raking procedure.[10]

The covariates will serve another function: all categories in all covariates will be the units of subgroup estimation. The reasons for estimating at the subgroup level are two-fold. First, studies make comparisons between Web surveys and traditional surveys typically at the total population level. Post-survey adjustments may correct the errors in the total population estimates, but not necessarily in the subgroup estimates. The second reason reflects the more realistic analytical interests – analyses are often done at the subgroup level to obtain more insightful conclusions than simply at the population level. For these reasons, this chapter expands the scope of estimation to the subgroup level.

4.3 Nonresponse Error Adjustment

Nonresponse error examined in this section focuses on the nonresponse on this particular Web survey among the full sample units (not the cumulative nonresponse for the entire panel). In this section, the full sample will be treated as a simple random sample of the target population and the weights will not be included in deriving estimates

[9] In multiple imputation, x_1, x_2, and x_9 are assigned to age, education, and gender, as the first two are considered as continuous and the last dichotomous. Ethnicity and region are polytomous variables with 4 (=*k*) categories, which require 3 (=*k*-1) binary response variables. Thus, x_3, x_4, x_5 are assigned to ethnicity and x_6, x_7, x_8 to region.

[10] KN's original adjustment includes one additional covariate, household income. However, there are many missing cases for the household income in the CPS. This item will be excluded from the analysis.

of means. The sample-level response rate, 57.5%, indicates the potential for the presence of nonresponse errors.

Table 4.1. Full Sample and Unadjusted Respondent Estimates of Percentages and Means

	Full Sample		*Unadjusted Respondents*		
	Estimate	***SE***	***Estimate***	***SE***	***Deviation*** [a.]
Computer Ownership (%)	79.6	0.98	81.4	1.25	1.8*
Prior Web Experience (%)	72.0	1.09	71.2	1.45	-0.8
Unemployment (%)	3.9	0.47	4.1	0.63	0.2
Household Size	4.2	0.03	4.1	0.04	-0.1**

*p<.05 **p<.01 ***p<.001

a. $Deviation = \hat{\bar{y}}_{Unadjusted\ Respondent} - \bar{y}_{Full\ Sample}$

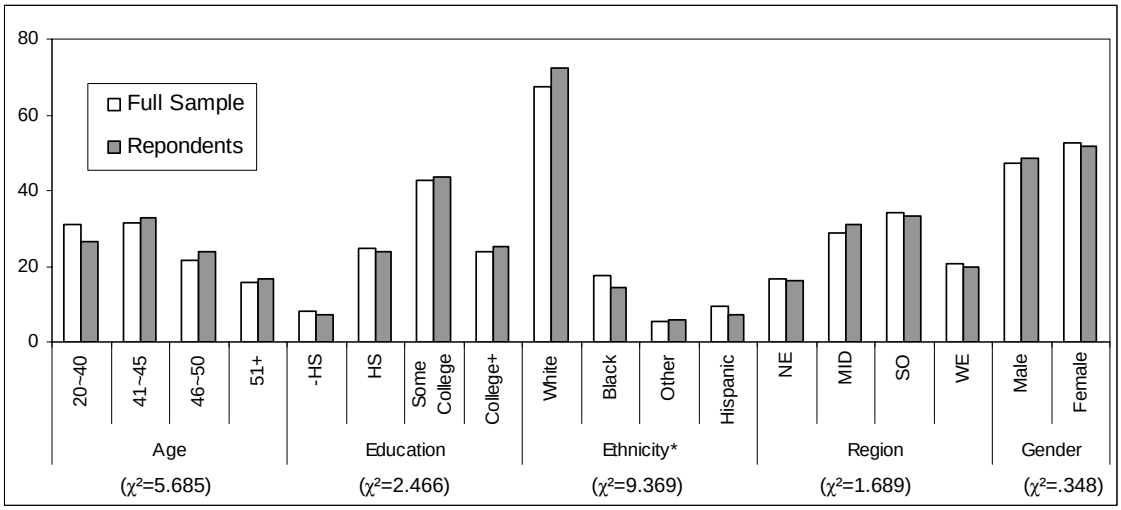
Table 4.1 compares the distribution of total population level estimates for the unadjusted respondents to those of the full sample and includes the initial deviations, $\hat{\bar{y}}_{Unadjusted\ Respondent} - \bar{y}_{Full\ Sample}$. Contrary to the initial speculation, the deviations of unadjusted statistics are surprisingly small. Since the estimates for the full sample and the respondents are not independent, variances of the deviations are calculated as follows:

$$\bar{y}_{F(Full\ Sample)} = \frac{r}{n}\hat{\bar{y}}_{R(Unadjusted\ Respondent)} + \frac{n-r}{n}\hat{\bar{y}}_{N(Unadjusted\ Nonrespondent)},$$

and, therefore,

$$\mathrm{var}\left(\hat{\bar{y}}_R - \bar{y}_F\right) = \mathrm{var}\left[\frac{n-r}{n}\left(\hat{\bar{y}}_R - \hat{\bar{y}}_N\right)\right] = \left(\frac{n-r}{n}\right)^2\left[\mathrm{var}\left(\hat{\bar{y}}_R\right) + \mathrm{var}\left(\hat{\bar{y}}_N\right)\right], \qquad (4.1)$$

where there are n units in the full sample and r respondents and $\mathrm{cov}\left(\hat{\bar{y}}_R, \hat{\bar{y}}_N\right) = 0$ is assumed. This is possible because information on nonrespondents is available from the profile data set. The deviations for computer ownership and household size, although statistically significant, do not appear meaningful.



*p<.05

Figure 4.2. Distributions of Covariates for Full Sample and Unadjusted Respondents

The distributions of the five covariates are shown in Figure 4.2. The two comparison groups are fairly identically distributed. Based on the Chi-square test for equality of distributions, only ethnicity is differently distributed. There are more Whites but fewer Blacks and Hispanics in the respondents than in the full sample, but these gaps are not large. Almost perfect comparability of the unadjusted estimates examined in Table 4.1 and Figure 4.2 may suggest that the respondents represent the full sample, i.e., the nonresponse occurs completely at random. One important implication from the identical covariate distributions is that the statistical adjustments using these covariates will not correct for any biases that may exist in variables that are not examined in this chapter, because the benchmark distributions are the same as the initial ones.

4.3.1 Sample-level Ratio-raking Adjustment

Ratio-raking adjustment is a popular modification of post-stratification which follows the iterative steps described in Deming and Stephan (1940). Unlike cell

weighting, ratio-raking controls the marginal distributions of covariates. This decreases difficulties that arise with unknown benchmarks or zero observation in cross-classified cells. The marginal counts of the five covariates from the full sample are used as benchmarks. For this study, ratio-raking was performed using WesVar™ 4.0 (Westat, 2000). Post-survey weights that adjust for sample-level nonresponse are generated and used in the estimation.

4.3.2 Multiple Imputation

Multiple imputation was first suggested by Rubin (1978) for item nonresponse. Although this chapter does not examine item nonresponse, unit nonresponse in this Web survey may be regarded as item nonresponse in some sense – there is enough background information for survey respondents and nonrespondents. Multiple imputation incorporates the frequentist concept of estimate variability evaluation into a Bayesian imputation approach.

Values for the missing observations are imputed by specifying an explicit model that produces posterior predictive distributions of the missing data, conditional on the distribution of the observed data. The models for the three dichotomous variables, y_1, y_2, y_3 are specified in the following way:

$$y_i \sim Bernoulli(\theta_i), \quad \text{logit}(\theta_i) = \alpha_i + \sum_{j=1}^{9} \beta_{ij} x_j + \varepsilon_i,$$

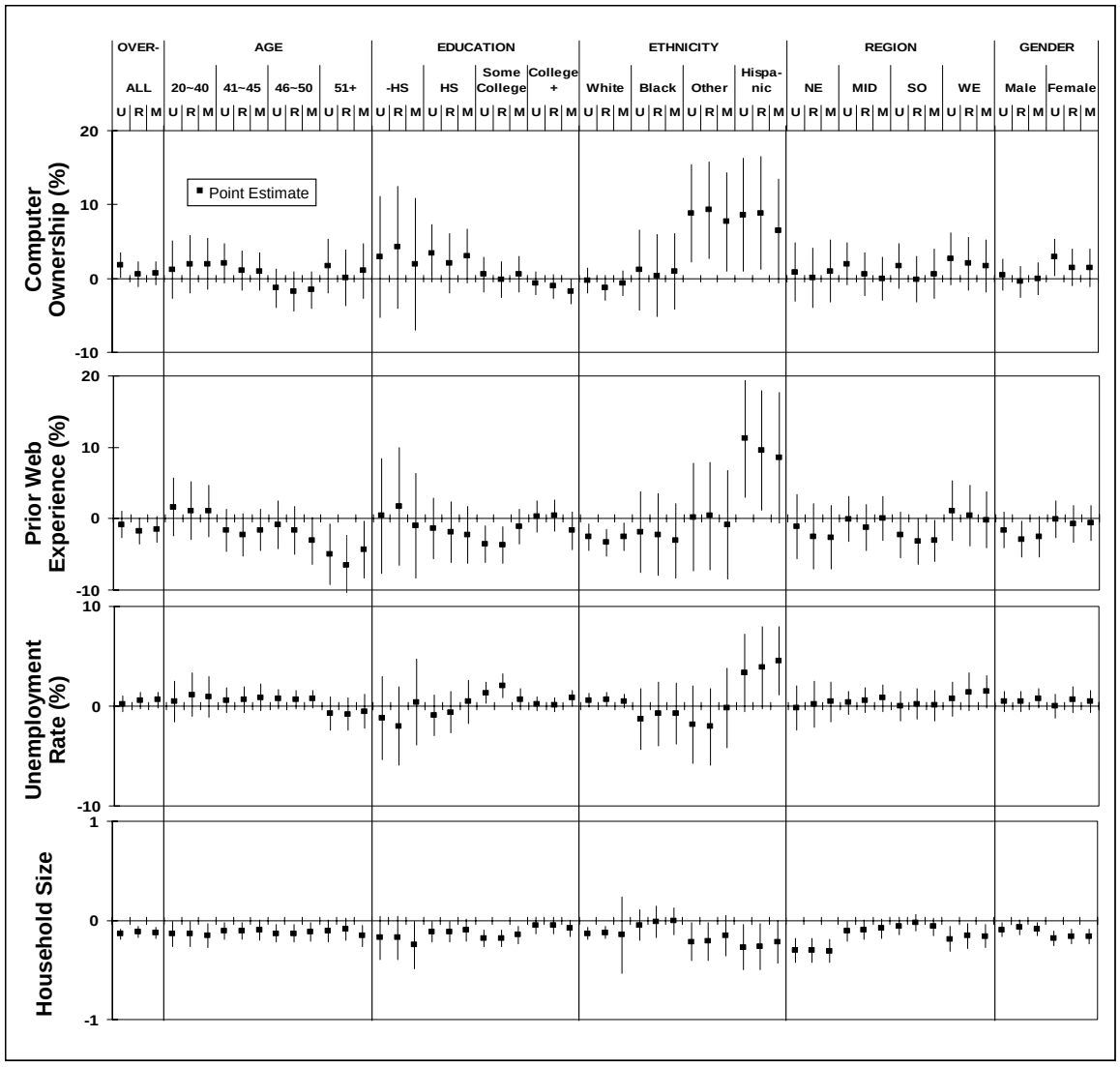
where $\alpha_i, \beta_{ij} \sim Normal(0,1)$ and ε_i's are random errors with a mean of zero for $j = 1,...,9$ and $i = 1,2,3$. Note that the same covariates are adopted here as in the ratio-raking procedure above. Since the y_i's are categorical, they are modeled as having

Bernoulli distributions determined by the parameters, θ_i. The θ_i's are predicted by the covariates known for both respondents and nonrespondents. The model parameters, α_i's and β_{ij}'s, have normal prior distributions – with mean 0 and variance 1. Similarly, the continuous variable, y_4, is modeled as follows;

$$y_4 \sim Normal(\theta_4, \upsilon),\ \theta_4 = \alpha_4 + \sum_{j=1}^{9} \beta_{4j} x_j + \varepsilon_4,$$

where θ_4 is the prior of y_4 predicted in a linear function of the same series of covariates, using prior information, $\upsilon \sim Gamma(0.5,\ 1)$, $\alpha_4, \beta_{4j} \sim Normal(0,1)$ for $j = 1,...,9$, and ε_4 a random error. Note that the model fit and modification are not considered here, because the purpose of this chapter is to compare sample-level ratio-raking adjustment and multiple imputation, given the same auxiliary information.

Winbugs 1.4 (Spiegelhalter, *et al.*, 1999) is used for the multiple imputation. The prior distributions of the model parameters are updated by the profile data. Missing values are predicted by the updated values of model parameters. Each missing value for each nonrespondent is imputed using five different initial values, which result in five different predicted values. Each model stated above is run in 10,000 iterations using the Markov Chain Monte Carlo method (details in Gelman, *et al.*, 1995, Ch.1). In order to use samples that produce convergent statistics among different initial values, the first 2,999 iterations were regarded as burn-in. For each chain, imputed values for nonrespondents are combined with observed values from respondents. The estimation and inference follows the procedure in Rubin (1987, Ch.3).



U: Unadjusted Respondents R: Ratio-raking Adjusted Respondents, M: Multiply Imputed Respondents

Figure 4.3. 95% Confidence Intervals of Deviations of Respondent Estimates from Full Sample Estimates

Figure 4.3 displays the 95% confidence intervals for the deviation of unadjusted (U), ratio-raking adjusted (R), and multiply imputed (M) estimates from the true values. Estimation for standard errors follows expression (4.1). More specifically, $\text{var}\left(\hat{\bar{y}}_N\right)$ and

$\text{var}\left(\hat{\bar{y}}_R\right)$ for the unadjusted $\hat{\bar{y}}_R$ are calculated based on the variance formula for simple random samples. For the ratio-raking adjusted $\hat{\bar{y}}_R$, $\text{var}\left(\hat{\bar{y}}_R\right)$ are obtained from WesVar™ 4.0. Variance estimation for the multiple-imputation adjusted $\hat{\bar{y}}_R$ uses procedure described in Rubin (1987). If the intervals contain zero, the deviations are not statistically significant, leading to the conclusion that the nonresponse error is negligible.

Figure 4.3 shows that most deviations are not significant both at the total population and the subgroup level. The deviation in household size appears to be statistically significant but not so much meaningful. When examined by subgroup, estimates for different racial/ethnic groups are likely to diverge the most from the true values. It is interesting to note that U, R, and M estimates are not very different from each other, especially given a sample nonresponse rate of 42.5%. In terms of deviation and variance, performance of ratio-raking and that of multiple imputation are almost equivalent. Recall that the preliminary analysis showed that the unadjusted estimates for all variables match the full sample values well. Nonresponse adjustments on these variables might have been unnecessary after all.

4.4 Coverage Error Adjustment

Coverage error in this analysis is not due solely to problems with the frame coverage per se. It also includes the combined response status from the four pre-survey stages. Unlike traditional surveys where full samples represent the target populations through sampling frames, this Web survey may not have a reliable sampling frame,

because there are multiple chances to systematically lose potentially eligible people. In other words, the frames built only on the active panel members may be biased to begin with. The population values used for comparison are calculated by applying the final weights provided in the CPS public use data and will be assumed as true values. The full sample Web survey estimates are calculated by applying the base design weights to the 1,700 cases in the full sample dataset.

Since the sample design variables are not provided in the CPS public use data and the CPS data analyzed for this study are truncated, direct calculation of the standard error for the CPS estimates is impossible. Instead, the following ad-hoc formula is used for calculating the standard error of the biases:

$$\begin{aligned} se(y_{CPS} - y_{Sample}) &= \sqrt{\text{var}(y_{CPS}) + \text{var}(y_{Sample})} \\ &= \sqrt{k\,\text{var}(y_{Sample}) + \text{var}(y_{Sample})} \\ &= \sqrt{k+1} \times se(y_{Sample}), \end{aligned} \tag{4.2}$$

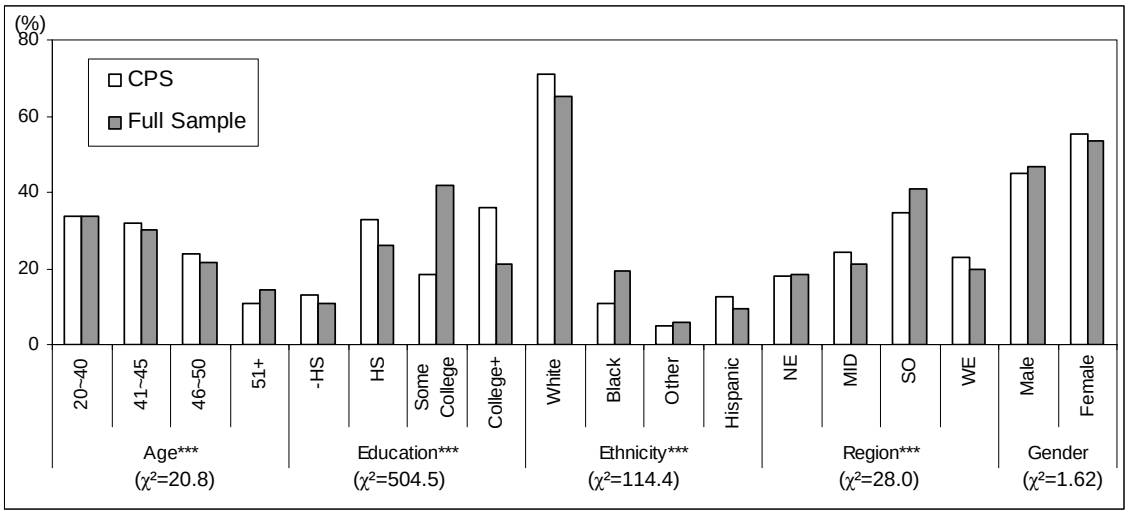
where $se(y_{Sample})$ is the standard error of the full sample estimate and k is some constant based on the ratio of the Web survey sample size to the CPS size. It should be noted that (4.2) is a crude approach to derive variance estimates because it assumes that the variability of an estimate is a function of the sample sizes.

Table 4.2. Population and Unadjusted Full Sample Estimate

	CPS		***Unadjusted Full Sample*** [a.]		
	Estimate	***SE***	***Estimate***	***SE***	***Deviation*** [b.]
Computer Ownership (%)	80.85	0.57	77.45	1.29	-3.40**
Prior Web Experience (%)	65.81	0.62	70.91	1.39	5.10***
Unemployment (%)	2.59	0.19	4.11	0.59	1.52**
Household Size	4.34	0.02	4.19	0.04	-0.15***
[a.] Design weighted full sample.			*p<.05	**p<.01	***p<.001

[b.] $Deviation = \hat{\bar{y}}_{Unadjsuted\ Full\ Sample} - \bar{y}_{CPS}$

Unlike the previous section, the comparison between the true values and the unadjusted full sample estimates suggests potential coverage problems as shown in Table 4.2. The weighted full sample estimates, when not adjusted, significantly stray from the population. This is more obvious for the computer ownership and prior Web experience. People in the frame are less likely to own computers but more likely to have Web experience. Moreover, remarkable inconsistencies in covariates can be found in Figure 4.4, especially for education and ethnicity. It becomes imperative to remedy these discrepancies.



*p<.05 **p<.01 ***p<.001

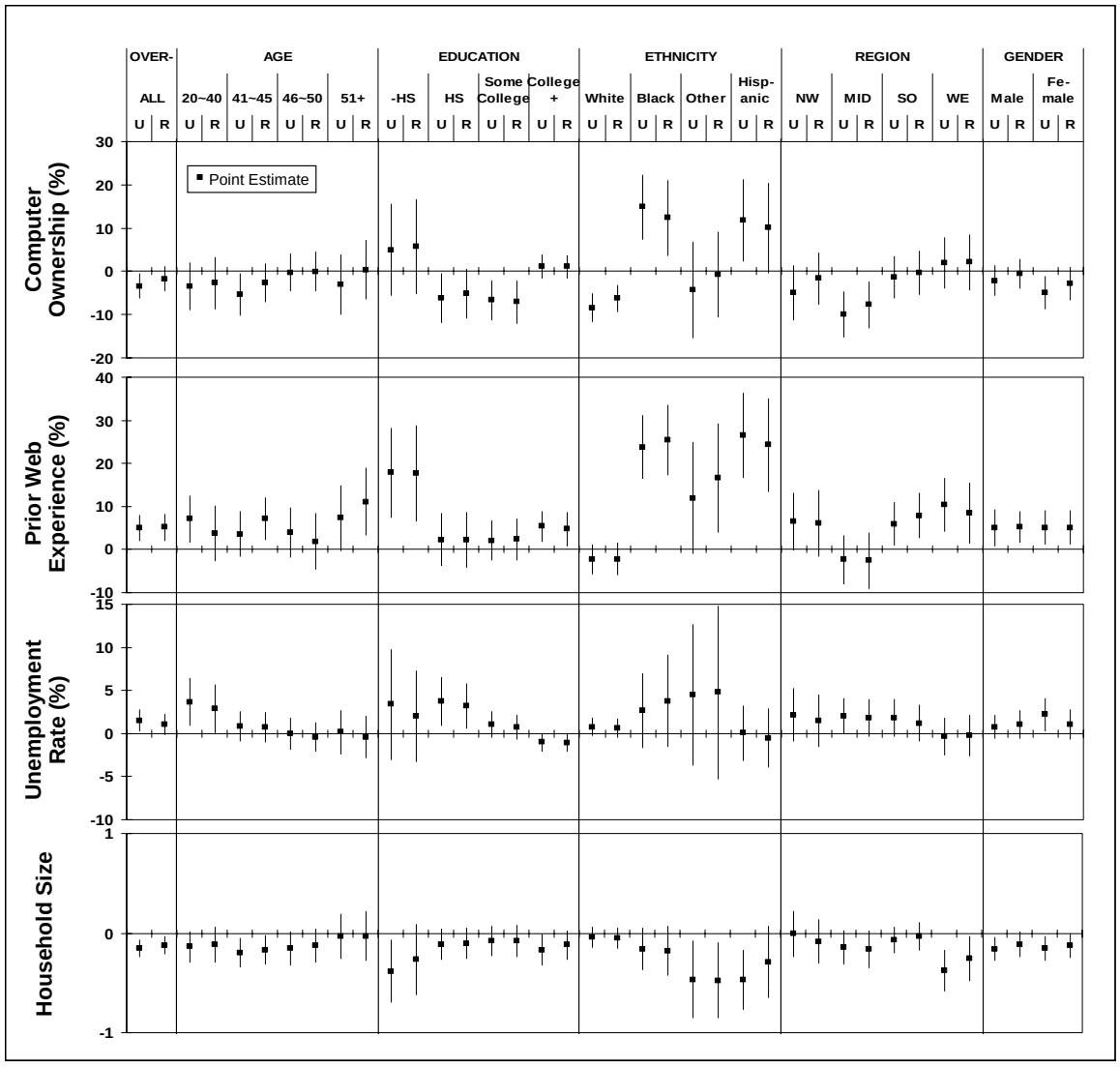
Figure 4.4. Distributions of Covariates for CPS and Unadjusted Full Sample

The coverage properties are examined by replicating the same ratio-raking procedure used in the previous section at the population level. The final adjustment weights are computed by ratio-raking the Web survey base weights to covariate marginal

counts from the CPS. The base weights are provided by KN. Both base weights and ratio-raking weights are simultaneously included in the estimation, using WesVar™ 4.0.

Imputation is not used for the evaluation of coverage error. This is because imputation is developed for data with item nonresponse. More specifically, we need to have some information about the units whose values are to be imputed. In this case, we do not have any information about the units in the target population other than ones in the full sample. Therefore, it is impossible to impute any values for the nonsampled units in the target population.

The 95% confidence intervals of the deviations of the unadjusted (U) and ratio-raking adjusted (R) estimates from the population values are shown in Figure 4.5. If the ratio-raking procedure is effective in reducing bias in estimates due to coverage error, Figure 4.5 would show confidence intervals of the deviations more likely to contain zero for the R estimates than for the U estimates. Roughly speaking, the adjustment seems to make a trivial improvement. The adjusted values are still closer to the unadjusted ones than to the population figures. Significant deviations still exist and they become more conspicuous for the subgroup estimates. Discrepancies are most prevalent for the education and ethnicity subgroups. This coincides with the divergence found in Figure 4.4. Although this divergence is supposed to be corrected by ratio-raking, estimates for subgroups formed by these covariates are still distant from the true values.



U: Unadjusted Full Sample, R: Ratio-raking Adjusted Full Sample

Figure 4.5. 95% Confidence Intervals of Deviations of Full Sample Estimates from CPS Comparison Estimates

Persons with less than a high school education report having prior Web experience at a far higher rate in the Web survey than in the CPS. In fact, its percentage in the Web sample is about 20 percentage points higher than in the CPS. One explanation may be a misunderstanding by persons in the Web sample what "Web" experience means. Another explanation may be that people with lower education in the

Web sample, before they join the KN panel, tend to own fewer computers and are more likely to be unemployed, but have had experience with the Internet at a higher rate than their counterpart in the population. These people are likely to have more time, thus, more potential opportunities to access the Internet, but are less able to afford computers because they are unemployed. This may make their reaction to obtaining free access to the Web more positively than persons with higher education, inducing them to stay active on the panel to maintain the access.

The discrepancies in computer ownership and Web usage by ethnicity warrant attention. The Web sample seems to include higher proportions of technology-savvy Blacks and Hispanics at a higher level than the CPS does. Both unadjusted and adjusted sample estimates of the computer ownership for Blacks and Hispanics are 10 percentage points higher than the population values. Equivalent racial/ethnic groups in the Web sample have higher levels of Web experience than their counterpart in the population – the full sample overestimates the Web experience by far over 20 percentage points. Interestingly, Whites in the Web sample are somehow less technologically experienced than those in the population as measured by computer ownership and Web experience. This suggests that the Web sampling frame coverage is systematically different from the population with respect to ethnicity. Ratio-raking does not seem a sufficient solution.

4.5 Discussion

This chapter is one of the first examinations of statistical adjustment approaches for Web surveys. The respondents in the particular survey studied seemed to represent

the full sample well, although the completion rate was fairly low. Consequently, the sample-level nonresponse adjustment was not even necessary for at least the variables examined in this chapter. This is similar to the recent findings about nonresponse (e.g. Curtin et al., 2000; Keeter *et al.*, 2000; Merkle and Edelman, 2002). Additionally, the covariate distributions for the respondents and the full sample were very close. This implies that adjustments based on these characteristics may not improve the estimates based on respondents much.

However, it does not seem safe to conclude the Web sample frame adequately covers the population. Estimates for the subgroups whose population and sample covariate distributions showed inconsistencies tended to deviate significantly from the population values. Traditional adjustments like raking had a limited effect in correcting for this deviation. Thus, this result failed to support the assumption of ignorability of the coverage mechanism inherent in the ratio-raking procedure.

Three points should be made about the implications of this chapter. First, they apply only to this particular type of Web survey and this particular topic. Other Web survey protocols targeting the general population are considered less scientific, as they often rely on convenience or volunteer samples, and, thus, may have completely different error structures. Second, coverage and nonresponse errors are properties of a statistic, not of a survey. Other statistics may show different nonresponse and coverage properties. Statistics in this chapter were selected because they are available at the respondent, the full sample, and the population level. Third, the target population of this chapter is very specific, parent figures with at least one teen household member. This population may have different nonresponse and coverage properties in this Web panel sample from other

populations. Findings in this chapter can serve as a window behind those error mechanisms, but cannot be generalized.

This chapter found that the coverage errors of this Web panel survey were more severe than nonresponse errors conditional on the RDD survey response. However, the full sample already includes multiple stages of nonresponse prior to the survey, which were captured under the coverage error examination in this chapter. Coverage errors from nonresponse or non-cooperation in the procedures of recruiting and maintaining panel members may be more serious than ones in the actual survey. Further investigations to statistically disentangle the coverage and nonresponse mechanisms at each stage would be informative. If consistent evidence against ignorability of the error mechanism is found, more innovative adjustment methods will be needed for sound inferences from Web survey data.

Chapter 5: Propensity Score Adjustment

5.1 Introduction

One of common methods for presenting scientific research results is group comparison. Especially in medical research reporting, it is not unusual to encounter such comparisons. For example, a report may claim that a health survey found people who consume a recommended amount of vegetables have a lower risk of cancer than people who do not. One notable fact about the comparison is that it tacitly implies a causal relationship. This report may seem reasonable prima facie, although the study design, an observational survey, does not necessarily accommodate grounds for such a finding. A closer examination may reveal that the claim relies on an assumption that sufficient vegetable consumption alone may decrease the cancer risk, whereas the control on other factors is not assured in the study.

A fundamental problem of the comparison above is that the two groups, high and low vegetable consumers, may be different with respect to not only the diet pattern but also other characteristics, such as age, gender, race, education, health status, etc. This occurs because this study uses observational data in which the assignment of the study subjects to the two groups to be compared is not guaranteed to be random. Unless the study sufficiently controls for conditions other than the experimental factor under study so that study subjects are balanced with respect to those other conditions, the difference in cancer prevalence between the groups may not be any more than an artifact.

Randomization, although desirable, is impractical, unethical or impossible in many cases. In a controlled lab experiment for the effect of vegetable consumption, randomization may be possible, but the generalization of such experimental findings may be problematic. The experiment may be unethical, when considering the study outcome may have a direct link with the cancer risk. Observational studies are the only alternative in this example, and it becomes impossible for the researcher to make one randomly assigned group of people eat more vegetables and the other eat less. The control is out of the researcher's hand, and those unrandomized conditions may lead to confounding the effect of interest with other uncontrolled effects. Now, the researcher is confined to what is available. In order to solve this problem, the researcher may use a statistical approach to control for the undesirable confounding effects.

In the context of Web surveys, the experimental treatment is translated into 'being in a Web survey' or 'having Web access.' The selection of people under this condition is assumed to be nonrandom. The control treatment is the complement but persons receiving the controls are assumed to be randomly selected from the target population. By the same statistical approach used to remedy the confounder described above, the experimental group may be adjusted to resemble the control group so that the randomness in the control group is borrowed for the Web survey group.

5.2 Treatment Effect in Observational Studies

5.2.1 Theoretical Treatment Effect

In this section, we summarize some of the considerations in estimating treatment effects based on Rosenbaum and Rubin (1983). Let the theoretical underlying treatment effect in the superpopulation, $\mathcal{U}$, be denoted as $\tau = \tau_1 - \tau_0$. The outcome under the experimental condition is τ_1, which is the mean of τ_{1i}, the outcome of all individuals in $\mathcal{U}$, where $i \in \mathcal{U}$ and $\mathcal{U}$ has $\mathcal{N}$ units. The control group outcome is τ_0, the mean of τ_{0i}. Theoretically, the treatment effect is obtainable for each unit *i* in $\mathcal{U}$ as $\tau_i = \tau_{1i} - \tau_{0i}$. The overall treatment effect is calculated over all units in $\mathcal{U}$ as $\tau = \frac{1}{\mathcal{N}} \sum_{i \in \mathcal{U}} \tau_i$.

In the finite population approach, the treatment effect is realized as t, the mean of the individual treatment effect, t_i, where the unit *i* belongs to the population, U, as $i = 1, ..., N$. Therefore, $t = t_1 - t_0 = \frac{1}{N} \sum_{i \in U} (t_{1i} - t_{0i}) = \frac{1}{N} \sum_{i \in U} t_i$. Theoretically, the treatment effect, t, is obtained when all units in the population are exposed to *both* control and experimental condition so that the realization of treatment effect for the i^{th} unit alone, $t_i = t_{1i} - t_{0i}$, is computable. In reality, whether the study is experimental or observational, only a set of sampled units from the population is examined and the study subjects are exposed to only one condition. We observe either t_{1i} or t_{0i} for the i^{th} unit, but not both. Assume that study units in an experiment come from two separate simple random samples, one under the experimental condition (s_1) with n_1 units, the other under the controlled condition (s_0) with n_0 units. From such a study, we obtain an estimate of the

treatment effect such that $\hat{t} = \hat{t}_1 - \hat{t}_0 = \frac{1}{n_1}\sum_{i \in s_1} t_{1i} - \frac{1}{n_0}\sum_{i \in s_0} t_{0i}$. Therefore, the computation of treatment effect always involves some degree of speculation about the unobserved components and unexamined population units.

Let M be a mechanism that all experimental/control treatments are repeatedly assigned to all units an infinite number of times. Under this mechanism, we may expect $E_M(t_1) = \tau_1$, $E_M(t_0) = \tau_0$, and $E_M(t_1 - t_0) = \tau$, where $E_M(\cdot)$ is the expected value over M. The mechanism M is assumed to be satisfied as $N \to \infty$. What we need is to link our sample estimates, $\hat{t}_1$ and $\hat{t}_0$, to the finite population quantities, t_1 and t_0, that approximate the underlying superpopulation figures, τ_1 and τ_0, through M. This linkage may be guaranteed under randomization of the treatment assignment distribution, denoted as π, such that $E_\pi(\hat{t}_1) = t_1$, $E_\pi(\hat{t}_0) = t_0$, and $E_\pi(\hat{t}_1 - \hat{t}_0) = t$. As long as the condition of the i^{th} unit is not dependent on that of the j^{th} unit in the same sample, implying that there is non-interference between subjects, the average treatment effect becomes

$$E_M E_\pi(\hat{t}_1 - \hat{t}_0) = \tau, \tag{5.1}$$

where $E_\pi(\cdot)$ is the expected value over the randomized assignment mechanism, π. The requirement for (5.1) is that we must be able to estimate $E_M E_\pi(\hat{t}_1)$ and $E_M E_\pi(\hat{t}_0)$ from the observed data, s_1 and s_0. Note that τ is the intended effect – not the actual effect. The actual effect may have an unintended effect arising from the imperfect or incomplete

randomization, as study units may opt to drop out from the study, cross over the assigned groups, or affect one another.

In order to estimate the average treatment effect from observed data, a stable unit treatment value assumption (SUTVA) must hold. Under SUTVA, $t_i = t_{1i}$, if $g_i = 1$ (treatment group) for all units, and $t_i = t_{0i}$, if $g_i = 0$ (control group). Thus, the outcome for the i^{th} unit can be expressed as $t_i = t_{1i} g_i + t_{0i}(1 - g_i)$, where g_i is 0 or 1. SUTVA implies that there is no interference among study subjects, meaning that potential outcomes for each unit are not related to the treatment status of other units. In addition to SUTVA, independence between the outcome and the treatment assignment is needed. When two random variables, x and y, are independent, we symbolize this by $x \perp y$. If $(t_1, t_0) \perp g$, $E_M E_\pi(\hat{t} \mid g = 1) = E_M E_\pi(\hat{t}_1 \mid g = 1) = E_M E_\pi(\hat{t}_1)$ and $E_M E_\pi(\hat{t} \mid g = 0) = E_M E_\pi(\hat{t}_0 \mid g = 0) = E_M E_\pi(\hat{t}_0)$. Thus, the estimated average treatment effect is equal to τ:

$$E_M E_\pi(\hat{t} \mid g = 1) - E_M E_\pi(\hat{t} \mid g = 0) = E_M E_\pi(\hat{t}_1) - E_M E_\pi(\hat{t}_0) = \tau . \qquad (5.2)$$

The unbiasedness in the estimation of treatment effect in (5.2) is guaranteed only under randomization with large samples.

5.2.2 Inherent Problems of Treatment Effect Estimation in Observational Studies

The unbiasedness in (5.2) does not hold in observational studies, because factors affecting the group assignment, g, are beyond researchers' control, as examined in Section 5.1. The resulting treatment effect estimates may inherently have discrepancies

between the treatment and the control group with respect to some demographic characteristics, behaviors, and/or attitudes. These attributes may confound the true treatment effect, τ.

Let us return to the example in Section 5.1 – the study on the effect of vegetable consumption on cancer risk. Suppose the researcher finds that high vegetable consumption decreases cancer risk. But he also finds that there are more females in the high vegetable consumption group and that females show a lower level of cancer than males. The question becomes whether the differentiation in cancer risk level is attributable to the amount of vegetables eaten or the gender. A sensible step to solve this dilemma is to compare the cancer risk between the groups within the same gender.

Generally speaking, lab experiments or cross-national surveys do not collect data for only one study variable. Often times, the data are analyzed for underlying relationships among variables. This means that the collected data readily contain variables that are related to the study variables – namely covariates. When the covariate means are different in two comparison groups, standard practice is to adjust for such differences when comparing means of outcome variables. Analogous to controlling for the gender effect in the example above, one can imagine adjustments on the treatment effect using auxiliary information in most studies.

5.3 Bias Adjustment Using Auxiliary Information

5.3.1 Covariates for Bias Adjustment

If the study subjects differ systematically with respect to a set of some covariates, $\mathbf{x}$, other than the assigned group characteristics, g, the realized outcome of the i^{th} unit in the treatment group and that of the j^{th} unit in the control group can be modeled as follows:

$$\begin{aligned} t_{1i} &= \tau_1 + u\left(\mathbf{x}_{1i}\right) + e_{1i} \\ t_{0j} &= \tau_0 + u\left(\mathbf{x}_{0j}\right) + e_{0j} \end{aligned}, \tag{5.3}$$

where $u(\mathbf{x})$ is a function of $\mathbf{x}$, a matrix of auxiliary variables; and e_{1j} and e_{0j} are random residuals with zero means. This implies that t_{1i}, the outcome of the i^{th} unit in the experimental treatment group, may deviate from τ_1, the true study outcome of the same group, by $u\left(\mathbf{x}_{1i}\right)$, its own distinctive characteristics, and e_{1i}, some random effect. The same is true for the individual unit outcome in the control group. The comparison of the outcomes should reflect the grouping characteristics only. Otherwise, the imbalance in the distribution of $\mathbf{x}_1$ and $\mathbf{x}_0$ confounds the comparison.

When this confounding effect of covariates is not adjusted out, the expected treatment effect becomes biased:

$$E_M\left(t_1\right) - E_M\left(t_0\right) = \tau_1 - \tau_0 + \left(\bar{u}_1 - \bar{u}_0\right) = \tau + \left(\bar{u}_1 - \bar{u}_0\right), \tag{5.4}$$

where $\bar{u}_1 = \int u(\mathbf{x})\phi_1(\mathbf{x})\,d\mathbf{x}$ and $\bar{u}_0 = \int u(\mathbf{x})\phi_0(\mathbf{x})\,d\mathbf{x}$; and $\phi_1(\mathbf{x})$ and $\phi_0(\mathbf{x})$ are the frequency functions of the covariates in the comparison groups. The expected value in (5.4) is over repeated applications of the treatments to units. Note that the expected

effect in (5.4) assumes that there is no interaction between the treatment effect and the covariates. By comparing (5.1) and (5.4), it is clear that the treatment effect is biased by $|\bar{u}_1 - \bar{u}_0|$.

This bias may be removed or reduced by balancing the covariates between the two groups. The problem of achieving the balance in estimating τ arises when $\mathbf{x}$ takes a high dimension. It is not practical to obtain equivalent distribution on many covariates, although theoretically desirable. An alternative is to summarize all covariates into one quantity and either balance or adjust based on this summary measure. Propensity score adjustment is the effective and intuitive method that serves this purpose, as it uses available covariate information and provides a scalar quantity for each unit, while requiring a minimal set of assumptions.

5.3.2 Balancing Score

For treatment effect estimation, covariates may be balanced on a function, $b(\mathbf{x})$. An appropriately constructed balancing score $b(\mathbf{x})$ has the property that the treatment assignment is conditionally independent of the covariates given $b(\mathbf{x})$. That is, the distribution of $\mathbf{x}$ conditional on $b(\mathbf{x})$ is the same for both treatment groups. It can be mathematically expressed as

$$\mathbf{x} \perp g \mid b(\mathbf{x}), \tag{5.5}$$

where $b(\mathbf{x})$ is called a balancing score as it balances out the distributional imbalance in covariates between the comparing groups. The finest balancing score is $\mathbf{x}$, the covariates themselves, but this is not practical as discussed above. While many functions of $\mathbf{x}$ can

serve as balancing scores, the propensity score, $e(\mathbf{x}) = f\{b(\mathbf{x})\}$, is frequently used. The propensity score takes the coarsest form of the balancing score. We discuss these scores in the next section.

5.3.3 Propensity Score

5.3.3.1 Bias Reduction by Propensity Scores

A propensity score is simply the probability of a unit being assigned to the treatment group ($g = 1$) given a set of covariates and is denoted as

$$e(\mathbf{x}_i) = \Pr(g_i = 1 \mid \mathbf{x}_i), \tag{5.6}$$

where $\Pr(g_1,...,g_n \mid \mathbf{x}_1,...,\mathbf{x}_n) = \prod_{i=1}^{n} e(\mathbf{x}_i)^{g_i} \{1 - e(\mathbf{x}_i)\}^{(1-g_i)}$ is assumed and $e(\mathbf{x}_i)$ is a scalar with a value between 0 and 1. Since the propensity score is a type of balancing score, the conditional independence holds as (5.5); $\mathbf{x} \perp g \mid e(\mathbf{x})$.

Returning to the earlier model in (5.3), if $u(\mathbf{x}) = e(\mathbf{x})$ and if the unit *i* from the treatment group and the unit *j* from the control group have the same propensity scores, the difference between these two units becomes confounder-free because

$$t_{1i} - t_{0j} = \tau_1 + e(\mathbf{x}_{1i}) + e_{1i} - \left[\tau_0 + e(\mathbf{x}_{0j}) + e_{0j}\right] = \tau_1 - \tau_0 + (e_{1i} - e_{0j}). \tag{5.7}$$

Omitting the subscripts, i and j, the expected value over model (5.3) is then

$E_M(t_1 - t_0) = \tau_1 - \tau_0 = \tau$, because $E_M(e_1) = E_M(e_0) = 0$. More formally, following Rosenbaum and Rubin (1983, Sec. 2.2), when a treatment and control unit have the same

propensity score, $e(\mathbf{x})$, and the treatment assignment is strongly ignorable (see Section 5.3.3.2),

$$\begin{aligned} & E_M\left(t_1 \mid e(\mathbf{x}), g=1\right)-E_M\left(t_0 \mid e(\mathbf{x}), g=0\right) \\ & \quad =E_M\left(t_1 \mid e(\mathbf{x})\right)-E_M\left(t_0 \mid e(\mathbf{x})\right) \\ & \quad =E_M\left(t_1-t_0 \mid e(\mathbf{x})\right). \end{aligned} \tag{5.8}$$

That is, the expected difference in observed responses for two units with the same $e(\mathbf{x})$ is equal to the average treatment effect at the propensity score, $e(\mathbf{x})$. When averaged over the distribution of the propensity score in the population, we have

$$\begin{aligned} & E_{e(\mathbf{x})}E_M\left(t_1 \mid e(\mathbf{x}), g=1\right)-E_{e(\mathbf{x})}E_M\left(t_0 \mid e(\mathbf{x}), g=0\right) \\ & \quad =E_{e(\mathbf{x})}E_M\left(t_1-t_0 \mid e(\mathbf{x})\right) \\ & \quad =\tau_1-\tau_0 \\ & \quad =\tau, \end{aligned} \tag{5.9}$$

since, by definition, the effect of the treatment is the average of the effects for the individuals in the population. As long as $e(\mathbf{x})$ contains all potential confounders, the adjustment based on propensity score will lead to an unbiased estimate of treatment effect in expectation.

In words, strong ignorability means that given a score, $e(\mathbf{x})$, the assignment of a unit to the treatment or control group ($g=1$ or 0) and the outcome for the unit (t_{1i} or t_{0i}) are independent. If a group of units with the same propensity score were randomly divided between the treatment and control group, (5.8) implies that we will get an unbiased estimate of the treatment effect for units that all have the same propensity score.

As discussed earlier, treatment means 'being in a Web survey' in the Web survey context. In Chapter 6, we will apply the propensity score adjustment to create groups of units with approximately the same propensity of being in a Web survey within each group. The aim is to create groups so that τ_1 for the Web sample persons equals τ_0 for the non-Web sample within each propensity score group, thus, allowing the Web sample to be used to make inference for the target population.

5.3.3.2 Assumptions in Propensity Score Adjustment

When propensity score is used to adjust for biases in observational studies, bias reduction is attainable as long as five assumptions hold. First, any propensity score should meet the strong ignorability assumption:

$$(t_1, t_0) \perp g \mid e(\mathbf{x}), \tag{5.10}$$

and $0 < \Pr(g = 1 | e(\mathbf{x})) < 1$ (Rosenbaum and Rubin, 1983; Rosenbaum, 1984a). Expression (5.10) indicates that the study outcomes, (t_0, t_1), and the assigned condition, g, are conditionally independent given the covariates in $e(\mathbf{x})$. It should be emphasized that (5.10) will hold only when the treatment assignment is *ignorable*. It is certain that this ignorability holds in randomized trials, while not necessarily in nonrandomized trials. This is why the strongly ignorable assumption is needed to develop the propensity score adjustments for nonrandomized experiments. Only under this assumption, the difference between outcomes (t_0, t_1) is unbiased for the average treatment effect, given a propensity score. Related to the strong ignorability, propensity score adjustment requires another assumption – no contamination among study units. A treatment assigned to one unit does

not affect the outcome for any other unit. Third, there should be nonzero probabilities of units being assigned to either experimental or control condition for any configuration of $\mathbf{x}$. Fourth, the observed covariates included in propensity score models represent the unobserved covariates (Rosenbaum and Rubin, 1983b), because balance is not achieved on unobserved covariates. The last assumption is that the assigned treatment does not affect covariates (Rosenbaum, 1984b).

The meanings of these assumptions must be adapted to apply to Web surveys. For example, the first assumption (strong ignorability) says that, given a propensity score, the persons in the Web survey and the persons who are not have the same means on the variables measured in the survey. This would be true on average if the persons in the Web survey with a particular propensity score were a random selection from all persons with that score. If the means are the same, then the Web sample can be used to make inferences that include the non-Web cases. In a volunteer panel, the equality of means could be violated if some important covariates used in modeling $e(\mathbf{x})$ are omitted, implying that the propensity score was not modeled correctly. The third assumption (non-zero probability of assignment) would be violated if there were certain groups of people who did not have Web access. If an important covariate, e.g., education, were omitted from the model for $e(\mathbf{x})$ and the Web sample persons and non-Web persons had different distributions of number of years of education, then assumption four would be violated.

5.3.3.3 Modeling Propensity Scores

Propensity scores have to be specified in a model and estimated from the observed data. In principle, the model for propensity score should be derived from data

for the whole population, which is not possible. However, Rubin and Thomas (1992, 1996) showed that the estimated propensity scores from sample data perform more efficiently than the true population propensity scores. A range of parametric models can be used to estimate propensity scores: logistic regression, probit model, generalized linear model, generalized additive model and classification tree model. Among them the most commonly used is logistic regression. In that case, the propensity score is modeled as:

$$\log\left[\frac{e(\mathbf{x})}{1-e(\mathbf{x})}\right] = \alpha + \mathbf{B}'f(\mathbf{x}), \tag{5.11}$$

where $f(\mathbf{x})$ is some function of covariates. There has to be enough overlap between the distributions of the propensity scores of the two comparison groups to estimate the parameters of (5.11). Otherwise, statistically reliable comparisons cannot be carried out.

Whenever covariates are used for estimation, the variable selection becomes an issue, because the predictability of the covariates in the model matters. According to Rosenbaum and Rubin (1984, p.522), $\mathbf{x}$ is required to be related to both the response and treatment assignment in order to satisfy the assumption of ignorability. Rubin and Thomas (1996) argued that there is no distinction between highly predictive covariates and weakly predictive ones in the performance of propensity score adjustment. The authors' recommendation is to include all covariates, even if they are not statistically significant, unless they are unrelated to the treatment outcomes or inappropriate for the model. In practice, however, some procedures are usually used for covariate selection. For example, a number of papers adopted stepwise regression (e.g., Rosenbaum and Rubin, 1983; Rosenbaum and Rubin, 1984; Berk and Newton, 1985; Lieberman *et al.*, 1996). Some choose one-step covariate selection based on theoretical and/or logical

relevance (e.g., Stone, et al. 1995; Duncan and Stasny, 2001). There are no clear-cut criteria for selecting variables for propensity score model building.

Drake (1993) in her simulation study showed misspecifying the model for propensity score adjustment, such as mistakenly adding a quadratic term or dropping a covariate, is not very serious. In fact, the misspecification on the propensity score model leads to only a small bias compared to the misspecification of the response model which was used to simulate the response distribution.

5.3.4 Other Adjustment Methods for Bias Reduction

So far, propensity score adjustment has been discussed as the main method for reducing selection bias in observational studies. There are other methods using covariates, and it is worthwhile to briefly examine these in comparison to propensity score adjustment (see Obenchain and Melfi, 1998 and Crown, 2001 for details).

In econometrics, Heckman (1979) proposed parametric selection bias models for bias reduction. This method has been used to evaluate the effectiveness of educational training programs in the labor market (Heckman, 1976; Heckman and Smith, 1995) Unlike a discrete variable with two levels in the propensity score adjustment, that is, $g = 1$ or 0, Heckman's selection model requires an underlying normally distributed variable, g^*, that determines treatment selection mechanism, such that, $g = 1$ when $g^* > threshold$, and $g = 0$ when $g^* < threshold$. Suppose g here defines the eligibility to a certain job training program; g^* is working hours per week; and the threshold for eligibility is 10. If a person is working more than 10 hours per week, he is automatically entitled to enroll in the program.

Another bias reduction method in econometrics incorporates instrumental variables and is known as the Rubin Causal Model. This was first outlined by Angrist, Imbens and Rubin (1996) for situations where the treatment is randomly assigned to the units, but study units comply with the assignment imperfectly, resulting in nonignorable reception of the treatment. The initial assignment here is used as an instrumental variable. The influence of the instrumental variable on the fundamental treatment outcome is assumed to go only through the actual compliance. In other words, the instrumental variable is highly correlated with the treatment receipt but not with the treatment outcome. The example for such a case is the military lottery example of the authors' article. Under a set of assumptions listed in Angrist, Imbens and Rubin (1996), the treatment effect incorporating both treatment assignment and reception identifies the average causal effect without selection bias.

Both econometric methods have not been applied extensively due to their shortcomings compared to the propensity score adjustment. More specifically, Heckman's approach uses a two-step approach to construct a variable that controls for the bias due to unobserved sources associated with treatment selection, and its sample selection models account for unobserved factors of bias only if distributional assumptions are valid. The variable that controls for selection bias should be correlated with the selected treatment but not with the treatment outcomes (Crown, 2001). The instrumental variable estimation has been criticized for strong behavioral assumptions that may not hold in reality (Heckman, 1997). Another limitation is that this method derives the causal effect only for the compliers, hence, ignores the other nonignorable components in the treatment receipt. As in Heckman's method, the instrumental variable method also

requires variables that control for selection bias should be correlated with the study variable but uncorrelated with the treatment outcomes (Crown, 2001). It is not easy to find such variables. Moreover, the two econometric model methods require less realistic distributional assumptions, are very sensitive to model specification details, and quickly become complex (see Obenchain and Melfi, 1998). These limitations lower the applicability of econometric selection methods. Thus, these are excluded from further discussion.

Outside of econometrics, Cook and Goldman (1989) compared analyses based on propensity score method to a multivariate confounder score method in epidemiological unrandomized research. The authors found that propensity score method is less affected by the high correlation between the treatment (or exposure) level and the confounders than the multivariate confounder score.

5.4 Methods for Applying Propensity Score Adjustment

Three application methods of propensity score adjustment are identified from the literature. The first approach matches two units based on the propensity score – one from the treatment group and the other from the control group, and forms a pair. The group comparison is done within a given propensity score, and the average treatment effect is calculated over all matching propensity scores. Subclassification is the second application method. From a combined pool of subjects from both conditions, units are stratified based on the propensity score so that $e(\mathbf{x})$ are approximately constant for all units in each stratum. The expected difference between the two assignments at a given

propensity score is equal to the average treatment effect. In the third application method, propensity scores are applied by adjusting covariance in a linear response model. The detailed operationalization of the three methods will be discussed below (see Rosenbaum and Rubin, 1983, 1984 and D'Agostino, 1998 for a review).

5.4.1 Matching by Propensity Scores

Matching is a natural approach to bias reduction when the cost of experimentation is high and when a large reservoir of units under control condition is available. In fact, most methodological studies of propensity scores application are concentrated on matching, especially pair matching. This may be because the propensity score adjustment is originated from causal inference studies, where only a small portion of the population is exposed to the experimental condition, which makes the size of the control group much larger than that of the treatment group. The basic idea in matching is compare all experiment treated units only with controlled units whose covariates show similar distributions.

The illustration of matching is first carried out in terms of univariate covariate x as in Rubin (1973). Suppose that there is a random sample of size n from some population of a treatment group ($g=1$) P_1 and denote the sample as S_1; and a larger random sample of size $m=kn$ with $k \geq 1$ from a control group ($g=0$)population P_0, denoted as S_0. It is further assumed that x is recorded for all subjects in S_1 and S_0. All subjects in S_1 are to be matched to their counterparts selected from S_0. Based on x, a subsample of size *n* is drawn from S_0, denoted by S_0' such that each unit in S_0' has an equivalent value of x to a certain unit in S_1. The treatment effect is estimated from S_1

and S_0'. If $k=1$, a purposeful matching is not attainable, as S_0' is in essence a random sample of P_0. In this case, the bias due to imbalanced x is retained. If $k \to \infty$, a perfect matching between S_1 and S_0' is highly feasible – the bias may be reduced, if not removed.

Rubin (1973) documented three simple approaches of constructing S_0' for pair matching. They all assign $\{s_{1i}\} \in S_1$ with $i=1,...,n$ the closest match from the unmatched units of $\{s_{0j}\} \in S_0$ with $j=1,.....,m$ with $m=kn$ and $k \geq 1$ base on x. The selection mechanism of S_0' is completely defined by how the order of $\{s_{0j}\}$ is specified: 1) random ordering (units are randomly ordered); 2) low-high ordering (a unit not yet matched with the lowest *x* score is matched next); and 3) high-low ordering (a unit not yet matched with the highest *x* score is matched next). All three methods show similar bias reduction patterns. Unless the ratio of the treatment group variance to that of the control group for the matching variable is larger than 1, all three ordering methods attain sizable bias reduction (Rubin, 1973).

So far, matching has been examined in terms of a univariate x. In practice, matching is done in a multivariate fashion, because the exact matching on all covariates is impossible. Instead, matching is carried out by using propensity score $\hat{e}(\mathbf{x})$ from (5.6) in order to equalize all covariate distributions between the treatment and the control group. The matching methods and bias reduction patterns examined for univariate x should apply similarly when using $\hat{e}(\mathbf{x})$.

Rosenbaum and Rubin (1985) compared three multivariate matching methods using propensity scores. Each of these methods has similarities to the nearest neighbor hot-deck method used for imputation in sample surveys (Little and Rubin, 2002, p.69). They differ by the level of importance given to the estimated propensity score relative to the other auxiliary variables in $\mathbf{x}$. Under the first method, nearest available matching, the first subject in randomly ordered S_1 is matched with the subject in S_0 having the nearest $\hat{e}(\mathbf{x})$. Both subjects are removed from the lists and the same matching procedure continues for the remaining unmatched subjects in S_1. The remaining two matching methods rely on the Mahalanobis metric quantity using all auxiliary variables and propensity scores, calculated from the Mahalanobis distance function: $d(\mathbf{u},\mathbf{v}) = (\mathbf{u}-\mathbf{v})' C_{S_0}^{-1} (\mathbf{u}-\mathbf{v})$, where $\mathbf{u}$ and $\mathbf{v}$ are values of $\{\mathbf{x}', \hat{e}(\mathbf{x}')\}$; and $C_{S_0}^{-1}$ is the sample covariance matrix of $\{\mathbf{x}', \hat{e}(\mathbf{x}')\}$ in S_0. The second method uses nearest available Mahalanobis metric matching. Units are matched as in the first method but with respect to Mahalanobis distance quantity. The third approach is nearest available Mahalanobis metric matching within calipers. For a unit in the randomly ordered S_1, create a subset of S_0 with all available subjects whose $\hat{e}(\mathbf{x})$ is within the range of a specified constant. This specified range is the caliper. Then, find the subject in the subset of S_0 that has the closest match to the unit in S_1 with respect to the Mahalanobis distance. Rosenbaum and Rubin (1985) demonstrated that the third method is superior to the other two with respect to the balance in covariates and in propensity scores. This is a reasonable finding, since nearest available Mahalanobis metric matching within calipers uses all covariates and propensity scores and takes advantage of the first two matching methods.

There is an issue with the degree of closeness of a matched pair. Rosenbaum and Rubin (1985b) compared inexact matching (failure to match on the exact covariate score) with incomplete matching (failure to match all units in the treatment group). The study showed that incomplete matching has a higher likelihood of retaining severe bias in the treatment effect than inexact matching. The authors recommended using an appropriate nearest multivariate matching to complete the matching, even if this may leave some residual due to inexact matching.

Pair marching of a Web sample to the nonsampled part of a population has limited relevance to finite population estimation. While matching of Web respondents to nonresponding or nonsampled cases from a larger pool constructed based on randomization might be feasible, the interest is in estimating population means, totals, and other population quantities. No data, other than covariates, are available for the nonsampled units. Also, estimating the difference between the Web sample and nonsample quantities is not possible nor is it of interest.

5.4.2 Subclassification by Propensity Scores

All units in the treatment and control groups may be combined into one and partitioned into a number of subclasses based on the covariate distributions such that each subclass has a restricted range of covariate values. This idea was first presented in Cochran (1968) with the underlying rationale that the units within one subclass become comparable with respect to the covariates. The major advantage of subclassification is that the treatment effect can be adjusted by restructuring subclass weights based on the covariate distribution without assumptions about response surface modeling. Propensity score adjustment using the subclassification method appears frequently in clinical trials

(e.g., Rubin and Rosenbaum, 1984; Hoffer *et al.*, 1985; Lavori and Keller, 1988; Cook and Goldman, 1989; Czajka *et al.*, 1992; Stone *et al.*, 1995; Lieberman *et al.*, 1996; Rubin, 1997; Benjamin, 2001). Its popularity is not surprising, when considering (1) that subclassification allows easier operation than matching, (2) that the number of the control group units need not be larger than that of the treatment group, and (3) that subclassification uses all study subjects, unlike matching where unmatched units in the control group are discarded.

Returning to the initial demonstration of (5.2), suppose that there is a univariate x available in the data. All units from both conditions first need to be sorted by x in order to use subclassification. Let the boundaries of x be x_{c-1} and x_c for the c^{th} subclass; and the sample means of study outcome of the c^{th} subclass for the two conditions be t_{1c} and t_{0c}. The expected outcomes from the experimental and control groups are

$$\begin{aligned} E_M E_\pi\left(t_{1c}\right) &= \tau_1 + \bar{u}_{1c} \\ E_M E_\pi\left(t_{0c}\right) &= \tau_0 + \bar{u}_{0c} \end{aligned}, \tag{5.12}$$

where $\bar{u}_{1c} = \dfrac{\int_{x_{c-1}}^{x_c} u(x)\phi_1(x)dx}{\int_{x_{c-1}}^{x_c} \phi_1(x)dx}$, $\bar{u}_{0c} = \dfrac{\int_{x_{c-1}}^{x_c} u(x)\phi_0(x)dx}{\int_{x_{c-1}}^{x_c} \phi_0(x)dx}$, and $u(x)$ is a function of the covariates; and c is an indicator of the subclass with $c = 1,...,C$. It becomes clear from (5.12) that the initial bias of the average treatment effect is $\bar{u}_0 - \bar{u}_1 = \sum_{c=1}^{C}\left(\bar{u}_{1c} - \bar{u}_{0c}\right)$, i.e., the cumulative difference in the covariate means.

All units within one subclass are comparable with respect to the covariates included in the propensity score model. By allocating appropriate weights, the overall

treatment effect can be adjusted. The treatment effect is the weighted mean of the differences between the experimental and control group units given a subclass. The weight for each subclass is derived, for example, based on the proportion of each subclass in the experimental or control group. After adjusting for the distributional differences in x, the remaining bias is

$$\sum_{c=1}^{C} w_c \left(\overline{u}_{1c} - \overline{u}_{0c} \right), \tag{5.13}$$

where w_c is a weight assigned to the c^{th} subclass. The proportion of the initial bias reduced by the adjustment is, therefore,

$$\theta = 100 \times \left(1 - \frac{\sum_{c=1}^{C} w_c \left(\overline{u}_{1c} - \overline{u}_{0c} \right)}{\left(\overline{u}_1 - \overline{u}_0 \right)} \right). \tag{5.14}$$

Five implications may be drawn from (5.12) and (5.14) – the bias reduction in subclassification adjustment depends on (1) the function of the covariate, $u(x)$; (2) the shape of frequency functions, $\phi_0(x)$ and $\phi_1(x)$; (3) the number of subclasses, c; (4) the division points, x_i; and (5) the choice of weights.

In practice, more than one covariate is likely to be used for reducing bias. Subclassification based on multiple collateral variables is not easy to carry out, because the number of subclasses increases exponentially with an increase in the number of covariates and/or their categories. This may lead to a number of subclasses with zero observation. Using propensity scores instead of multiple covariates becomes a sensible choice as they represent all covariates included in the model approximately.

The same procedure in subclassification on a univariate x holds for the subclassification using estimated propensity scores (see Rosenbaum and Rubin, 1984, for the full illustration). It is possible to create numerous subclasses so as to refine units in each subclass to have almost identical propensity scores. Cochran (1968) found that five subclasses are often sufficient to remove over 90% of the bias and that having more than five subclasses does not add much bias reduction. It seems to be a norm to adopt five subclasses, more specifically quintiles of the propensity scores, in existing literature (e.g., Rosenbaum and Rubin, 1984; Terhanian et al. 2000a). This seems more reasonable as one would want to create subclasses of approximately the same size. Each subclass needs to have at least one unit from both conditions to meet the assumption of assignment ignorability. In addition, there should be enough observations from both conditions within each subclass in order to derive less volatile weights.

Propensity score adjustment by subclassification examined above resembles post-stratification. According to Kish (1965, p.92), the initial error in any sample survey estimate is $\sum_{h=1}^{H} w_h y_h - \sum_{h=1}^{H} W_h Y_h$, where there are H strata in the population; y_h and Y_h are the sample estimate and the population quantity for the h^{th} stratum; and w_h and W_h are proportion of the h^{th} stratum in the sample and population. What post-stratification aims to achieve is to obtain w_h that is close to W_h, i.e., $w_h \approx W_h$, so that the error becomes

$$\sum_{h=1}^{H} W_h \left(y_h - Y_h \right). \tag{5.15}$$

From (5.15), it is clear that the magnitude of error is related not only to the choice of the weight, w_h , but also to the difference between the sample estimate and the

population quantity. When the sample selection is random, $(y_h - Y_h)$ is free from bias. However, when not, much bias is retained (Kish, 1965). That is, even if $w_h = W_h$, the error can be sizable because y_h is not guaranteed to approximate Y_h. Relating this to propensity score subclassification, we are trying to derive weights in (5.13), w_c, that minimize the quantity $\sum_{c=1}^{C} w_c(\bar{u}_{1c} - \bar{u}_{0c})$. However, if the covariate included in the function, $u(x)$, is different for the treatment and control group, we may not expect a substantial reduction in the overall bias. The only difference from post-stratification is that the weights are calculated by adopting explicit models.

Therefore, subclassification based on propensity scores may be regarded as model-based post-stratification. The former is more efficient than the latter, as it incorporates multi-dimensional covariates without concerns about the convergence and allows explicit modeling for adjustment.

In sample surveys, classes constructed based on propensity scores may be included in the post-survey adjustment, such as calibration adjustment. The requirements in this case are that the population or reference data must be available, unlike the marginal or cell counts in the traditional adjustments, and contain all variables included in the propensity model. While this may serve as an alternative, applying this strategy may be limited due to the requirements in the reference data.

5.4.3 Covariance/Regression Adjustment by Propensity Scores

The bias in treatment effect may be reduced by regression adjustment using propensity scores. Under this method, the expected value of the responses are modeled as

$$\begin{aligned} E_M E_\pi (t_1) &= \tau_1 + \beta_1 x_1 \\ E_M E_\pi (t_0) &= \tau_0 + \beta_0 x_0 \end{aligned}, \tag{5.16}$$

and the expected treatment effect is

$$E(r_1 - r_0) = \tau_1 - \tau_0 + \beta_1 x_1 - \beta_0 x_0 = \tau + \beta(x_1 - x_0), \tag{5.17}$$

when the response surfaces in both conditions are parallel ($\beta_1 = \beta_0$). The bias in treatment effect is $\beta(x_1 - x_0)$, which may be removed when $x_1 = x_0$.

When multiple covariates are used, propensity scores provide a convenient alternative. It only requires finding the regression of the responses on the propensity scores in the treatment and control groups and uses the regression for treatment effect estimation. If propensity scores are used in (5.16) instead of x, the expected treatment effect (5.17) becomes bias-free given a propensity score as

$$E(r_1 - r_0 \mid e(x)) = \tau_1 - \tau_0 + \beta_1 e(x) - \beta_0 e(x) = \tau, \tag{5.18}$$

where the response surfaces of the two groups are parallel, i.e., $\beta_1 = \beta_0$.

What is the difference between removing bias using propensity scores in the regression and performing regression adjustment directly on the responses using all covariates? Rosenbaum and Rubin (1984) illustrated that the 'point estimate of treatment effect from an analysis of covariance adjustment for $\mathbf{x}$… is equal to the estimate

obtained from a univariate covariance adjustment for the sample linear discriminant based on $\mathbf{x}$, whenever the same sample covariance matrix is used for both the covariance adjustment and the discriminant analysis.' D'Agostino (1998) noted that the propensity adjustment is more convenient. Fitting complicated propensity score models is not as difficult as fitting complicated response surface models, because the goal of propensity score modeling is to get good estimates of the probability of receiving a certain treatment, not to obtain parsimony.

The covariance adjustment using propensity scores is not as widely applied in the literature as subclassification or matching due to two reasons. First, there is a restriction imposed on the response surfaces. As in (5.17) and (5.18), the response surfaces in the two group assignments should be parallel, which may be difficult to verify. Second, there are many cases where a regression adjustment performs poorly and increases biases. When the linear discriminant in response surfaces is not a monotone function of the propensity score (i.e. the covariance matrices in the experimental and control groups are unequal), the covariance adjustment may seriously increase the expected squared bias, because it implicitly adjusts for a poor approximation to the propensity score (Rubin, 1979). For the nonlinear response surfaces, univariate covariance adjustment can either increase the bias or overcorrect bias if the variances of x in the two conditions differ (Rubin, 1973). Therefore, unless the requirements are well met and the linear discriminant is highly correlated with the propensity score, matching or subclassification may serve the bias reduction better.

The theoretical underpinnings of propensity score adjustment and its application methods have been examined in this chapter. Propensity score adjustment may serve as a

potential post-hoc adjustment method for bias reduction, when the sample selection mechanism is not guaranteed to follow a random fashion. In order to utilize propensity score adjustment legitimately, the five assumptions examined previously need to hold. They are strong ignorability, no contamination among study units, nonzero probability of being assigned to both experimental and control conditions, the observed covariates' representativeness of the unobserved covariates, and no effect of assigned treatment on the covariates. It should be noted that propensity score adjustment achieves the balance on covariates averaging over repeated studies. This implies that not all studies using propensity score adjustment necessarily achieve the balance.

Chapter 6: Alternative Adjustments for Volunteer Panel Web Survey Data

The focus in this and later chapters will be on the estimation of population means and totals. To do this, we will determine alternative sets of weights, $\{w_i\}_{i=1}^{n}$, that (1) adjust for imbalance in the distribution of covariates between the Web survey sample and a reference survey data set and (2) use auxiliary data to produce weights that are properly scaled for estimating totals in addition to means. The first purpose will be served by the use of adjustment subclasses formed on the basis of propensity scores as described in Section 6.1 and 6.2. Auxiliary or covariate data will be used to further adjust weights in calibration estimation, discussed in Section 6.3. Both the propensity score and calibration adjustments are mainly intended to reduce biases caused by nonrandom sample selection and deficient coverage in Web surveys.

6.1 Problems in Volunteer Panel Web Surveys

Volunteer panel Web surveys are conducted among a set of people who have Web access and self-select to join the panel. The overall survey protocol is described in Section 2.1 and is depicted in Figure 6.1. As the colors in this figure suggest, people under each step are not guaranteed to resemble one another. Therefore, the relationships between steps are not necessarily known. The greatest threat to a Web survey is the uncertain and incomplete coverage of the frame, because one must have Web access and voluntarily join the panel in order to be eligible for the survey participation. Unless the population of interest is the volunteer panel itself, the protocol in Figure 6.1 does not

allow construction of frames with known coverage of the population of interest. This problem with coverage yields the next problem that it is impossible to draw samples from the full population with known probabilities and to assign selection weights to the sample units in the ways normally done in sample surveys. Moreover, poor response rates of this type of Web survey leave more room for survey errors.

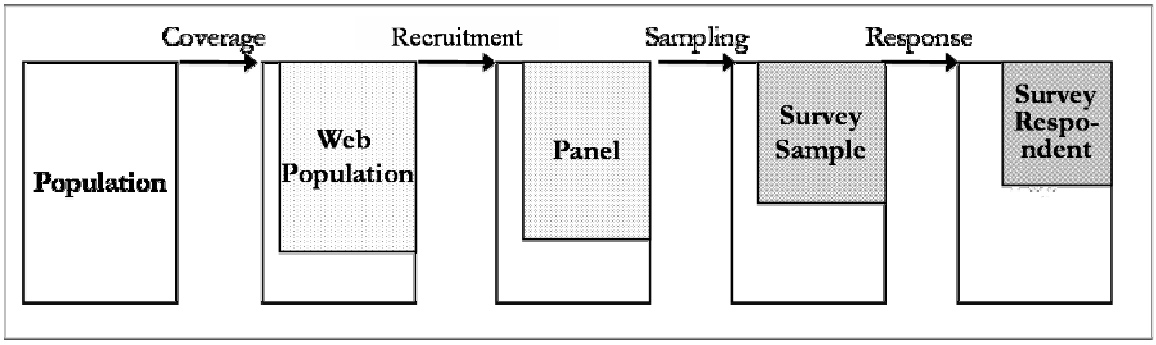


Figure 6.1. Volunteer Panel Web Survey Protocol

Estimates from this type of Web survey may suffer from a combination of noncoverage, nonprobability sampling, and nonresponse. Chapter 4 examined whether some parts of these errors may be corrected by traditional adjustment methods. The finding indicated the limitations of the traditional methods and the needs for more innovative adjustments. When integrating the causal inference views in Chapter 5, these problems may be summarized into one simple term – selection bias. The respondents' self-selection mechanism from one step to the next in Figure 6.1 is not guaranteed to be random, which causes biased survey estimates. As in Chapter 5, propensity score adjustment may be adopted as a post-hoc remedy to diminish the bias. In this case, we will model the propensity of being in the responding Web sample.

Ideally, propensity score adjustment is not necessary in survey data analysis to correct initial selection bias, as most surveys rely on randomized sample selection. Samples are assumed to represent the characteristics of the desired population. In theory, survey estimates are expected to be design unbiased or consistent estimates of the population quantities. On the other hand, propensity score adjustment is not novel in survey statistics, especially in post-survey adjustment. It has been used to derive adjustment weights for reducing biases in survey estimates arising from coverage problems (e.g., Duncan and Stasny, 2001), late response (e.g., Czajka *et al.*, 1992) and nonresponse (e.g., Smith *et al.*, 2000; Vartivarian and Little, 2003).

Focusing on volunteer panel Web surveys, this chapter will propose a two-stage adjustment method in combination with the survey protocols in Figure 6.1. The first stage adjustment will be examined in Section 6.2, which will provide a detailed mathematical presentation of how to use propensity score adjustment for the volunteer panel Web surveys. The adjustment will require a reference survey that is conducted parallel to the Web survey (Terhanian and Bremer, 2000). The reference survey is required to have more desirable coverage and sampling properties and higher response rates than the Web survey. For instance, the reference survey may be conducted using traditional survey modes, such as random digit dialing telephone method in Harris Interactive's case. As Figure 6.2, the reference survey data are used as a source for benchmarks for the first-stage adjustment. This benchmarking is carried out via propensity score adjustment as it balances the covariate distribution between the Web and reference survey samples. A reference survey needs to collect only the covariate information needed to compute propensity scores. Through this method described in

detail in Section 6.2, it is hoped to use the strength of the reference survey and reduce biases in the Web survey estimates. However, it should be noted that the employment of the reference survey implicitly disregards the dissimilar measurement properties due to mode difference between the Web and reference surveys.

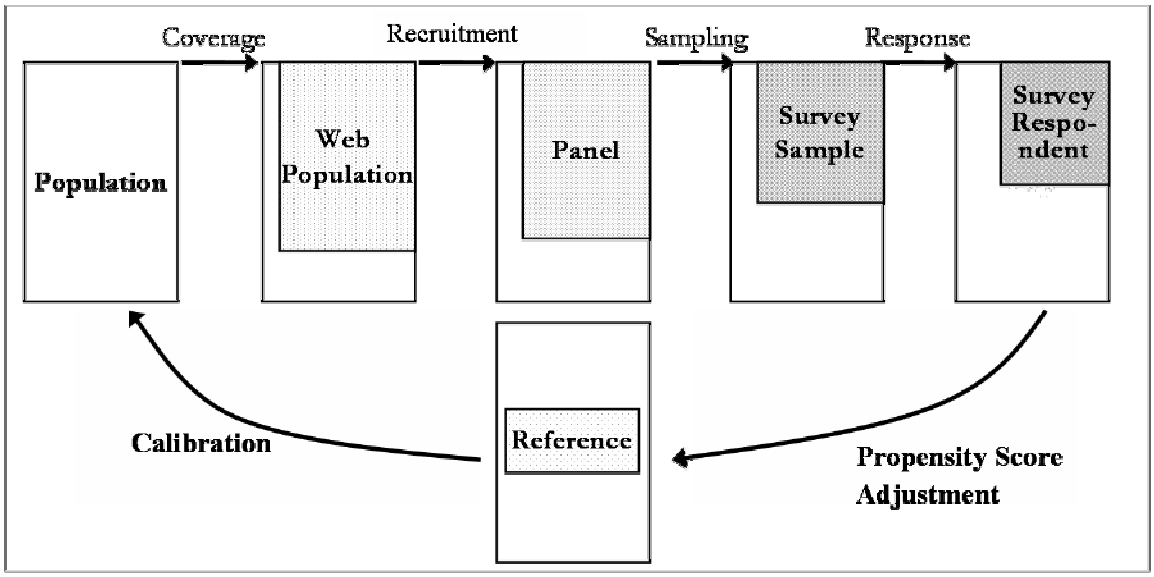


Figure 6.2. Proposed Adjustment Procedure for Volunteer Panel Web Surveys

Section 3 will introduce calibration adjustment as the second stage adjustment. The remaining disparities in covariates between the population and propensity-score-adjusted Web sample are expected to be tuned by adding another layer of weights by calibration adjustment.

Section 4 will summarize the combination of the propensity score and calibration adjustments and provide a theoretical illustration on how the bias properties are modified through the course of adjustment application.

6.2 Adjustment to the Reference Survey Sample: Propensity Score Adjustment

Subclassification is most applicable and practical for Web survey situations among the three application methods of the propensity score adjustment examined in Chapter 5. Although pair matching is a possibility when comparing treatment and control groups, how to apply the method to estimate finite population quantities is unclear, as discussed in Section 5.4.1. Therefore, we do not regard pair matching method as feasible. As noted earlier, a larger reservoir of control units is needed for pair matching in the analysis of quasi-experimental designs. If a larger reference survey is conducted in a traditional mode parallel to a Web survey only to acquire covariate information, it will be more logical to discard the Web survey and collect information on all variables in the reference survey. However, a large reference survey, like the Current Population Survey, conducted by an established survey organization with high coverage and response rates, can be quite useful. Regression adjustment using propensity scores is possible, but the restrictions associated with building response models make this approach less appealing. The requirements for the response surface examined in Section 5.4.3 are difficult to be achieved. Therefore, the subsequent discussion on the application of propensity score adjustment will be focused on subclassification.

Suppose that there are two samples – a volunteer panel Web survey sample (s^W) with n^W units each with a base weight of d_j^W, where $j = 1,...,n^W$; and a reference survey sample (s^R) with n^R units each with a base weight of d_k^R, where $k = 1,...,n^R$. Note that these base weights will not be inverses of selection probabilities, since the volunteers are not obtained by probability sampling. First, the two samples are combined into one,

$s = \left(s^W \cup s^R \right)$ with $n = n^W + n^R$ units. We need to calculate the propensity score from the combined sample, s. The propensity score of the i^{th} unit, where $i = 1,...,n$, is the likelihood of the unit participating in the volunteer panel Web survey rather than the reference survey and is defined as $e\left(\mathbf{x}_i\right) = \Pr\left(i \in s^W \mid \mathbf{x}_i, i = 1,...,n\right)$. The propensity score is estimated in a logistic regression as in (5.11) using covariates collected in both the Web and reference surveys ($\mathbf{x}_{obs}$). If all relevant covariates are included in both surveys, then $\mathbf{x}_i = \mathbf{x}_{obs,i}$ for each unit *i*. A critical assumption in doing this is that the combined sample can legitimately be used to estimate the probability of being in the volunteer panel. Given a set of covariate values, a person must have some nonzero probability of being in the Web survey or not, and that probability must be estimable from the combined sample, s.

Based on the predicted propensity score, $\hat{e}(\mathbf{x})$, the distribution of the Web sample units is rearranged so that s^W resembles s^R in terms of $\mathbf{x}_{obs}$ included in the propensity model. Mechanically, this is first done by sorting the combined data (s) by the predicted propensity score of each unit and partitioning s into *C* subclasses, where each subclass has about the same number of units. Based on Cochran (1968), the conventional choice in practice is to use five subclasses based on quintile points. Ideally, all units in a given subclass will have about the same propensity score or, at least, the range of scores in each class is fairly narrow. This is so that (5.8) and (5.9) will apply approximately. In the c^{th} subclass in the merged data denoted as s_c, there are $n_c = n_c^W + n_c^R$ units, where n_c^W is the number of units from the Web survey data, and n_c^R from the reference survey. The total number of units in the merged data remains the same because

$$\sum_{c=1}^{C}\left(n_c^W + n_c^R\right) = \sum_{c=1}^{C} n_c = n\,.$$

Second, we compute the following adjustment weights to all units in n_c^W, the c^{th} subclass of the Web survey data:

$$f_c = \frac{\sum_{k\in\left(s_c^R\right)} d_k^R \Big/ \sum_{k\in\left(s^R\right)} d_k^R}{\sum_{j\in\left(s_c^W\right)} d_j^W \Big/ \sum_{j\in\left(s^W\right)} d_j^W}, \tag{6.1}$$

where s_c^R and s_c^W are the sets of units in the reference sample and Web sample of the c^{th} subclass. If the weights in (6.1) are the inverse of selection probability, it can be expanded to:

$$f_c = \frac{\sum_{k\in\left(s_c^R\right)} d_k^R \Big/ \sum_{k\in\left(s^R\right)} d_k^R}{\sum_{j\in\left(s_c^W\right)} d_j^W \Big/ \sum_{j\in\left(s^W\right)} d_j^W} \equiv \frac{\hat{N}_c^R \big/ \hat{N}^R}{\hat{N}_c^W \big/ \hat{N}^W}\,.$$

The adjusted weight for unit j in class c in the Web sample is

$$d_j^{W.PSA} = f_c d_j^W = \frac{\hat{N}_c^R \big/ \hat{N}^R}{\hat{N}_c^W \big/ \hat{N}^W} d_j^W\,. \tag{6.2}$$

When the base weights are equal for all units or are not available, one may use an alternative adjustment weight as follows:

$$f_c = \frac{n_c^R \big/ n^R}{n_c^W \big/ n^W}\,. \tag{6.3}$$

The adjustment using (6.3) does not allow populations totals to be estimated since the weights are not appropriately scaled, unless the population sizes for both the reference survey and Web survey are known.

From (6.2), we can see that $\sum_{j \in (s_c^W)} f_c = n_c^W f_c = n^W \frac{n_c^R}{n^R}$. The weights from (6.1) or (6.3) may make the distribution of the Web survey sample equal to the reference survey sample in terms of propensity scores. For example, the estimated number of units in class c from the Web sample using the adjusted weights is

$$\begin{aligned} \hat{N}_c^{W.PSA} &= \sum_{j \in (s_c^W)} d_j^{W.PSA} \\ &= \hat{N}^W \frac{\hat{N}_c^R}{\hat{N}^R} \quad . \end{aligned}$$

In words, the estimated number of units from the Web survey, $\hat{N}^W$, is distributed among the classes according to the distribution from the reference survey, $\hat{N}_c^R / \hat{N}^R$.

The estimator for the mean of a study variable, *y*, for the Web survey sample (s^W) becomes

$$\hat{\bar{y}}^{W.PSA} = \frac{\sum_c \sum_{j \in (s_c^W)} d_j^{W.PSA} y_j}{\sum_c \sum_{j \in (s_c^W)} d_j^{W.PSA}}.$$

Note that the reference sample units are not used in deriving $\hat{\bar{y}}^{W.PSA}$ after adjustment weights, $d_j^{W.PSA}$, are assigned. Therefore, the reference sample is required to have only the covariate data, not necessarily the variables of interest. The algorithm for propensity score adjustment is computationally implemented in **`psa.fcn`** using **R**© (Venables *et al.*, 2003) shown in Appendix 1.1 (part of the code comes from Obenchain, 1999).

The set of covariates typically includes similar kinds of demographic variables to those used in post-stratification. Harris Interactive includes both demographic and nondemographic variables in the propensity models (e.g., Terhanian *et al.*, 2000; Taylor *et al.*, 2001). The importance of covariates in propensity score adjustment should be understood in relation to the substantive study variable, y, and the group assignment variable, g (Rosenbaum and Rubin, 1984). How important it is to include nondemographic variables in propensity score adjustment for Web surveys is unclear due to two facts: (1) the inclusion of more variables automatically increases the predictive power of the model and (2) the nondemographic, e.g., attitudinal, covariates can often be explained by demographic variables.

6.3 Adjustment to the Target Population: Calibration Adjustment

The second stage adjustment makes the adjusted Web survey sample resemble the target population. More specifically, this section will examine calibration using generalized regression estimators (GREG) in Deville and Särndal (1992) and Deville *et al.* (1993) as a method of deriving the second stage weights.

Suppose that an initial set of weights, $\left\{w_j^0\right\}_{j\in\left(s^W\right)}$, is available and that a population total, t_y, of the variable of interest, y, is estimated as

$$\hat{t}_y^W = \sum_{j\in\left(s^W\right)} w_j^0 y_j ,$$

By using calibration with GREG, we may modify the set of initial weights, $\left\{w_j^0\right\}_{j\in\left(s^W\right)}$, in order to find a new set of calibration weights, $\left\{\tilde{w}_j\right\}_{j\in\left(s^W\right)}$, that produces $\tilde{t}_y^W = \sum_{j\in\left(s^W\right)} \tilde{w}_j y_j$, while respecting

$$\tilde{\mathbf{t}}_{\mathbf{z}} = \mathbf{t}_{\mathbf{z}}, \tag{6.4}$$

where

$$\mathbf{z}_j = \left\{z_{j1}, z_{j2}, ..., z_{jp}, ..., z_{jP}\right\}' \tag{6.5}$$

is a vector of values for P auxiliary variables for the unit j in the Web survey; $\mathbf{t}_{\mathbf{z}} = \sum_{i\in U} \mathbf{z}_i = \left(t_{z_1}, t_{z_2}, ..., t_{z_p}, ..., t_{z_P}\right)$, is the set of the population (U) marginal totals of all P covariates with $U = \left\{1, 2, ..., N\right\}$; and $\tilde{\mathbf{t}}_{\mathbf{z}} = \sum_{j\in\left(s^W\right)} \tilde{w}_j \mathbf{z}_j$, estimates of $\mathbf{t}_{\mathbf{z}}$ adjusted by the calibration weights, $\tilde{w}_j$. If these population total for the p^{th} auxiliary variable is known and fixed, such as the number of males in the U.S., then the total for that variable becomes $t_{z_p} \equiv T_{z_p} = \sum_{i=1}^{N} z_{ip}$. When the population totals are not readily available, such as the number of persons with some disability, t_{z_p} may be replaced with estimates from a larger independent survey where the estimates may be more reliable than the survey on hand (Deville *et al.*, 1993, p.1015). The initial weights (w_j^0), in our case, will be $d_j^{W.PSA}$ ($j \in \left(s^W\right)$) in (6.1), (6.3), when propensity score adjustment is applied initially, or sample design weights, where the simplest form may be $w_j = N^W / n^W$, when no adjustment is applied beforehand.

The GREG algorithm minimizes the measures of distance between w_j^0 and $\tilde{w}_j$, that is,

$$\sum_{i\in(s^w)} G^*\left(w_j^0, \tilde{w}_j\right), \tag{6.6}$$

subject to the constraint (6.4), where G^* is a distance function associated with generalized least squares (GLS), such that

$$G^*\left(w_j^0, \tilde{w}_j\right) = w_j^0 G\left(\frac{\tilde{w}_j}{w_j^0}\right) = \frac{w_j^0}{2}\left(\frac{\tilde{w}_j}{w_j^0} - 1\right)^2. \tag{6.7}$$

We seek to find $\left\{\tilde{w}_j\right\}$ by minimizing (6.6) with (6.7), while respecting (6.4). This operation is equivalent to minimizing the quantity $\sum_{j\in(s^w)} w_j^0 G^*\left(w_j^0, \tilde{w}_j\right) - \boldsymbol{\lambda}'\left(\tilde{\mathbf{t}}_{\mathbf{z}} - \mathbf{t}_{\mathbf{z}}\right)$, where $\boldsymbol{\lambda} = \left(\lambda_1, \lambda_2, ..., \lambda_p, ..., \lambda_P\right)'$ is a P vector of Lagrange multipliers. This minimization leads to the desired calibration weights, $\tilde{w}_j = w_j^0 F\left(\mathbf{z}_j'\boldsymbol{\lambda}\right)$, where $F\left(\mathbf{z}_j'\boldsymbol{\lambda}\right)$ is the inverse function of $g^*\left(w_j^0, \tilde{w}_j\right) = dG^*\left(w_j^0, \tilde{w}_j\right)/d\tilde{w}_j$. For the GLS distance function in (6.7), $F(u) = g^*\left(u\right)^{-1} = 1 + u$.

In order to compute $\tilde{w}_j$, $\boldsymbol{\lambda}$ must be determined by solving the calibration equation

$$\sum_{j\in(s^w)} \tilde{w}_j \mathbf{z}_j = \sum_{j\in(s^w)} w_j^0 F\left(\mathbf{z}_j'\boldsymbol{\lambda}\right)\mathbf{z}_j = \mathbf{t}_{\mathbf{z}}, \tag{6.8}$$

where the vector $\boldsymbol{\lambda}$ is the only unknown component. Following Deville and Särndal (1992), we rearrange (6.8) and define

$$\phi_{s^W}(\boldsymbol{\lambda}) = \sum_{j \in (s^W)} w_j^0 \left\{ F\left(\mathbf{z}_j'\boldsymbol{\lambda}\right) - 1 \right\} \mathbf{z}_j = \mathbf{t}_{\mathbf{z}} - \hat{\mathbf{t}}_{\mathbf{z}} . \tag{6.9}$$

Based on the iterative procedure, such as Newton's method, $\boldsymbol{\lambda}$ is solved for as follows. First, expand $\phi_{s^W}\left(\boldsymbol{\lambda}^{t+1}\right)$ around $\boldsymbol{\lambda}^t$, where $\boldsymbol{\lambda}^t$ is the value at the t^{th} iteration and $\boldsymbol{\lambda}^{t+1}$ at the $(t+1)^{\text{st}}$, such that

$$\phi_{s^W}\left(\boldsymbol{\lambda}^{t+1}\right) = \phi_{s^W}\left(\boldsymbol{\lambda}^t\right) + \phi'_{s^W}\left(\boldsymbol{\lambda}^t\right)\left(\boldsymbol{\lambda}^{t+1} - \boldsymbol{\lambda}^t\right), \tag{6.10}$$

where

$$\phi'_{s^W}\left(\boldsymbol{\lambda}^t\right) = \left.\frac{d\phi_{s^W}\left(\boldsymbol{\lambda}^t\right)}{d\boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^t} = \sum_{j \in (s^W)} w_j^0 F'\left(\mathbf{z}_j'\boldsymbol{\lambda}^t\right)\mathbf{z}_j\mathbf{z}_j' \tag{6.11}$$

is the $P \times P$ matrix of partial derivatives and $F'\left(\mathbf{z}_j'\boldsymbol{\lambda}\right)$ is the derivative with respect to the argument $\mathbf{z}_j'\boldsymbol{\lambda}$. Using (6.9), (6.10) and (6.11), we obtain

$$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \left[\phi'_{s^W}\left(\boldsymbol{\lambda}^t\right)\right]^{-1}\left[\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{t}}_{\mathbf{z}} - \phi_{s^W}\left(\boldsymbol{\lambda}^t\right)\right]$$

as in the equation (3.5) in Deville and Särndal (1992). For an unrestricted GLS distance function, a closed form solution for the Lagrange multiplier can be obtained as

$$\boldsymbol{\lambda} = \left[\sum_{j \in (s^W)} w_j^0 \mathbf{z}_j \mathbf{z}_j'\right]^{-1}\left(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{t}}_{\mathbf{z}}\right).$$

One problem associated with the GLS distance function is that the final weight may be negative or extremely large (see Deville and Särndal, 1992 for detail). In order to avoid such situations, this study uses the truncated linear (L, U) distance function presented in Deville *et al.* (1993) and Jayasuriya and Valliant (1996):

$$G(x)=\begin{cases}1/2(x-1)^2, & \text{if } L<x<U\\ \infty, & \text{otherwise}\end{cases},$$

using two fixed constants L and U. The corresponding F function becomes

$$F(u)=\begin{cases}L, & \text{if } u<L-1\\ 1+u, & \text{if } L-1\le u\le U-1\\ U, & \text{if } u>U-1\end{cases}.$$

Define three subsets of the sample as

$$s_A^W=\left\{j\in\left(s^W\right):\mathbf{z}_j'\boldsymbol{\lambda}<L-1\right\},$$

$$s_B^W=\left\{j\in\left(s^W\right):L-1\le\mathbf{z}_j'\boldsymbol{\lambda}\le U-1\right\},\text{ and}$$

$$s_C^W=\left\{j\in\left(s^W\right):\mathbf{z}_j'\boldsymbol{\lambda}>U-1\right\}.$$

Expression (6.9) can be decomposed as

$$\begin{aligned}\boldsymbol{\phi}_{s^W}\left(\boldsymbol{\lambda}^t\right)&=\boldsymbol{\phi}_{s_A^W}\left(\boldsymbol{\lambda}^t\right)+\boldsymbol{\phi}_{s_B^W}\left(\boldsymbol{\lambda}^t\right)+\boldsymbol{\phi}_{s_C^W}\left(\boldsymbol{\lambda}^t\right)\\&=(L-1)\sum_{j\in\left(s_A^W\right)}w_j^0\mathbf{z}_j+\sum_{j\in\left(s_B^W\right)}w_j^0\left(\mathbf{z}_j'\boldsymbol{\lambda}\right)\mathbf{z}_j+(U-1)\sum_{j\in\left(s_C^W\right)}w_j^0\mathbf{z}_j,\end{aligned}\tag{6.12}$$

.

Since $F'(u)=0$ for s_A^W and s_C^W, and $F'(u)=1$ for s_B^W, (6.11) becomes

$$\boldsymbol{\phi}'_{s^W}\left(\boldsymbol{\lambda}^t\right)=\sum_{j\in\left(s_B^W\right)}w_j^0\mathbf{z}_j\mathbf{z}_j'.\tag{6.13}$$

The computational algorithm is implemented in **`cal.fcn`** using **R**© shown in Appendix 1.2 as:

(1) Assign starting value $\boldsymbol{\lambda}^0=\mathbf{0}_{P\times1}$.

(2) Evaluate (6.10), substituting (6.12) and (6.13) to compute the components.

(3) Check whether the convergence criterion is achieved by using $\max\left[\left(\boldsymbol{\lambda}^{t+1}-\boldsymbol{\lambda}^{t}\right)/\boldsymbol{\lambda}^{t}\right]\le\varepsilon$, where ε is some small constant, say $\varepsilon=10^{-2}$.

(4) If convergence is obtained, go to step (5), otherwise repeat (2) and (3).

(5) Evaluate the final weights as $\tilde{w}_j=w_j^0F\left(\mathbf{z}_j'\boldsymbol{\lambda}^*\right)$, where $\boldsymbol{\lambda}^*$ is the converged value.

Because $\tilde{w}_j$ satisfies (6.4) and (6.9), all population constraints are satisfied even with the restriction placed on the range of weights.

The resulting restricted regression estimator of t_y for the Web survey is then

$$\tilde{t}_y^{W.Cal}=\sum_{j\in\left(s^W\right)}\tilde{w}_jy_j\,.$$

If the *L* and *U* restrictions have no effect, then the estimator reduces to the GREG defined as

$$\tilde{t}_y^{W.Cal}=\hat{t}_y^W+\left(\mathbf{t}_\mathbf{z}-\hat{\mathbf{t}}_\mathbf{z}^W\right)'\hat{\mathbf{B}}_{ws}\,. \tag{6.14}$$

The regression coefficient, when there are no weight restrictions, can be estimated directly, using weighted least squares, as

$$\hat{\mathbf{B}}_{ws}=\left(\sum_{j\in\left(s^w\right)}w_j^0\mathbf{z}_j\mathbf{z}_j'\right)^{-1}\left(\sum_{j\in\left(s^w\right)}w_j^0\mathbf{z}_jy_j\right). \tag{6.15}$$

In terms of matrices and vectors, (6.15) becomes

$$\hat{\mathbf{B}}_{ws}=\left(\mathbf{Z}_s^{W\prime}\mathbf{W}_s^W\mathbf{Z}_s^W\right)^{-1}\left(\mathbf{Z}_s^{W\prime}\mathbf{W}_s^W\mathbf{Y}_s^W\right)=\left(\mathbf{A}_s^W\right)^{-1}\left(\mathbf{Z}_s^{W\prime}\mathbf{W}_s^W\mathbf{Y}_s^W\right), \tag{6.16}$$

where $\mathbf{Z}_s^W$ is a $n^W\times P$ matrix of covariates for the n^W cases in the Web sample; $\mathbf{W}_s=diag\left(\tilde{w}_j\right)$; $\mathbf{Y}_s^W$ a $n^W\times1$ vector of the study variable in the Web sample; and

$\mathbf{A}_s^W = \mathbf{Z}_s^{W\prime}\mathbf{W}_s^W\mathbf{Z}_s^W$. In the general case with calibration and weight restriction, the estimator of the mean becomes

$$\tilde{\bar{y}}^{W.Cal} = \frac{\sum_{j\in(s^W)} \tilde{w}_j y_j}{\sum_{j\in(s^W)} \tilde{w}_j},$$

where the calibration weight is

$$\tilde{w}_j = w_j^0 F\left(\mathbf{z}_j'\boldsymbol{\lambda}^*\right).$$

6.4 Theory for Propensity Score Adjustment and Calibration Adjustment

The bias properties of Web survey estimates will be examined in this section with respect to population total estimates. Two structural models are considered: one in which the population follows a stratified model with strata defined by propensity score subclasses and the other in which covariates are used. The unadjusted Web survey estimate, $\hat{t}_y^W$, will generally be biased under either of these models.

6.4.1 Stratification Model

Suppose that there is an underlying structural model *M* that produces

$$E_M\left(y_i\right) = \mu_c,$$

where $i \in U_c$; and U_c is subclass *c* in the universe, U. Under this model, the expected value of the population total is

$$E_M\left(t_y\right) = \sum_{c=1}^{C} N_c\mu_c.$$

Because this model uses subclasses formed based on the quintiles of $e(\mathbf{x})$, we interpret N_c to be the count that would be obtained if the propensity score adjustment were applied to the entire population.

The Web survey estimate without adjustment is $\hat{t}_y^W = \sum_{c=1}^{C} \sum_{j \in (s_c^W)} d_j^W y_j$ and its expectation over the model is

$$\begin{aligned} E_M\left(\hat{t}_y^W\right) &= \sum_{c=1}^{C} \mu_c \sum_{j \in (s_c^W)} d_j^W \\ &= \sum_{c=1}^{C} \hat{N}_c^W \mu_c. \end{aligned} \tag{6.17}$$

The bias in (6.17) with respect to *M* is

$$E_M\left(\hat{t}_y^W - t_y\right) = \sum_{c=1}^{C} \mu_c \left(\hat{N}_c^W - N_c\right). \tag{6.18}$$

Suppose that there is a mechanism π that describes how persons voluntarily become part of the Web sample. In particular, suppose that $\delta_i = \begin{cases} 1, \text{ if unit } i \text{ in Web sample} \\ 0, \text{ otherwise} \end{cases}$, and that $E_\pi(\delta_i) = \pi_i^W$. The π may be difficult or impossible to model, although the propensity score modeling is an attempt to do this. If $E_\pi\left(\hat{N}_c^W\right) = N_c$, the model bias (6.18) averages to zero over the voluntary mechanism:

$$E_\pi E_M\left(\hat{t}_y^W - t_y\right) = E_\pi\left\{\sum_{c=1}^{C} \mu_c \left(\hat{N}_c^W - N_c\right)\right\} = 0.$$

Only under both the model *M* and the volunteering mechanism π, the unadjusted Web sample estimate $\hat{t}_y^W$ becomes unbiased. Note that it is quite likely that $\pi_i^W = 0$ for some persons, because they would never volunteer to participate in a Web survey. Generally,

$E_\pi\left(\hat{N}_c^W\right)=\sum_{i\in(U_c)} d_i^W \pi_i^W$. If $d_i^W = 1/\pi_i^W$, then $E_\pi\left(\hat{N}_c^W\right)=N_c$, but if $\pi_i^W = 0$ for any persons, this cannot hold. As a result, we expect $\hat{t}_y^W$ to be biased.

By applying propensity score adjustment weights, we obtain a new Web survey estimate,

$$\hat{t}_y^{W.PSA}=\sum_{c=1}^{C}\sum_{j\in\left(s_c^W\right)} d_j^{W.PSA} y_j \,, \tag{6.19}$$

where $d_j^{W.PSA}$ is from (6.2). The *M*-expected value of this estimate is

$$\begin{aligned} E_M\left(\hat{t}_y^{W.PSA}\right) &= \sum_{c=1}^{C}\sum_{j\in\left(s_c^W\right)} \frac{\hat{N}_c^R/\hat{N}^R}{\hat{N}_c^W/\hat{N}^W} d_j^W \mu_c \\ &= \frac{\hat{N}^W}{\hat{N}^R}\sum_{c=1}^{C}\hat{N}_c^R\mu_c \,, \end{aligned}$$

and its model bias is $E_M\left(\hat{t}_y^{W.PSA}-t_y\right)=\sum_{c=1}^{C}\mu_c\left(\frac{\hat{N}^W}{\hat{N}^R}\hat{N}_c^R-N_c\right)$. If the weights in the reference sample and the Web sample are scaled so that $\hat{N}^W=\hat{N}^R$, and if the application of the π distribution, appropriate to the reference sample, produces $E_\pi\left(\hat{N}_c^R\right)=N_c$, $\hat{t}_y^{W.PSA}$ will be an unbiased estimator of the population total in the sense that $E_\pi E_M\left(\hat{t}_y^{W.PSA}-t_y\right)=0$. Therefore, the role of the reference survey sample is as important as the propensity score model that attempts to describe π.

Another approach to analyzing the propensity score adjustment estimator is to consider the correction factor, f_c, described earlier to be a response propensity adjustment factor. That is, $\pi_j^{W.PSA}=\left(f_c d_j^W\right)^{-1}=\left(d_j^{W.PSA}\right)^{-1}$ is the estimated propensity of

being in the Web sample. Since all $j \in s_c^W$ get the same weight adjustment factor, f_c, we could use

$$\overline{e}_c = \frac{1}{n_c} \sum_{j \in s_c^W} e\left(\mathbf{x}_j\right)$$

as an alternative to $1/f_c$, although this option is not pursued in this study. If $1/d_j^{W.PSA}$ can be interpreted as an inverse inclusion probability, then (6.19) becomes analogous to a Horvitz-Thompson estimator and is unbiased with respect to the volunteering mechanism because

$$E_\pi\left(\delta_j\right) \doteq \pi_j^{W.PSA}.$$

Consequently, the propensity score adjustment estimator is *M*-π unbiased, if $\hat{N}_c^R$ is a π-unbiased estimator and is unbiased with respect to the volunteering mechanism, and if $1/d_j^{W.PSA}$ is an inclusion probability.

6.4.2 Regression Model

A more elaborate model would be one that accounts for covariates which are good predictors of *y*. To that end, suppose that there are covariates that affect the study variable in the following way:

$$E_M\left(y_i\right) = \boldsymbol{\beta}'\mathbf{z}_i, \tag{6.20}$$

where $i \in U_c$ and $\mathbf{z}_i = \left\{z_{i1}, z_{i2}, ..., z_{ip}, ..., z_{iP}\right\}'$ similarly defined as in (6.5). Here, the model bias of an unadjusted Web sample estimate for the population total is

$$
\begin{aligned}
E_M\left(\hat{t}_y^W - t_y\right) &= \sum_{c=1}^{C}\sum_{j\in\left(s_c^W\right)} E_M\left(d_j^W y_j\right) - \sum_{c=1}^{C}\sum_{i\in U_c} E_M\left(y_i\right) \\
&= \left(\sum_{c=1}^{C}\sum_{j\in\left(s_c^W\right)} d_j^W \boldsymbol{\beta}'\mathbf{z}_j\right) - \left(\sum_{c=1}^{C}\sum_{i\in U_c} \boldsymbol{\beta}'\mathbf{z}_i\right) \\
&= \boldsymbol{\beta}'\hat{\mathbf{t}}_{\mathbf{z}}^W - \boldsymbol{\beta}'\mathbf{t}_{\mathbf{z}} \\
&= \boldsymbol{\beta}'\left(\hat{\mathbf{t}}_{\mathbf{z}}^W - \mathbf{t}_{\mathbf{z}}\right),
\end{aligned}
\tag{6.21}
$$

where $\mathbf{t}_{\mathbf{z}} = \sum_{c=1}^{C}\sum_{i\in U_c}\mathbf{z}_i$ and $\hat{\mathbf{t}}_{\mathbf{z}}^W = \sum_{c=1}^{C}\sum_{j\in\left(s_c^W\right)} d_j^W\mathbf{z}_j$. If volunteering, i.e., the π mechanism, satisfies $E_\pi\left(\hat{\mathbf{t}}_{\mathbf{z}}^W - \mathbf{t}_{\mathbf{z}}\right) = 0$, $\hat{t}_y^W$ becomes M-π unbiased. However, as noted in the previous section, this assumption is unrealistic. Consequently, we can expect the unadjusted estimator, $\hat{t}_y^W$, to be biased.

The combination of calibration adjustment using GREG without the weight constraints and propensity score adjustment produces the following estimator from (6.14):

$$
\tilde{t}_y^{W.Cal} = \hat{t}_y^{W.PSA} + \hat{\mathbf{B}}_{ws}'\left(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{t}}_{\mathbf{z}}^{W.PSA}\right). \tag{6.22}
$$

Based on the model (6.20), its model expectation is

$$
E_M\left(\tilde{t}_y^{W.Cal}\right) = E_M\left(\hat{t}_y^{W.PSA}\right) + E_M\left(\hat{\mathbf{B}}_{ws}'\left(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{t}}_{\mathbf{z}}^{W.PSA}\right)\right). \tag{6.23}
$$

The expectation of this regression coefficient is

$$
\begin{aligned}
E_M\left(\hat{\mathbf{B}}_{ws}\right) &= \left(\mathbf{A}_s^W\right)^{-1}\mathbf{Z}_s^{W\prime}\mathbf{W}_s^W E_M\left(\mathbf{Y}_s^W\right) \\
&= \left(\mathbf{A}_s^W\right)^{-1}\mathbf{Z}_s^{W\prime}\mathbf{W}_s^W\mathbf{Z}_s^W\boldsymbol{\beta} \\
&= \boldsymbol{\beta},
\end{aligned}
\tag{6.24}
$$

i.e., $\hat{\mathbf{B}}_{ws}$ is M-unbiased of $\boldsymbol{\beta}$. Using (6.24), (6.23) becomes

$$
\begin{aligned}
E_M\left(\tilde{t}_y^{W.Cal}\right) &= E_M\left(\hat{t}_y^{W.PSA}\right) + E_M\left(\hat{\mathbf{B}}_{ws}{}'\left(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{t}}_{\mathbf{z}}^{W.PSA}\right)\right) \\
&= \sum_{j\in\left(s_c^W\right)} d_j^{W.PSA}\boldsymbol{\beta}'\mathbf{z}_j + \boldsymbol{\beta}'\left(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{t}}_{\mathbf{z}}^{W.PSA}\right) \\
&= \boldsymbol{\beta}'\hat{\mathbf{t}}_{\mathbf{z}}^{W.PSA} + \boldsymbol{\beta}'\mathbf{t}_{\mathbf{z}} - \boldsymbol{\beta}'\hat{\mathbf{t}}_{\mathbf{z}}^{W.PSA} \\
&= \boldsymbol{\beta}'\mathbf{t}_{\mathbf{z}} \\
&= E_M\left(t_y\right).
\end{aligned}
\tag{6.25}
$$

Important facts from above are (1) that (6.25) holds even if $\hat{\mathbf{t}}_{\mathbf{z}}^{W.PSA}$ has a π-bias, because they cancel out each other in the *M* expectation but (2) that $\mathbf{t}_{\mathbf{z}}$ does have to be correct. If $\mathbf{t}_{\mathbf{z}}$ contains estimates from some other survey, the model bias will have the form $\boldsymbol{\beta}_{sub}{}'\left(\mathbf{t}_{\mathbf{z},sub}^{*} - \mathbf{t}_{\mathbf{z},sub}\right)$, where the subscript "*sub*" denotes the part of the $\mathbf{z}$-vector whose totals come from that survey, and $\mathbf{t}_{\mathbf{z},sub}^{*}$ is the vector of covariate estimates from that survey. If $E_{\pi^*}\left(\mathbf{t}_{\mathbf{z},sub}^{*} - \mathbf{t}_{\mathbf{z},sub}\right) = 0$, where, in this case, E_{π^*} is the expectation over the selection mechanism for the other survey, then $\tilde{t}_y^{W.Cal}$ is *M*-π^* unbiased. Therefore, the calibration adjustment will produce *M*-unbiased estimates (or possibly, *M*-π^* unbiased estimates), when the model (6.20) holds.

In a case where the propensity score adjustment successfully adjusts for the probability of being in the Web survey sample, under the assumption of $E_\pi\left(\delta_j\right) = 1/d_j^{W.PSA} = \pi_j^{W.PSA}$, we obtain $E_\pi\left(\hat{t}_y^{W.PSA}\right) = t_y$, $E_\pi\left(\hat{\mathbf{t}}_{\mathbf{z}}^{W.PSA}\right) = \mathbf{t}_{\mathbf{z}}$, and $E_\pi\left(\hat{\mathbf{B}}_{ws}\right) \doteq \mathbf{B} = \left(\mathbf{Z}_N{}'\mathbf{Z}_N\right)^{-1}\mathbf{Z}_N{}'\mathbf{Y}_N$, which is the finite population version of the regression slope (6.23). These three expectations lead to $E_\pi\left(\tilde{t}_y^{W.Cal}\right) \doteq t_y$, when the population total $\mathbf{t}_{\mathbf{z}}$ is used in deriving $\tilde{t}_y^{W.Cal}$. Therefore, (1) if the propensity score adjustment correctly

accounts for the volunteering mechanism π, $\tilde{t}_y^{W.Cal}$ is π-unbiased, and (2) if the model M is correct, $\tilde{t}_y^{W.Cal}$ is M-unbiased. If unbiased estimates from another survey are used, then the calibration estimator will be M-π^* unbiased.

Suppose that the propensity score adjustment does not fully adjust for π-bias. Then $E_\pi\left(\hat{t}_y^{W.PSA}\right) = t_y + b_y$ and $E_\pi\left(\hat{\mathbf{t}}_{\mathbf{z}}^{W.PSA}\right) = \mathbf{t}_{\mathbf{z}} + \mathbf{b}_{\mathbf{z}}$, where b_y is the bias which can take a positive or negative direction and $\mathbf{b}_{\mathbf{z}} = \left(b_{z_1}, ...b_{z_p}, ..., b_{z_P}\right)$ whose components can also be positive or negative. The π expectation of the calibration adjusted estimate is then

$$\begin{aligned} E_\pi\left(\tilde{t}_y^{W.Cal}\right) &\doteq t_y + b_y + E_\pi\left(\hat{\mathbf{B}}_{ws}{}'\left(\mathbf{t}_{\mathbf{z}} - \left(\mathbf{t}_{\mathbf{z}} + \mathbf{b}_{\mathbf{z}}\right)\right)\right) \\ &= t_y + b_y - E_\pi\left(\hat{\mathbf{B}}_{ws}{}'\mathbf{b}_{\mathbf{z}}\right). \end{aligned} \tag{6.26}$$

Expression (6.26) is not equal to t_y, unless $b_y = E_\pi\left(\hat{\mathbf{B}}_{ws}{}'\mathbf{b}_{\mathbf{z}}\right)$, which is not true in general. When propensity score adjustment is not correct, $\tilde{t}_y^{W.Cal}$ will not generally be π-unbiased, meaning that the estimate is not unbiased with respect to the volunteering mechanism. However, it can be model unbiased as long as y_i follows a linear model M in (6.20) which we specify correctly.

Chapter 7: Application of the Alternative Adjustments for Volunteer Panel Web Surveys

7.1 Introduction

This chapter will document the performance of the proposed adjustment in Chapter 6 for volunteer panel Web surveys. The role of the adjustment is to decrease the bias occurring from the possibly nonrandom mechanism in the selection of panel Web survey respondents. In order to examine the degree of bias reduction, it is necessary to apply the adjustment for more than one sample realization. A logical approach for this purpose is to adopt simulation studies that use pseudo-populations whose population values are known.

This chapter will employ two survey data sets: the 2002 General Social Survey (GSS) and the 2003 Michigan Behavioral Risk Factor Surveillance System (BRFSS). Each of these will be used as a pseudo-population data set. The reasons for using these data are two-fold. First, one interesting feature of both surveys is that they contain an Internet supplement which provides information about whether respondents have Internet access or not. Since the volunteer panels in Web surveys are required to have their own Web access, information on Web access ownership becomes essential for constructing a pool of units potentially eligible to be included the Web surveys. The full sample of GSS and BRFSS themselves are capable of serving as populations as well as potential pools of reference survey sample units. Second, unlike existing research where the focus of adjustment is placed on polling and election outcomes, having two data sets will expand the scope of the examination to a wide range of different substantive study variables.

More specifically, GSS provides attitudinal information toward general social issues, and BRFSS gives factual information about health-related behaviors.

Two case studies comprise this chapter, where one study utilizes GSS data and the other utilizes BRFSS data. While both case studies will examine the performance of propensity score adjustment and calibration adjustment as bias reduction techniques, the emphasis of each study will differ. Section 7.2 will present the first study using GSS data, where the focus will be on the effectiveness of adjustment. Propensity-score-adjusted Web survey sample estimates will be compared to the reference survey sample estimates, and calibration-adjusted estimates will be compared to the population values. The second case study is presented in Section 7.3. It will use BRFSS data and expand the examination to multiple dimensions: the impact of covariate selection both in propensity score adjustment and calibration adjustment, the effectiveness of combining calibration adjustment with propensity score adjustment, and the calculation of variance estimates when multiple adjustment weights are applied.

7.2 Case Study 1: Application of Propensity Score Adjustment and Calibration Adjustment to 2002 General Social Survey Data

7.2.1 Construction of Pseudo-population and Sample Selection for Simulation

In order to assess the performance of bias reduction as described above, three different data sets are required: a population, a reference survey and a Web survey data set.

Samples mimicking the respondents in the Harris Interactive volunteer panel Web survey will be drawn based on subclass proportions from a real Harris Interactive Web

survey data set (obtained via a personal communication with Matthias Schonlau, see Schonlau *et al.*, 2004). The cells are formed by four demographic variables: age, gender, education and race. These proportions of Harris Interactive data are displayed in Table 7.1 along with the same cross-classified cell proportions of all respondents and respondents who use the Internet in the 2002 General Social Survey (GSS) data.

Table 7.1. Distribution of Age, Gender, Education and Race of GSS Full Sample, GSS Web User and Harris Interactive Survey Respondents

		High School or Less		*Some College or Above*	
		White	***Nonwhite***	***White***	***Nonwhite***
A. GSS Full Sample (n=2,746) [a]					
≤ 40 yrs	Female	9.76%	6.61%	5.51%	1.79%
	Male	9.65%	4.18%	4.41%	1.37%
41 yrs +	Female	16.75%	4.75%	8.39%	1.44%
	Male	13.14%	3.15%	7.75%	1.37%
				Sum	100.00%
B. GSS Web Users (n=1,692) [b]					
≤ 40 yrs	Female	11.68%	6.08%	7.97%	2.62%
	Male	10.52%	3.22%	6.69%	2.01%
41 yrs +	Female	11.50%	2.31%	11.01%	1.64%
	Male	9.49%	1.46%	10.16%	1.64%
				Sum	100.00%
C. Harris Interactive Respondents (n=8,195)					
≤ 40 yrs	Female	2.03%	1.64%	13.28%	13.37%
	Male	0.85%	0.61%	7.58%	9.09%
41 yrs +	Female	2.45%	0.48%	15.58%	4.58%
	Male	1.70%	0.24%	20.82%	5.71%
				Sum	100.00%

a. This sample size reflects the exclusion of 19 cases where some of the four covariates is missing.

b. This is the subset of Web users from the original 2002 GSS sample.

The 2002 GSS is a part of on-going biennial survey conducted by National Opinion Research Center with core funding from the National Science Foundation. The data were gathered in order to measure contemporary American society targeting

noninstitutionalized adults 18 years old and older. A representative national sample was drawn using multi-stage area probability sampling. Respondents were surveyed in a 90-minute in-person interview. The reported response rate for the 2002 GSS is 70%.[11] The protocol for the Harris Interactive Web surveys was discussed in Section 2.1 and 6.1.

From Table 7.1, we can examine the distributions of the 2002 GSS sample, its Web user subgroup and the Harris Interactive respondents. There is a noteworthy gap not only between the GSS sample and the two Web samples but, surprisingly, also between the two Web samples. The GSS full sample includes fewer young people and those with higher education than the two Web samples. The most notable disparity between the Harris Interactive data and the two GSS data is in the educational attainment level. While less than a half of the GSS and its Web users have some college or higher education, the same group of people makes up 90% of the Harris Interactive respondent data. Also, Harris Interactive respondents include more minorities, especially educated minorities, than the GSS samples. If a sample distributed like the Harris Interactive respondents is to provide unbiased estimates for the general population or even the population with Web access, some major weighting adjustment will be required.

The creation of the full pseudo-population starts from the GSS data set (U) which contains 2,746 cases with complete information on four stratifying variables in Table 7.1 and the Web usage variable.[12] The propensity score adjustment is feasible when all cases in the merged data have information on covariates included in the propensity score models. Otherwise, propensity scores for the units where some of the covariates are

[11] Information about the GSS is available at http://webapp.icpsr.umich.edu/GSS/ and http://norc.org/projects/gensoc3.asp.

[12] 19 units where the information on these five variables is missing are excluded from the original GSS data with 2,765 units.

missing cannot be computed, which hinders the adjustment. For this problem, missing values on the 14 covariates described in Table 7.2 that are used to build the propensity score models are imputed within the cell defined in Table 7.1 using hot-deck method. A larger population will facilitate testing of methods by simulation. By bootstrapping U with simple random sampling with replacement, the full pseudo-population (P^F) is created with a size of 20,000 persons.

As discussed earlier the 2002 GSS collected information about e-mail[13] and Internet usage.[14] Based on this information, people who are classified as Web users from P^F are retained for the pseudo-Web population (P^W), which results in the size of 12,306.[15] This pseudo-Web population will allow us to draw different types of Web samples, especially the one resembling Harris Interactive Web survey respondents, since Web usage is the prerequisite for the panel members in those surveys.

Using the two pseudo-populations, a reference sample and two types of Web sample are drawn in each simulation. The reference survey sample (s^R) is drawn from P^F by simple random sampling for the size of $n^R = 200$ using **`ref.sam`** function created in **R** (see Appendix 1.3). Since the 2002 GSS was conducted in the face-to-face mode, these reference samples will serve as face-to-face reference samples with known probabilities of selection.

Two types of Web samples are drawn from P^W by Poisson sampling with selection probabilities equal to cell proportions in Table 7.1.B and 7.1.C. For example,

[13] Question wording: "About how many minutes or hours per week do you spend sending and answering electronic mail or e-mail?"
[14] Question wording: "Other than e-mail, do you ever use the Internet or World Wide Web?"
[15] The proportion of the Web users in the full pseudo-population is the same as that in the original GSS data at 61%.

for the first Web sample, White female with high school education or less who were 40 years old or less were selected with a probability of 0.1168. Thus, the two samples were allocated according to the covariate distributions from Table 7.1.B and 7.1.C, where each cell serves as a stratum. The first Web sample, $s^{W.ST}$, is assumed to resemble the pseudo-Web population (Table 7.1.B), and the second, $s^{W.HI}$, the Harris Interactive respondents (Table 7.1.C). Both Web samples are drawn using **`pois.sam`** in Appendix 1.4 for the desired size of $n^{W.ST} = n^{W.HI} = 800$.[16] This procedure of selecting the three samples (s^R, $s^{W.ST}$ and $s^{W.HI}$) is repeated 2,000 times.

7.2.2 Propensity Score Adjustment

This study examines two variables: (1) y_{blks}: the proportion of people indicating warm feelings towards Blacks; and (2) y_{vote}: the proportion of people who voted in the 2000 presidential election. The estimates of y_{blks} and y_{vote} from the simulated Web samples, $s^{W.ST}$ and $s^{W.HI}$, are corrected by applying propensity score adjustment described in Section 6.2. There are 14 covariates used for adjusting y_{blks} and 13 for y_{vote}, where nine of each set of all covariates are demographic and the remainder are nondemographic characteristics.[17] As shown in Table 7.2, the significance of these covariates on y_{blks} and y_{vote} differs greatly. Some of the variables are continuous, while others are categorical with different numbers of categories.

[16] The actual Web sample sizes vary around 800, as Poisson sampling is used.
[17] The demographic/nondemographic nature of a given covariate is tentatively determined based on whether the variable is typically used in post-stratification or not.

Table 7.2. P-values of the Auxiliary Variables in Logit Models Predicting y_{blks} (Warm Feelings towards Blacks) and y_{vote} (Voting Participation in 2000 Presidential Election)[a]

			p-value	
Covariate	***Description***	***Type***	y_{blks}	y_{vote}
Demographic				
age	Age in years	Continuous	<.0001	<.0001
educ	Education in years	Continuous	<.0001	<.0001
newsize	Size of the residential area	Continuous	.2006	.1804
hhldsize	Household size	Continuous	.8318	.3496
income	Family income	Continuous	.4548	.0002
race	Race	4 categories	<.0001	.0002
gender	Gender	2 categories	<.0001	.1568
married	Marital Status	2 categories	.0616	.0280
region	Region of the residential area	4 categories	.0391	.2017
Nondemographic				
class	Self-rated social class	Continuous	.1435	<.0001
work	Employment status	2 categories	.6502	.1680
party	Political party affiliation	3 categories	.2174	<.0001
religion	Having a religion	2 categories	.1197	.8480
ethnofit	Opinion towards ethnic minorities	Continuous	<.0001	-

[a.] These analyses were done using the original GSS sample (n=2,746)

Based on the significance level (*p-value*) and the characteristics of the covariates (demographic or nondemographic) listed in Table 7.2, propensity score models are developed. The first model which serves as the base propensity model, *D1*, includes all demographic variables as main effects in a logistic regression as in (5.11)[18], such that

$$D1: \ \ln\left(\frac{\Pr(g=1)}{1-\Pr(g=1)}\right) = \alpha + \beta_1 age + \beta_2 educ + \beta_3 newsize + \beta_4 hhldsize + \beta_5 income + \beta_6 race + \beta_7 gender + \beta_8 married + \beta_9 region,$$

where $g=1$ for Web sample units and $g=0$ for reference sample units. The subsequent models are shown in Table 7.3, and their detailed specifications in **R**© are shown in

[18] This study focuses on the relationship between the substantive study variables and the covariates than on the relationship between the treatment variables and the covariates in constructing propensity score models.

Appendix 2. The respective effectiveness of different models will be compared in the following section. This will allow us to detect the importance of including highly predictive and/or nondemographic covariates in the propensity model.

Table 7.3. Propensity Score Models and Their Covariates by Variable [a.]

Covariate	***Propensity Score Models***					
	D1		***D2***		***D3***	
	All (1)		**Significant (2)**		**Nonsignificant (3)**	
Demographic (D)	y_{blks}	y_{vote}	y_{blks}	y_{vote}	y_{blks}	y_{vote}
age	√	√	√	√		
educ	√	√	√	√		
newsize	√	√			√	√
hhldsize	√	√			√	√
income	√	√		√	√	
race	√	√	√	√		
gender	√	√	√			√
Married	√	√		√	√	
Region	√	√	√			√
	N1		***N2***		***N3***	
	All (1)		**Significant (2)**		**Nonsignificant (3)**	
Nondemographic (N)	y_{blks}	y_{vote}	y_{blks}	y_{vote}	y_{blks}	y_{vote}
class	√	√		√	√	
Work	√	√			√	√
party	√	√		√	√	
Religion	√	√			√	√
ethnofit	√	-	√	-	√	-
	A1		***A2***		***A3***	
Demographics & Nondemographics (A)	**All (1)**		**Significant (2)**		**Nonsignificant (3)**	
	y_{blks}	y_{vote}	y_{blks}	y_{vote}	y_{blks}	y_{vote}
	D1+N1	***D1+N1***	***D2+N2***	***D2+N2***	***D3+N3***	***D3+N3***

[a.] Included covariates are indicated by check marks

Note: Propensity model ***4*** not shown in the table is the combination of ***D1*** and ***N2***.

The general steps for each simulation are:

(1) to combine the reference sample (s^R) and the Web sample ($s^{W.ST}$ or $s^{W.HI}$),

(2) to estimate the propensity $e(\mathbf{x}_i)$ of being in the Web sample rather than the reference sample for the i^{th} person in the combined sample,

(3) to divide the combined sample into five groups based on quintiles of the propensity scores, and

(4) to compute the weight $d_j^{W.PSA}$ defined in (6.2) for each person *j* in the Web sample.

7.2.3 Results Propensity Score Adjustment

Reference and Web samples are drawn using **`ref.sam`** and **`pois.sam.`** The adjustment and estimation described are carried out by **`psa.fcn`** function. The propensity score adjustment function includes the adjustment weight in (6.2). This is because the reference sample units have an equal probability of selection and the Web sample units are supposed to have unknown selection probabilities. The simulation is done over 2,000 times using **`psa.sim`** function in Appendix 1.5 which includes all functions introduced previously. Since the estimation benchmarks in this adjustment stage (propensity score adjustment) are from the reference sample (s^R) estimates, population values are not included in the discussion. However, for convenience, we will refer to the difference between the average of a Web sample estimate and the means of the reference sample estimates as a "bias."

Table 7.4. Simulation Means of Estimates by Different Samples before Adjustment

	s^R	$s^{W.ST}$	$s^{W.HI}$
y_{blks} : Proportion of warm feelings towards blacks (*M*=2000)	0.612	0.636	0.675
y_{vote} : Proportion of voters in 2000 election (*M*=1971) [a]	0.650	0.715	0.817

a. In simulations for y_{vote}, 29 simulations were not completed due to zero cases in subclasses in s_c^R defined by propensity scores, which resulted in inability to derive weights in (6.2).

Table 7.4 shows the respective unadjusted means of y_{blks} and y_{vote} from the three samples, s^R , $s^{W.ST}$, and $s^{W.HI}$ over all simulations. They are calculated as

$$\overline{y} = \sum_{m=1}^{M} y_m \Big/ M \ , \tag{7.1}$$

where M is the total number of simulated samples and y_m is an estimate from the m^{th} simulation. Web estimates deviate from the reference sample estimates, indicating that people in the Web samples are more likely to express warm feelings towards Blacks and more likely to have participated in the election than people in the reference sample. This result seems plausible when considering the cell proportions in Table 7.1 which were used to create $s^{W.ST}$ and $s^{W.HI}$. There is likely to be a higher proportion of people with higher levels of education and minorities in the Web samples than in s^R. The biases are even larger between $s^{W.HI}$ and s^R. For the voting behavior, the estimate from $s^{W.HI}$ is off by 16.7% from the reference sample estimate. Therefore, it becomes necessary to decrease the bias.

7.2.3.1 Performance of Propensity Score Adjustment

Correction for the deviations of Web sample estimates is carried out by applying propensity score adjustment. First the base propensity model (*D1*) which includes all demographic covariates was applied. Table 7.5 compares unadjusted and *D1* adjusted estimates in the relation to reference sample estimates. For example, the *D1* propensity score adjusted mean (*y.D1*) for y_{blks} is 0.623 based on $s^{W.ST}$ samples, which is closer to the reference sample mean (*y.R*: 0.615) than the unadjusted mean (*y.U*: 0.636). By incorporating adjustment weights, the Web estimates are closer to the reference sample values than the unadjusted estimates.

Throughout this section, the performance of propensity score adjustment can be evaluated with respect to three criteria: bias and its reduction, root mean square deviation and its reduction, and standard error.

7.2.3.1.A Bias and Percent Bias Reduction

As discussed above, the "bias" measure of the Web survey estimates compared to the reference survey estimates takes the following form:

$$bias\left(\bar{y}^{W}\right)=\sum_{m=1}^{M} y_m{}^{W}-\sum_{m=1}^{M} y_m{}^{R},$$

where $y_m{}^{R}$ and $y_m{}^{W}$ are the reference and Web estimates from the m^{th} simulation with $m=1,...,M$.

Additionally, percent bias reduction ($p.bias$) is calculated using an adapted form of (5.14) as

$$p.bias\left(\bar{y}^{W.PSA}\right)=\left[\frac{\left|bias\left(\bar{y}^{W.U}\right)\right|-\left|bias\left(\bar{y}^{W.PSA}\right)\right|}{\left|bias\left(\bar{y}^{W.U}\right)\right|}\right]\times 100, \qquad (7.2)$$

where $\bar{y}^{W.U}$ is the simulation mean of the unadjusted Web estimate and $\bar{y}^{W.PSA}$ is the simulation mean by propensity score adjustment (*PSA* is substituted by model names hereafter). It is expected that the unadjusted estimates have larger $bias$ than the adjusted ones. The larger the $p.bias$, the more effective the propensity score adjustment in reducing bias. A negative $p.bias$ indicates that the adjustment actually makes the estimates worse.

7.2.3.1.B Root Mean Square Deviation and Percent Root Mean Square Deviation Reduction

The second evaluation criterion is related to the root mean square deviation (*rmsd*) summarizes the deviation of Web estimates from the reference estimate over all simulations. This statistic is calculated as

$$rmsd\left(\bar{y}^{W}\right)=\sqrt{\sum_{m=1}^{M}\left(y_{m}{}^{W}-y_{m}{}^{R}\right)^{2}\Big/M}\ .$$

From this statistic, we may compare *rmsd* 's of the Web sample estimates derived from adjustments using different propensity models. Estimates with smaller *rmsd* may be considered less-deviated from the reference estimates than others.

Just like (7.2), the percent deviation reduction (*p.rmsd*) is also computed in order to provide the relative size of *rmsd* of the adjusted estimates to the unadjusted estimates as:

$$p.rmsd\left(\bar{y}^{W.PSA}\right)=\left[\frac{rmsd\left(\bar{y}^{W.U}\right)-rmsd\left(\bar{y}^{W.PSA}\right)}{rmsd\left(\bar{y}^{W.U}\right)}\right]\times 100\ .$$

This will provide the reduction of deviation in Web survey estimates achieved by propensity score adjustment.

7.2.3.1.C Standard Error

While applying adjustment in the estimation may reduce biases in the estimates, the variability introduced by the weights may increase the variability of the estimates. It is important to understand the trade-off between bias reduction and variance increase. The variability in estimates is calculated in the form of a standard error (*se*) of the simulation mean as:

$$se\left(\bar{y}^W\right) \approx \sqrt{\sum_{m=1}^{M}\left(y_m^W - \bar{y}^W\right)^2 \Big/ M} \ ,$$

where y_m^W is the Web sample estimate from the m^{th} simulation and $\bar{y}^W$ is the average of y_m^W defined in (7.1). This statistic allows us to examine the magnitude of added variability on the estimates due to the propensity score adjustment.

Table 7.5. Reference Sample and Unadjusted and Propensity Score Adjusted Web Sample Estimates for y_{blks} and y_{vote}

	$s^{W.ST}$						$s^{W.HI}$					
	estimate	***bias***	***p.bias***	***rmsd***	***p.rmsd***	***se***	***estimate***	***bias***	***p.bias***	***rmsd***	***p.rmsd***	***se***
y_{blks} (M=2,000)												
y.R	0.612					0.0339	0.612					0.034
y.U	0.636	0.024		0.045		0.0160	0.675	0.064		0.074		0.016
y.D1	0.623	0.012	52.4%	0.040	9.6%	0.0221	0.638	0.026	58.6%	0.052	29.4%	0.032
y_{vote} (M=1,971)												
y.R	0.650					0.034	0.650					0.034
y.U	0.715	0.065		0.075		0.015	0.817	0.167		0.171		0.013
y.D1	0.709	0.059	9.7%	0.069	8.3%	0.022	0.724	0.074	55.7%	0.086	50.0%	0.031

Note: *y.R*: Reference sample estimate.
y.U: Unadjusted Web sample estimate.
y.D1: Web sample estimate after propensity score adjustment using model *D1*.

Table 7.5 exhibits simulation estimates of y_{blks} and y_{vote} and their evaluation statistics when no adjustment and adjustment by *D1* model are applied for both $s^{W.ST}$ and $s^{W.HI}$ (see Appendix 3 for the same information for all adjusted estimates based on all propensity models). When *D1* adjustment is applied, biases and deviations in Web estimates from the reference sample estimates are decreased dramatically. The greatest advantage of using propensity score adjustment is in the samples mimicking Harris Interactive respondents – the larger bias reduction is in $s^{W.HI}$ than in $s^{W.ST}$ for both study

variables. This echoes the statement in Cochran *et al.* (1954, pp.246) that "adjustment will only be seriously helpful when the sampling procedure is not random…." The reductions in the *bias* of estimates from $s^{W.HI}$ are 58.6% and 55.7%. Their *p.rmsd* 's are also large at 29.4% and 50%. Nonetheless, the adjusted estimates have larger standard error, showing that the reduction in the bias and deviation comes at the cost of increased variability. The trade-off between the decrease in deviation and the increase in standard error will be discussed in detail shortly.

7.2.3.2 Effect of Covariates in Propensity Score Models

The choice of covariates can be on important factors in the performance of propensity score adjustment. Assessment of the role of covariates is carried out exclusively using $s^{W.HI}$ for several different sets of covariates. First, different propensity models are developed by the significance of the covariates predicting y_{blks} and y_{vote}. Using a cut-point of $p = .05$, covariates in Table 7.2 are classified by whether they are highly ($p < .05$) or weakly predictive ($p \geq .05$). As a result, there are three models related only to demographic variables as shown in Table 7.3: all demographic covariates are included in the base propensity model (*D1*); highly predictive covariates only (*D2*); and weakly predictive covariates only (*D3*).

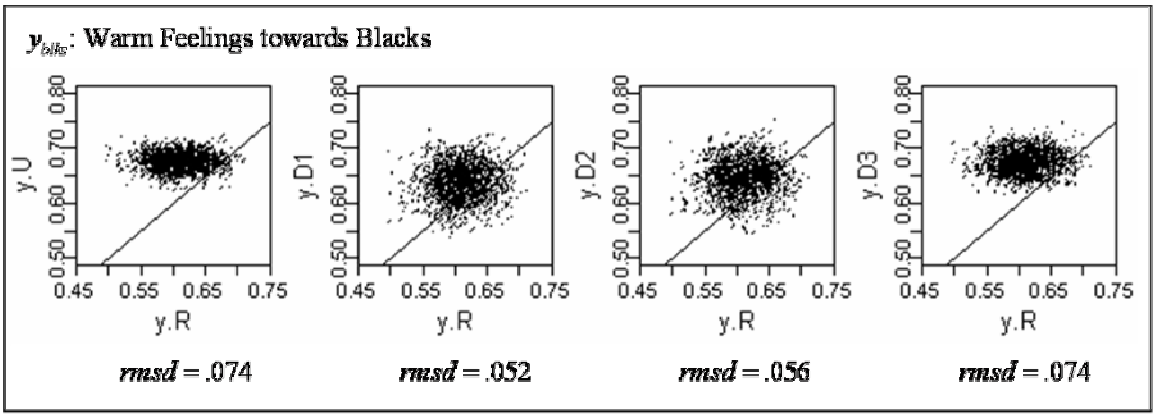


Figure 7.1. Relationship between the Distributions of the Different Web Sample Estimates and the Reference Sample Estimates for y_{blks} (Warm Feelings towards Blacks)

The unadjusted ($y.U$) and the adjusted Web estimates using *D1*, *D2*, and *D3* ($y.D1$, $y.D2$, and $y.D3$, respectively) are plotted against the reference sample estimate ($y.R$) for y_{blks} in Figure 7.1 and for y_{vote} in Figure 7.2 for all simulated samples (see Appendix 4 for all scatter plots of the estimates using all propensity models for both study variables in both $s^{W.ST}$ and $s^{W.HI}$). Underneath each scatter plot is displayed the corresponding *rmsd* for each adjustment. A diagonal $y = x$ reference line is drawn in each panel in Figure 7.1 and Figure 7.2. If the propensity score adjusted Web sample estimates were always equal to the reference sample estimates, then all points would fall on the reference line. Therefore, in the scatter plots, as the cluster of dots is approaching the reference line, the disparity of Web estimates is diminished. The scatter plot with the dots closest to the identity line indicates the best adjustment method in terms of deviation. Widely dispersed clusters are the evidence of increased variability.

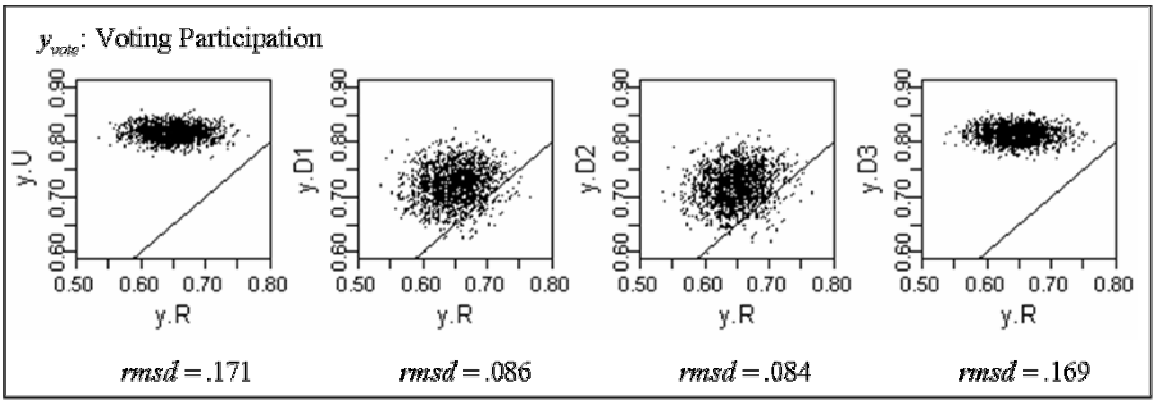


Figure 7.2. Relationship between the Distributions of the Different Web Sample Estimates and the Reference Sample Estimates for y_{vote} (Voting Participation)

Figure 7.1 and 7.2 convey the same messages. Among the three adjustments, *D1* and *D2* outperform *D3*. When the propensity score model is composed of only highly predictive covariates (*D2*), the level of adjustment is comparable to the base model that includes all variables (*D1*). The propensity score adjustment based on weakly predictive covariates (*D3*) does not improve the point estimates to any degree. The figures also illustrate the increased variability of estimates when using propensity score adjustment weights. Once the weights are incorporated, the scatter plots in the panel 2 and 3 show higher variability. In particular, estimates from the better performing models show widely scattered distributions. In the case of the propensity model *D3* for y_{blks}, the adjustment increases variability *without* decreasing the deviation to any degree, which ultimately worsens the quality of estimates in an absolute sense.

Next, we examine the importance of including nondemographic (or attitudinal) variables in the propensity score model by comparing four different models: all demographic covariates (*D1*), all nondemographic covariate (*N1*), all covariates (*A1=D1+*

N1), and all demographic and important nondemographic covariates (*4*). Again, Table 7.3 shows the variables included in each model. The distributions of the adjusted estimates using these models are displayed in Figure 7.3 along with those of the reference sample estimates ($y.R$) and the unadjusted estimates ($y.U$) (see Appendix 5 for all box plots of the estimates using all propensity models for both study variables in both $s^{W.ST}$ and $s^{W.HI}$).

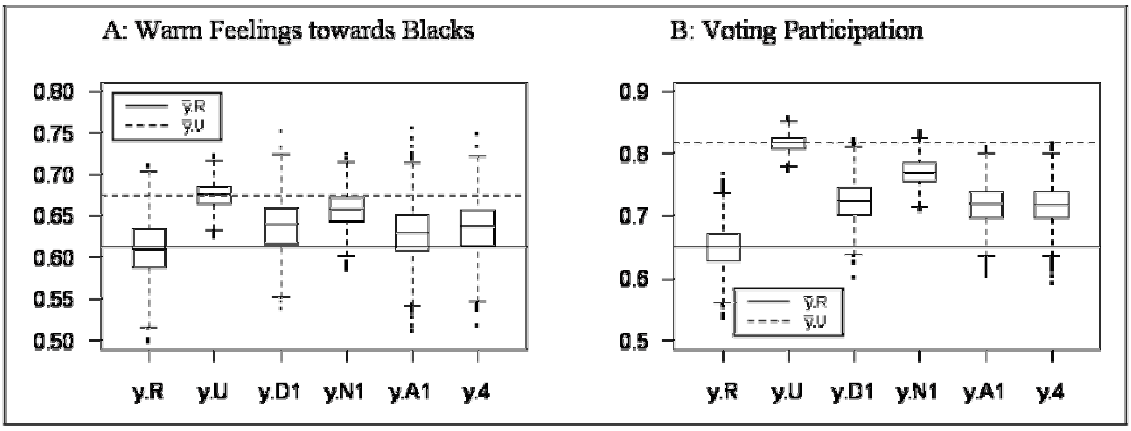


Figure 7.3. Distributions of the Web Estimates by Different Propensity Score Adjustments

For both study variables, the reference sample estimates ($y.R$) are more widely distributed than the unadjusted Web sample estimates ($y.U$). This is not surprising since the size of the Web samples is four times larger than the reference samples. However, the distributions of $y.U$ do not contain the simulation means of $y.R$. For y_{vote}, the distribution of $y.U$ and $y.R$ are almost non-overlapping. Among the four adjustment models, ones including demographic variables (*D1*, *A1* and *4*) produce less biased Web estimates than ones only with nondemographics (*N1*). The marginal effect of including

nondemographic variables in addition to demographic ones can be seen by comparing the box plots for *A1* with *D1*. Figure 7.3 shows that this effect is minimal, since the performance of *A1* and *D1* are comparable. Although the distributions of the adjusted estimates differ noticeably, none of the methods successfully removes the deviation.

As in Figure 7.1 and 7.2, the variance of the Web sample estimates increases when adjustment weights are applied and when the adjustments are more effective in reducing deviation. The increase in variance is primarily due to including the demographic covariates in the propensity models. This may be translated into the significance of these covariates in predicting the propensity score, $e(\mathbf{x}) = \Pr\left(i \in s^W \mid \mathbf{x}, i = 1,...,n\right)$. However, it should be noted that the variability of effective model estimates can be as large as that of the reference sample estimates, meaning that the precision obtained from the larger sample size in Web surveys may be completely lost.

7.2.3.3 Discussion

This section illustrates the exclusive application of propensity score adjustment for volunteer panel Web surveys. The adjustment decreases but does not eliminate the difference between the benchmark sample estimates and the Web sample estimates. This reduction comes at the cost of increased variance. The relationship between the covariates and the study variables is found to be important in forming propensity models, since the propensity models with weakly predictive covariates do not decrease the deviation but add to the variability. It seems to be a reasonable practice to include all available covariates from the given data set, as Rubin and Thomas (1996) suggest. The assertion that including nondemographic variables in the propensity models is useful is

not verified, as the value of including nondemographic variables appears limited in comparison to demographic ones. This may be due to the nature of the two study variables, warm feelings towards Blacks and voting behavior, which are highly correlated to demographic variables, such as race and education. Web sample estimates are compared to reference sample estimates, which may also be contaminated by sampling and nonresponse error. It seems to be a logical approach to combine the propensity score adjustment weights with additional weights that project the adjusted Web samples to the general population. For example, calibration adjustment using general regression estimation proposed in the previous chapter may be an alternative. The combination of the two weights may reduce selection bias in Web surveys to a greater degree. This will be examined in the following section.

7.2.4 Calibration Adjustment

In this section, we apply calibration adjustment as described in Section 6.3 using the propensity score adjusted weights as the starting point. More specifically, the weight in (6.16) is applied as in (6.15) in order to correct for remaining discrepancies between the propensity score adjusted Web sample estimates and the population values. This procedure makes the Web sample covariates that are already balanced to the probabilistically drawn reference sample further balanced to the target population. While the propensity score adjustment is on the reference sample level to attempt to correct for the nonprobability nature of Web samples, the calibration adjustment is on the population level for noncoverage and nonresponse problems in survey samples (refer to Figure 6.2).

Two different sets of covariates are respectively used in calibrating adjustment weights for each of y_{blks} and y_{vote}. For y_{blks}, the first calibration (Calibration *1*) uses *age*,

educ, *race*, *gender*, *region*, and *ethnofit* listed in Table 7.2, and the second (Calibration *2*) uses all but *ethnofit*. For y_{vote}, the first (Calibration *1*) includes *age*, *educ*, *race*, *gender*, *region*, and *party*, whereas the second (Calibration *2*) excludes *party*. The adoption of different sets of covariates in the adjustment is to assess the advantage of calibration – that is, including estimated population totals in addition to the known population values as the benchmarks in the adjustment. The first models will show the marginal effect of incorporating rather unconventional variables (*ethnofit* and *party*) in the adjustment. The **R** code for the calibration using linear distance function with trimmed upper and lower bounds is **`cal.fcn`** and the simulation is done over 2,000 iterations using **`cal.sim`** in Appendix 1.6.

7.2.5 Results of Calibration Adjustment

Adjustments are focused on $s^{W.HI}$ from this section on. For brevity, four different propensity models (*A1, A2, A3, 4*) will be combined with the two calibration adjustments (Calibration *1* and *2*) – resulting in 15 combinations of adjustment (= 5 (4 propensity score models + no propensity score adjustment) x 3 (2 calibrations + no calibration)). As a notational convention, we will denote an estimator of the mean by *y*.(propensity score adjustment type).(calibration type). The unadjusted estimator is denoted by *y.U* and the reference sample estimator by *y.R*. For instance, the Web estimate using the *A1* propensity score model and no calibration will be denoted as $y.A1.n$. As in the simulation in Section 7.2.1, each s^R was selected by simple random sampling without replacement with $n^R = 200$ and $s^{W.HI}$ was a Poisson sample of size $n^{W.HI} = 800$. Table

7.6 presents the population values and the summary statistics for estimates from the reference samples, and the unadjusted and adjusted Web samples.

7.2.5.1 Performance of Calibration Adjustment

The benchmarks of calibration adjustment are the population values. Therefore, we need different evaluation criteria than when propensity score adjustment alone is used

7.2.5.1.A Root Mean Square Error and Percent Root Mean Square Error Reduction

Since we have a fixed known value from the population, the first evaluation criterion is root mean square error (*rmse*) calculated as follows:

$$rmse\left(\bar{y}\right)=\sqrt{\sum_{m=1}^{M}\left(y_m-\bar{Y}\right)^2\Big/M}\,, \tag{7.3}$$

where y_m is the sample estimate from the m^{th} simulation in (7.1) and $\bar{Y}$ is the full pseudo-population mean. The magnitude of *rmse* reduction achieved in adjustment, compared to no adjustment can be compared across different adjustment methods and sets of covariates by percent root mean square error reduction ($p.rmse$):

$$p.rmse\left(\bar{y}^{W.A}\right)=\left[\frac{rmse\left(\bar{y}^{W.U}\right)-rmse\left(\bar{y}^{W.A}\right)}{rmse\left(\bar{y}^{W.U}\right)}\right]\times 100\,, \tag{7.4}$$

where $\bar{y}^{W.A}$ and $\bar{y}^{W.U}$ are the adjusted and unadjusted Web survey estimates, respectively. The larger the $p.rmse$, the smaller the error in $\bar{y}^{W.A}$. A negative $p.rmse$ indicates that the error is increased by the adjustment.

7.2.5.1.B Bias and Percent Bias Reduction

The error component that the propensity score and calibration adjustments attempt to decrease is the bias, which is the difference between the expected value of the sample estimate and the population value, that is,

$$bias\left(\bar{y}\right) = E\left(y\right) - \bar{Y} \,. \tag{7.5}$$

From one realization of samples, biases cannot be estimated from (7.5) because the expected sample estimate, $E\left(y\right)$, is not available. However, simulation makes (7.1) approximate the expected Web sample estimate. As usual, the square of the *rmse* can be decomposed into two components: *bias* squared and variance (*var*),

$$rmse\left(\bar{y}\right) = \sqrt{var\left(\bar{y}\right) + \left[bias\left(\bar{y}\right)\right]^2}$$

and produces

$$bias\left(\bar{y}\right) = \sqrt{\left(rmse\left(\bar{y}\right)^2 - var\left(\bar{y}\right)\right)} \,. \tag{7.6}$$

The standardized measure of bias reduction achieved by the adjustment is percent bias reduction (*p.bias*). This is computed like (7.2) as

$$p.bias\left(\bar{y}^{W.A}\right) = \left[\frac{bias\left(\bar{y}^{W.U}\right) - bias\left(\bar{y}^{W.A}\right)}{bias\left(\bar{y}^{W.U}\right)}\right] \times 100 \,. \tag{7.7}$$

Just like *p.rmse*, a larger *p.bias* indicates that the adjustment performed accomplishes bias reduction to a greater degree, and a negative *p.bias* indicates that the adjustment increases the bias.

7.2.5.1.C Standard Error and Percent Root Standard Error Reduction

The variability of estimates is measured by standard error (se) calculated as the following:

$$se\left(\bar{y}\right)=\sqrt{var\left(\bar{y}\right)}\,. \tag{7.8}$$

The adjusted estimates are expected to have larger standard errors than the unadjusted ones as the weights in the adjustment are likely to introduce extra variability in the estimates. The impact of adjustment on the variability can be measured with percent standard error increase ($p.se$):

$$p.se\left(\bar{y}^{W.A}\right)=\left[\frac{se\left(\bar{y}^{W.A}\right)-se\left(\bar{y}^{W.U}\right)}{se\left(\bar{y}^{W.U}\right)}\right]\times 100\,. \tag{7.9}$$

Since the estimates with a smaller variability are considered better, the estimates with smaller $p.se$ are preferred. A negative $p.se$ indicates that the variance is decreased by the adjustment.

From Table 7.6, it is observed that calibration adjustment applied to the propensity score adjusted estimates improves the accuracy of the Web estimates, as the bias reduction is larger than when propensity score adjustment alone is used. The effectiveness of combining calibration is striking when the propensity score adjustment alone is not successful. For example, in *A3*, the improvement by adding calibration is more apparent. Among the two calibration methods, Calibration *1* that includes a substantive variable shows better bias reduction than Calibration *2*.

Table 7.6. Comparison of Population Values, Reference Sample Estimates and Web Sample Estimates for y_{blks} and y_{vote}

	estimate	*rmse*	*bias*	*se*	*p.rmse*	*p.bias*	*p.se*
y_{blks}							
y.pop	0.614	-	-	-	-	-	-
y.R	0.612	0.034	-0.002	0.034	-	-	-
y.U	0.675	0.064	0.062	0.016	0.0%	0.0%	0.0%
y.A1.n	0.629	0.036	0.016	0.032	44.2%	74.7%	103.5%
y.A1.1	0.621	***0.033***	***0.008***	0.032	***47.5%***	***87.1%***	107.0%
y.A1.2	0.625	0.035	0.012	0.033	45.3%	81.0%	109.1%
y.A2.n	0.642	0.043	0.029	0.032	33.1%	53.7%	101.4%
y.A2.1	0.632	0.036	0.018	0.032	42.8%	70.5%	101.2%
y.A2.2	0.636	0.039	0.022	0.032	39.2%	64.2%	102.8%
y.A3.n	0.669	0.059	0.055	***0.021***	6.9%	10.3%	***35.7%***
y.A3.1	0.638	0.037	0.024	0.028	41.7%	60.6%	79.2%
y.A3.2	0.647	0.043	0.033	0.028	32.0%	45.9%	76.2%
y.4.n	0.635	0.038	0.021	0.032	39.7%	65.6%	104.0%
y.4.1	0.626	0.035	0.012	0.032	45.5%	79.9%	106.7%
y.4.2	0.630	0.037	0.016	0.033	42.7%	73.7%	108.6%
y_{vote}							
y.pop	0.648	-	-	-	-	-	-
y.R	0.650	0.034	0.002	0.034	-	-	-
y.U	0.817	0.169	0.169	0.013	0.0%	0.0%	0.0%
y.A1.n	0.718	0.078	0.070	0.032	54.3%	58.3%	151.2%
y.A1.1	0.713	0.072	0.066	0.030	57.6%	61.2%	130.8%
y.A1.2	0.715	0.074	0.067	0.031	56.6%	60.6%	142.3%
y.A2.n	0.716	0.075	0.068	0.032	55.7%	59.8%	148.9%
y.A2.1	0.711	***0.069***	***0.063***	0.030	***59.0%***	***62.8%***	130.0%
y.A2.2	0.712	0.071	0.064	0.031	58.1%	62.1%	140.0%
y.A3.n	0.818	0.171	0.170	***0.014***	-0.7%	-0.7%	***10.2%***
y.A3.1	0.755	0.109	0.107	0.022	35.7%	36.9%	74.7%
y.A3.2	0.766	0.120	0.118	0.022	29.0%	30.0%	74.9%
y.4.n	0.718	0.077	0.070	0.032	54.6%	58.7%	150.4%
y.4.1	0.712	0.071	0.064	0.030	58.2%	61.9%	129.6%
y.4.2	0.714	0.073	0.066	0.031	57.2%	61.2%	141.5%

Note: The figure for the best estimate (excluding *y.R* and *y.U*) is highlighted in bold/Italic in each column.

However, calibration does tend to increase the standard errors compared to the unadjusted estimates. Figure 7.4 plots *p.bias* against *p.se* from all Web sample estimates and depicts the trade-off between two – a surprisingly clear positive

relationship. The fitted linear regression shows a high capability of explaining the variability. This again confirms the earlier finding that the increased accuracy from the adjustment comes at the cost of increased variability.

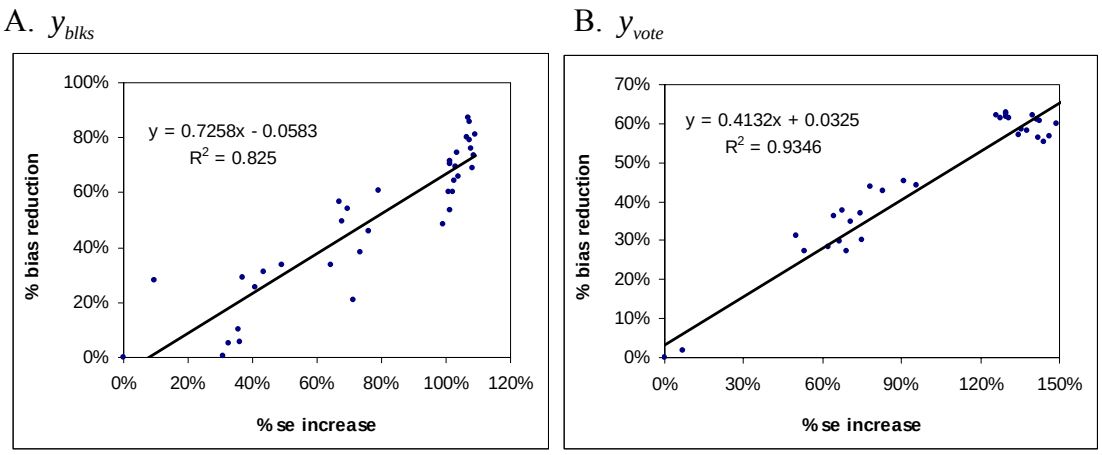


Figure 7.4. Relationship between Percent Bias Reduction and Percent Standard Error Increase in Unadjusted and Adjusted Web Sample Estimates

7.2.5.2 Discussion

The combination of propensity score adjustment and calibration adjustment seems to serve the aim of adjustment better than using only propensity score adjustment. Three things can be improved in the subsequent case study. First, the degrees of bias decrease and variability increase due to the adjustments are only speculated in this section. A statistical test is needed to verify the extent to which the inference of this argument holds. Second, the significance of the covariates in this section is examined only in relation to the substantive study variables, *y*, not to the treatment variable, *g*. Covariate and model selection may be modified by incorporating both *y* and *g*. It will allow us to examine the role of covariates more extensively. Lastly, the subclassification based on the propensity scores for voting behavior was not completed in 29 out of 2,000 simulations due to

subclasses having zero counts of units from the reference sample. Suppose that the reference sample data were originally collected for the general population and that only a subset (like veterans of the military) were to be used because the Web survey target population was a subgroup of the general population. In this case, one may have a small number of cases in the reference sample for that particular Web survey. Therefore, the reference survey should have a large enough size so that the reference samples for any Web survey target populations will have sufficient number of observations for forming the quintile subclassification.

7.3 Case Study 2: Application of Propensity Score Adjustment and Calibration Adjustment to 2003 Michigan Behavioral Risk Factor Surveillance Survey Data

7.3.1 Construction of Pseudo-population and Sample Selection for Simulation

More elaborate examination of the adjustment is carried out in this case study with the data from the 2003 Michigan Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is a collaborative project of the Centers for Disease Control and Prevention and U.S. states, Washington, D.C., and territories and is designed to measure behavioral risk factors in the adult population (18 years of age or older) living in households (CDC, 1998). The 2003 Michigan BRFSS data consist of four quarterly data sets collected by the Institute for Public Policy and Social Research at Michigan State University. The respondents were selected by random digit dialing method with disproportionate allocation for strata defined by geographic area, phone bank density, and probability of each phone number being listed (Michigan Department of Community

Health, 2003). The original 2003 Michigan BRFSS data contain 3,551 units. Among them, 3,410 cases without item nonresponse on Web access ownership and the four stratifying variables (age, gender, race and education) are retained for the study to form the pseudo-population data.

Table 7.7. Distribution of Age, Gender, Education and Race of BRFSS Full Sample, BRFSS Web User and Harris Interactive Survey Respondents

		High School or Less		*Some College or above*	
		White	***NonWhite***	***White***	***NonWhite***
A. BRFSS Full Sample (n=3,410)					
≤ 40 yrs	Female	5.01%	1.35%	10.32%	2.35%
	Male	3.58%	1.09%	6.48%	1.23%
41 yrs +	Female	16.57%	2.49%	20.29%	2.23%
	Male	10.29%	1.17%	14.13%	1.41%
			Sum	100%	
B. BRFSS Web Users (n=1,250)					
≤ 40 yrs	Female	5.28%	0.83%	13.98%	2.22%
	Male	3.29%	0.83%	8.70%	1.44%
41 yrs +	Female	10.56%	0.97%	23.29%	2.08%
	Male	7.18%	0.65%	17.18%	1.53%
			Sum	100%	
C. Harris Interactive Respondents (n=8,195)					
≤ 40 yrs	Female	2.03%	1.64%	13.28%	13.37%
	Male	0.85%	0.61%	7.58%	9.09%
41 yrs +	Female	2.45%	0.48%	15.58%	4.58%
	Male	1.70%	0.24%	20.82%	5.71%
			Sum	100%	

Table 7.7 compares the distribution of the four stratifying variables among all respondents in BRFSS, Web users in BRFSS and Harris Interactive survey respondents (the same as in Case Study 1). The results from the comparison echo what has been observed previously. Harris Interactive survey respondents over-represent Nonwhites and more educated and younger people compared to BRFSS respondents. This tendency remains even when the Web access owners in BRFSS are compared to the Harris

Interactive survey respondents. When interpreting Table 7.7, one should bear in mind that the target population of BRFSS represents Michigan residents, while the Harris Interactive survey targets the general U.S. population. Even so, since the purpose of this study is not to show the discrepancy between the two data sets but to investigate whether this discrepancy can be reduced by the proposed statistical adjustment, this should not degrade the value of the study. The table shows that there are still considerable gaps in the distributions of education and race between the two sets of BRFSS respondents and the HI respondents.

As in Case Study 1, a BRFSS pseudo-population is created by bootstrapping the 3,410 BRFSS respondents with replacement for the size of 20,000. Among the pseudo-population, 12,674 people indicated that they have Web access at home[19] and these will be considered as the Web pseudo-population. This results in 63.4% of Web access owners in the pseudo-population, which is very close to 63.3% (=2,160/3,410) Web owners in the original BRFSS data.

The aims of Case Study 2 are slightly different than those of the previous case study. First, the emphases are placed on the propensity score model development in a more practical situation. Second, variance estimation methods are examined for estimates calculated with propensity score adjustment weights and/or calibration weights. Third, the effectiveness of different propensity score models and calibration methods is assessed with significance tests. Therefore, Case Study 2 uses slightly different simulation functions than the first case study.

[19] Question Wording: "Do you have access to the Internet at home?"

Using the full pseudo-population data, a reference sample of size 500 is drawn from the full pseudo-population using **`ref.sam`**. For the Web samples, instead of drawing different types of Web samples, samples resembling Harris Interactive survey respondents, i.e. $s^{W.HI}$, are examined. Web samples of size 1,500 are drawn from the Web pseudo-population with an allocation proportional to the stratum distribution for Harris Interactive respondents in Table 7.7 using **`pois.sam`**. The sample sizes here are larger than the ones used in the first case study in order to avoid situations where weighting based on propensity score adjustment becomes impossible due to zero observations in subclasses in (6.1). These samples are drawn in 3,200 simulations.[20]

7.3.2 Adjustments

7.3.2.1 Propensity Score Adjustment

Case study 1 used different PSA models for different variables. However, this type of modeling is unlikely to be exercised in a real setting, because it requires different weights for each study variable when estimating more than one variable. In practice, one propensity model is likely to be applied to derive weights for all study variables. In order to implement propensity score adjustment, the following modeling method is used in this study. First, one reference sample of size 500 and one Web sample of size 1,500 are drawn as described previously. Then, these two samples are merged into one data set of 2,000. The base propensity score model is constructed from this merged data based on the relationship between g and $\mathbf{x}$, not between y and $\mathbf{x}$ as in Case Study 1. Five different logistic models are used for propensity score adjustment as in (7.10). For the

[20] The number of simulation is increased in this case study in order to compute the variance and the confidence interval more reliably.

base model (Model *2*), the vector of $\mathbf{x}_i$ for person *i* includes all 30 covariates listed in Table 7.8, such that

$$\ln\left[\frac{\Pr(g_i=1)}{1-\Pr(g_i=1)}\right]=\alpha+\mathbf{B}'\mathbf{x}_i, \tag{7.10}$$

where $i=1,...,n$, $\mathbf{B}$ and $\mathbf{x}_i$ are 30×1 vectors, and *n* is the total number of cases in the merged data set. Model *1*, *3*, *4*, and *5* use subset of the covariates in Table 7.8. Model 3 retains marginally significant covariates with $p\le0.2$ in Model *2* and, thus, tests the role of significant covariates in predicting *g* in propensity score models. In order to detect the marginal effect of stratifying variables used in the sampling stage, Model *1*, *4* and *5* are constructed. Model *1* includes stratifying variables only; Model *4* excludes variables in Model *1* from Model *2* (i.e., Model *4* uses all variables except the stratifiers); and Model *5* excludes variables in Model *1* from Model *3* (i.e., Model *5* includes covariates significant at 20% but excludes the stratifiers). Details of these models are shown in Table 7.9.

Table 7.8. List of Covariates Used for Propensity Modeling

Covariate	***Type***	***Description***
Ghealth	Continuous	General Health (1: excellent, 2: very good, 3: good, 4: fair, 5: poor)
Coverage	2 categories	Having health care coverage (1: yes, 2: no)
Doctor	2 categories	Having personal doctor/health care provider (1: yes, 2: no)
cprevent	2 categories	Cost prevented from doctor's visit in the past 12 months (1: yes, 2: no)
phyact	2 categories	Participate in any physical activities other than regular job durin the past month (1: yes, 2: no)
diabete	2 categories	Ever told to have diabetes by a doctor (1: yes, 2: no)
cholest	2 categories	Ever checked blood cholesterol (1: yes, 2: no)
losewgt	2 categories	Trying to lose weight (1: yes, 2: no)
wgtadv	2 categories	Weight advice given by health professional in the past 12 months (1: yes, 2: no)

Table 7.8 (continued)

Covariate	*Type*	*Description*
asthma	2 categories	Ever told to have asthma by a doctor (1: yes, 2: no)
flushot	2 categories	Had a flu shot in the past 12 months (1: yes, 2: no)
pneumon	2 categories	Ever had a pneumonia shot (1: yes, 2: no)
sunburn	2 categories	Had a sunburn in the past 12 months (1: yes, 2: no)
age	Continuous	Age in years
educ	Continuous	Education
income	Continuous	Household income
weight	Continuous	Current weight
numphone	Continuous	Number of residential phone lines
gender	2 categories	Gender (1: male, 2: female)
jointsym	2 categories	Had any symptoms of pain, aching, or stiffness around joint in the past 30 days (1: yes, 2: no)
limitact	2 categories	Limited in any activities because of physical, mental or emotional problems (1: yes, 2: no)
modact	2 categories	Moderate activities for at least 10 minutes in a usual week when not working (1: yes, 2: no)
army	2 categories	Ever served on active duty in the United States Armed Forces (1: yes, 2: no)
cellphon	2 categories	Have a cell or mobile phone (1: yes, 2: no)
alcohol	Continuous	Amount of alcohol consumption
hhsize	Continuous	Household size
work	2 categories	Work full time (1: yes, 2: no)
marry	2 categories	Marital status (1: married, 2: others)
race	2 categories	Race (1: Whites, 2: others)
veggie	Continuous	Amount of vegetable consumption

Propensity score Models *1* through *5* are applied in deriving five sets of weights that are used for the three study variables, y_1: whether respondents have high blood pressure (HBP), y_2: whether respondents have smoked 100 or more cigarettes (SMOKE), and y_3: whether respondents do vigorous physical activities (ACT).[21] In

[21] Question wording for the three variables are as follows.
y_1 (HBP): "Have you ever been told by a doctor, nurse or other health professional that you have high blood pressure?"
y_2 (SMOKE): "Have you smoked at least 100 cigarettes in your entire life?"

order to examine the relationship between the covariates in propensity models and the study variables, the same sets of covariates in Model *1* through *5* are fitted to predict y_1, y_2, y_3, and g (whether the unit is from the Web survey sample or the reference survey sample) in the original BRFSS data (n=3,410). The p-value of each covariate in all models is shown in Table 7.9.

The performance of the model predictability is detected using Akaike Information Criteria (*AIC*). *AIC* is computed as $AIC = -2\log\left(\hat{L}\right) + 2K$, where $\hat{L}$ is the likelihood statistic and K is the number of parameters in the model. The smaller the *AIC*, the better fitting the propensity score model. The *AIC* penalizes model complexity by increasing as the number of parameters increase. Not surprisingly, propensity score Model *2* and *3* fit better across four dependent variables than the other models, as Model *2* contains more covariates and Model *3* contains only significant ones. Model 4, which includes all covariates except stratifying variables, is also competitive based on the *AIC*.

---

y_3 (ACT): "Now thinking about the vigorous physical activities you [fill in] in a usual week, do you do vigorous activities for at least 10 minutes at a time, such as running aerobics, heavy yard work, or anything else that causes large increase in breathing or hear rate?"

Table 7.9. Propensity Score Models and P-values of Covariates for Different Dependent Variables

	Dependent Variable			
	g : **WEB**	y_1 : **HBP**	y_2 : **SMOKE**	y_3 : **ACT**
MODEL *1*				
age	0.3218	<0.0001	0.1042	<0.0001
educ	0.0000	0.0065	<0.0001	<0.0001
gender	0.0726	0.6029	<0.0001	<0.0001
race	0.0000	0.0001	0.0160	<0.0001
AIC	2008.8	3757.6	4552.4	4278.3
MODEL *2*				
ghealth	0.1806	<0.0001	0.0004	<0.0001
coverage	0.4073	0.4904	0.0164	0.1199
doctor	0.1045	0.1435	0.5436	0.5270
cprevent	0.0221	0.3360	0.0013	0.3121
phyact	0.3604	0.3266	0.4718	<0.0001
diabete	0.0480	0.0001	0.1825	0.5117
cholest	0.4914	<0.0001	0.9139	0.0063
losewgt	0.0350	0.1837	0.5822	0.0296
wgtadv	0.2986	0.0008	0.8317	0.1817
asthma	0.4106	0.2167	0.9845	0.2333
flushot	0.8168	0.0021	0.0449	0.9012
pneumon	0.3888	0.8610	0.0660	0.1647
sunburn	0.1466	0.2629	0.0117	0.0215
age	0.6221	<0.0001	0.8366	<0.0001
educ	<0.0001	0.6441	<0.0001	0.2376
income	0.0097	0.1335	0.9379	0.1250
weight	0.5240	<0.0001	0.0557	0.1128
numphone	0.4489	0.5027	0.7071	0.4085
gender	0.1632	0.2615	0.0738	<0.0001
jointsym	0.8323	0.3379	0.0012	0.3371
limitact	0.0342	0.3663	0.0508	0.0048
modact	0.4473	0.8883	0.2546	<0.0001
army	0.1326	0.6561	<0.0001	0.6325
cellphon	0.0039	0.0786	0.2519	0.0972
alcohol	0.7995	0.2469	<0.0001	0.0096
hhsize	0.7981	0.0235	0.1913	0.0405
work	0.6905	0.1298	0.7322	0.0339
marry	0.2720	0.3171	0.4568	0.0220
race	<0.0001	0.1219	0.0083	0.1930
veggie	0.7797	0.1457	0.0448	<0.0001
AIC	2004.1	2729.8	3486.2	3139.2

Table 7.9 (continued)

	Dependent Variable			
	g : **WEB**	y_1 : **HBP**	y_2 : **SMOKE**	y_3 : **ACT**
MODEL *3*				
ghealth	0.0880	<0.0001	<0.0001	<0.0001
doctor	0.1940	<0.0001	0.8831	0.4741
cprevent	0.0116	0.2333	<0.0001	0.0052
diabete	0.1087	<0.0001	0.2517	0.0471
losewgt	0.0559	<0.0001	0.3762	0.0006
sunburn	0.3119	<0.0001	0.0085	<0.0001
educ	<0.0001	0.9828	<0.0001	0.0297
income	0.0006	0.0002	0.8408	0.0004
gender	0.1173	0.5012	0.0090	<0.0001
limitact	0.1297	0.0002	0.0219	<0.0001
army	0.1959	<0.0001	<0.0001	0.0009
cellphon	0.0016	0.4797	0.6566	0.0424
race	<0.0001	0.3687	0.0045	0.4833
AIC	1981.1	3312.6	3914.9	3777.2
MODEL *4*				
ghealth	0.0152	<0.0001	<0.0001	<0.0001
coverage	0.1835	0.8748	0.0059	0.0229
doctor	0.3021	0.0526	0.4018	0.7992
cprevent	0.0501	0.8298	0.0031	0.3112
phyact	0.0174	0.4261	0.1576	<0.0001
diabete	0.6122	<0.0001	0.1343	0.3520
cholest	0.8262	<0.0001	0.4369	0.2597
losewgt	0.0540	0.1638	0.3007	0.5248
wgtadv	0.4037	0.0022	0.8598	0.0730
asthma	0.5499	0.0532	0.9058	0.5355
flushot	0.9521	<0.0001	0.0514	0.2514
pneumon	0.2047	0.1082	0.0147	0.0193
sunburn	0.9247	0.0016	0.0009	<0.0001
income	<0.0001	0.0259	0.0783	0.0662
weight	0.6802	<0.0001	0.2382	0.1460
numphone	0.6040	0.2946	0.8867	0.2045
jointsym	0.2648	0.0508	0.0016	0.8521
limitact	0.0031	0.6016	0.1116	0.0079
modact	0.7416	0.4107	0.2555	<0.0001
army	0.0289	0.6521	<0.0001	0.3924
cellphon	0.0001	0.2652	0.0446	0.0875
alcohol	0.1280	0.2824	<0.0001	0.0003
hhsize	0.1190	<0.0001	0.2480	0.0009
work	0.2855	<0.0001	0.7928	<0.0001
marry	0.0879	0.0311	0.9053	0.0014
race	0.3698	0.5276	0.0020	<0.0001
AIC	2147.1	2786.0	3546.2	3198.9

Table 7.9 (continued)

	Dependent Variable			
	g : **WEB**	y_1 : **HBP**	y_2 : **SMOKE**	y_3 : **ACT**
MODEL *5*				
ghealth	0.0005	<0.0001	<0.0001	<0.0001
doctor	0.4131	<0.0001	0.7442	0.0725
cprevent	0.0154	0.2023	<0.0001	0.0140
diabete	0.7052	<0.0001	0.1861	0.0319
losewgt	0.0466	<0.0001	0.1201	0.0082
sunburn	0.8364	<0.0001	0.0004	<0.0001
income	<0.0001	0.0002	0.0173	<0.0001
limitact	0.0122	0.0001	0.0513	<0.0001
army	0.0088	<0.0001	<0.0001	0.5847
cellphon	0.0001	0.5391	0.1275	0.0840
AIC	2136.4	3307.9	3897.2	3819.1

7.3.2.2 Calibration Adjustment

Two different sets of calibration variables are employed to test the effect of including population estimates of substantive variables. The first calibration adjustment projects the weighted Web sample to the pseudo-population with respect to *age*, *gender*, *educ* and *race* (Calibration *1*). This resembles generalized ratio-raking using known population figures. Calibration *2* expands the first one by adding a key health variable, *ghealth*: "Would you say that in general your health is excellent, very good, good, fair or poor?" Although *ghealth* is a rather less traditional variable to be included in calibration, our three study variables are all highly health-related. Therefore, inclusion of *ghealth* in the calibration adjustment is expected to improve the adjustment. The propensity score adjustment weights are calculated in **`psa.fcn`**. These propensity score adjusted weights and base weights are modified in calibrating the sample covariate estimates to the pseudo-population benchmarks using **`newcal.fcn`** (see Appendix 1.7 for the **R**© code).

Applying calibration weights, Web sample estimates on y_1, y_2, and y_3 are calculated in each simulation. The whole simulation is done similarly to **`cal.sim`** in Appendix 1.6.

7.3.3 Results of Adjustments

7.3.3.1 Comparison of Adjusted Estimates

For each study variable, there are population values, reference sample estimates and Web sample estimates. Since reference samples do not require propensity score adjustment, there are three types of estimates reflecting calibration adjustment status: No Calibration, Calibration *1*, and Calibration *2*. For Web samples, there are 18 different combinations of adjustments: (No propensity score adjustment, propensity score Model *1*, *2*, *3*, *4*, and *5*) x (No Calibration, Calibration *1*, and Calibration *2*). The type of adjustment will be denoted after the name of variable. For instance, the unadjusted reference sample estimate of HBP will be denoted as *y1.R.n*; the Web sample estimate of SMOKE using no propensity score adjustment but Calibration *1* as *y2.n.1*; and the estimate of ACT using propensity score Model *5* and Calibration *2* as *y3.5.2*.

Table 7.10 presents the simulation means of the Web sample estimates using all adjustments and the reference sample estimates and the population values. The distribution of the reference sample and Web sample estimates over all simulations are shown in Figure 7.5 using box plots. While the reference sample estimates are distributed around the population values, the Web sample estimates are not necessary so. The unadjusted estimates (*y1.n.n*, *y2.n.n* and *y3.n.n*) are the most biased of all the alternatives. In fact, none of the ranges of unadjusted estimates from 3,200 simulations contains the true values. On average, when no adjustment is applied, people in the Web samples are less likely to have high blood pressure, less likely to have smoked 100 cigarettes and

more likely to do vigorous physical activities than the population. Once the adjustment is incorporated, the discrepancies between the Web estimates and the population figures tend to decrease. Some of the adjusted Web sample estimates are almost unbiased, since estimates, such as *y1.1.n*, *y1.1.2*, *y2.n.2, y2.1.1*, *y2.2.1*, *y2.2.2*, *y2.3.1*, *y2.3.2*, *y2.4.1*, *y2.4.2*, *y2.5.1*, *y2.5.2*, *y3.5.1*, and *y3.5.2* show almost symmetric distributions around the population values. The most striking bias reduction can be observed for SMOKE. When any combinations of propensity score and calibration adjustment are applied, the means of the Web sample estimates for the proportion of people who smoked 100 or more cigarettes are almost right on the population value. However, the introduction of adjustments causes estimates to be more variable, as evidence by larger interquartile ranges. The reduction in variance due to having large sample sizes in Web surveys is offset by the bias corrections.

Table 7.10. Population Values, Reference Sample Estimates and Web Sample Estimates for HBP, SMOKE and ACT

Adjustment Combination	***estimate***		
	y_1 : **HBP**	y_2 : **SMOKE**	y_3 : **ACT**
Pop	0.3201	0.5276	0.4349
R.n	0.3197	0.5278	0.4352
R.1	0.3197	0.5279	0.4351
R.2	0.3197	0.5278	0.4352
n.n	0.2766	0.4722	0.5117
n.1	0.2864	0.5194	0.4758
n.2	0.3022	0.5330	0.4653
1.n	0.3205	0.4985	0.4738
1.1	0.3114	0.5295	0.4660
1.2	0.3197	0.5356	0.4583
2.n	0.3042	0.4998	0.4604
2.1	0.3029	0.5319	0.4525
2.2	0.3050	0.5333	0.4502
3.n	0.2882	0.5012	0.4612
3.1	0.2934	0.5330	0.4524

Table 7.10 (continued)

Adjustment Combination	*estimate*		
	y_1 : **HBP**	y_2 : **SMOKE**	y_3 : **ACT**
3.2	0.2945	0.5337	0.4513
4.n	0.2898	0.4820	0.4575
4.1	0.3079	0.5320	0.4221
4.2	0.2996	0.5260	0.4251
5.n	0.2955	0.4831	0.4733
5.1	0.3102	0.5325	0.4402
5.2	0.3020	0.5269	0.4419

7.3.3.2 Performance of Adjustments on Error Reduction

The mechanisms of error reduction are examined to a greater depth in this section. Table 7.10 summarizes the error properties of all Web sample estimates calculated as in (7.3), (7.4), (7.6), (7.7), (7.8) and (7.9). It also includes standardized error properties, such as standardized *rmse* (*s.rmse*), standardized *bias* (*s.bias*), and standardized *se* (*s.se*). These are defined as *rmse* , *bias* , and *se* divided by the simulation mean in (7.1). These standardized error and the percentage figures allow unit-free comparisons on the magnitude of error reduction across all variables.

As discussed in the previous section, there is a notable reduction in bias by using propensity score and calibration adjustment. The reduction in bias is achieved at 68.2% on average, ranging from 17.7% for *y.4.n* of SMOKE to 99.2% for *y.1.n* for HBP. When both propensity score and calibration adjustment are employed, the average bias reduction is realized at 78.8%. Propensity score adjustment or calibration adjustment alone does not seem to remove bias as much as when the two are combined.

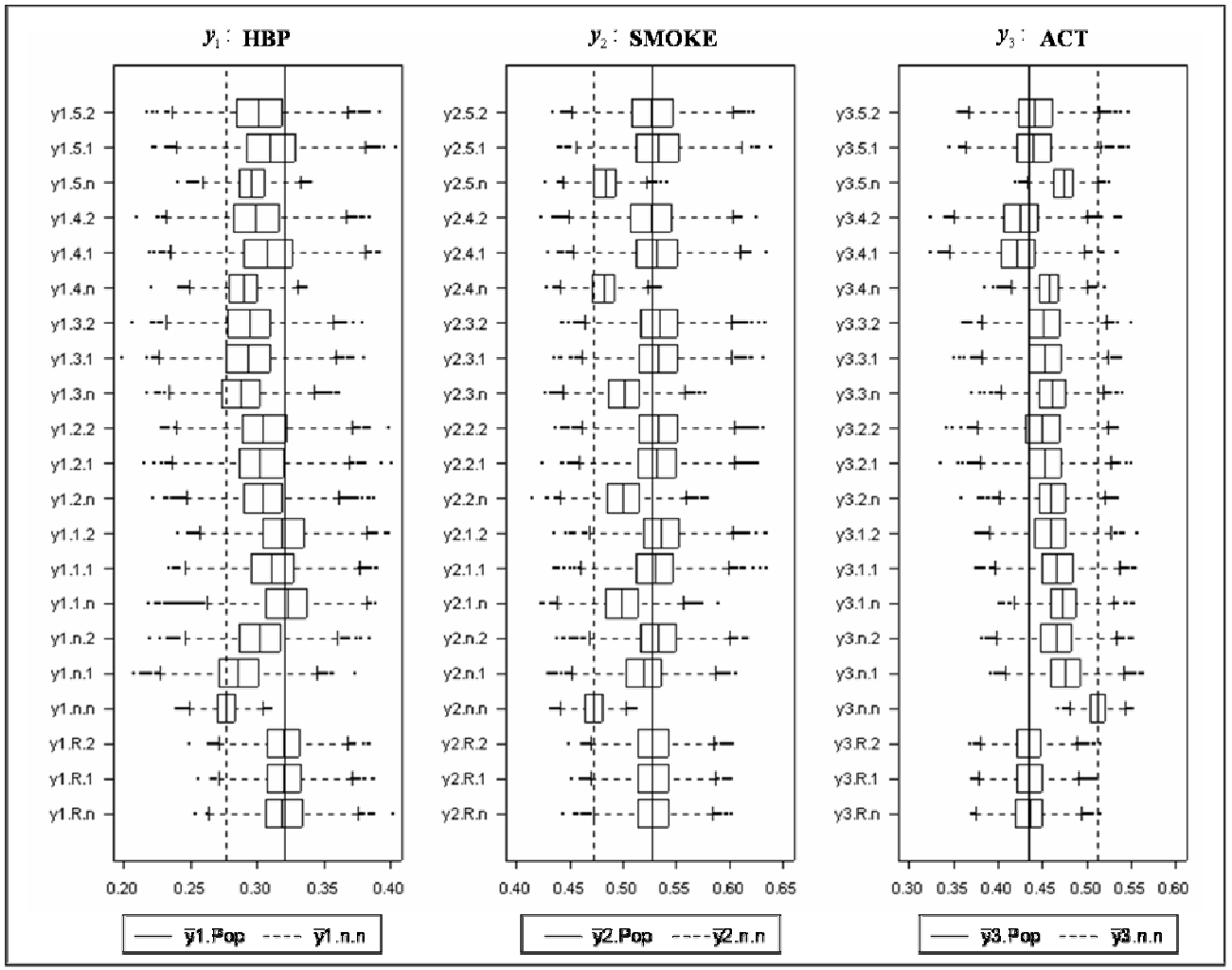


Figure 7.5. Simulation Means of All Web Sample Estimates and Reference Sample Estimates and Population Values

Although we utilize the adjustments to reduce bias, it is worthwhile to examine how they change the overall error structure. This is done by comparing *rmse*, *s.rmse* and *p.rmse* among different adjustment methods presented in Table 7.11.A, Table 7.11.B and Table 7.11.C. The magnitude of *rmse* reduction is smaller than that of bias reduction. While the adjustment on SMOKE decreases the bias as much as by 98.8%, the reduction in *rmse* is only half of that, although still substantial compared to no adjustment.

Table 7.11.A. Error Properties of Reference Sample and Web Sample Estimates for Proportion of People with High Blood Pressure

Adjustment Combina-tion	***rmse***	***s.rmse***	***p.rmse***	***bias***	***s.bias***	***p.bias***	***se***	***s.se***	***p.se***
y_1 : **HBP**									
R.n	0.0204			-0.0004			0.0204		
R.1	0.0192			-0.0004			0.0192		
R.2	0.0188			-0.0004			0.0188		
n.n	0.0448	0.1618	-	-0.0435	-0.1574	-	0.0104	0.0376	-
n.1	0.0404	0.1412	9.7%	-0.0337	-0.1178	22.5%	0.0223	0.0778	114.4%
n.2	0.0286	0.0946	36.1%	-0.0179	-0.0591	59.0%	0.0223	0.0739	114.8%
1.n	***0.0247***	***0.0770***	***44.9%***	***0.0004***	***0.0012***	***99.2%***	0.0247	0.0770	137.3%
1.1	0.0259	0.0832	42.1%	-0.0087	-0.0279	80.1%	0.0244	0.0784	134.8%
1.2	0.0232	0.0727	48.1%	-0.0004	-0.0012	99.1%	0.0232	0.0727	123.6%
2.n	0.0267	0.0878	40.3%	-0.0159	-0.0523	63.5%	0.0215	0.0705	106.4%
2.1	0.0302	0.0998	32.4%	-0.0172	-0.0570	60.4%	0.0248	0.0820	139.0%
2.2	0.0284	0.0930	36.7%	-0.0151	-0.0496	65.2%	0.0240	0.0786	130.6%
3.n	0.0380	0.1319	15.1%	-0.0319	-0.1107	26.7%	0.0206	0.0716	98.6%
3.1	0.0362	0.1233	19.2%	-0.0267	-0.0909	38.7%	0.0244	0.0833	135.1%
3.2	0.0347	0.1180	22.4%	-0.0256	-0.0869	41.2%	0.0235	0.0798	126.1%
4.n	0.0339	0.1169	24.3%	-0.0303	-0.1045	30.5%	0.0152	0.0524	46.1%
4.1	0.0294	0.0955	34.3%	-0.0122	-0.0395	72.0%	0.0268	0.0870	157.7%
4.2	0.0326	0.1087	27.2%	-0.0205	-0.0685	52.9%	0.0253	0.0845	143.5%
5.n	0.0283	0.0956	36.8%	-0.0246	-0.0831	43.6%	***0.0140***	***0.0473***	***34.6%***
5.1	0.0284	0.0915	36.6%	-0.0099	-0.0320	77.2%	0.0266	0.0858	155.9%
5.2	0.0310	0.1025	30.8%	-0.0181	-0.0600	58.4%	0.0251	0.0832	141.7%

Note: The figure for the best estimate (excluding *y.R* and *y.U*) is highlighted in bold/Italic in each column.

Table 7.11.B. Error Properties of Reference Sample and Web Sample Estimates for Proportion of People Who Smoked 100 Cigarettes or More

Adjustment Combination	*rmse*	*s.rmse*	*p.rmse*	*bias*	*s.bias*	*p.bias*	*se*	*s.se*	*p.se*
y_2 : **SMOKE**									
R.n	0.0222			0.0003			0.0222		
R.1	0.0218			0.0003			0.0218		
R.2	0.0217			0.0003			0.0217		
n.n	0.0566	0.1199	-	-0.0554	-0.1172	-	0.0119	0.0251	-
n.1	0.0265	0.0510	53.2%	-0.0082	-0.0158	85.2%	0.0252	0.0485	112.4%
n.2	***0.0256***	***0.0479***	***54.9%***	0.0055	0.0103	90.1%	0.0250	0.0468	110.4%
1.n	0.0368	0.0738	35.0%	-0.0290	-0.0582	47.6%	0.0226	0.0454	90.7%
1.1	0.0267	0.0504	52.9%	0.0020	0.0037	96.4%	0.0266	0.0502	124.1%
1.2	0.0272	0.0509	51.9%	0.0081	0.0151	85.4%	0.0260	0.0486	119.3%
2.n	0.0356	0.0712	37.2%	-0.0278	-0.0556	49.8%	0.0222	0.0444	87.0%
2.1	0.0278	0.0523	50.9%	0.0044	0.0082	92.1%	0.0275	0.0517	131.6%
2.2	0.0279	0.0522	50.8%	0.0058	0.0108	89.6%	0.0273	0.0511	129.7%
3.n	0.0341	0.0681	39.7%	-0.0264	-0.0526	52.4%	0.0217	0.0433	82.8%
3.1	0.0273	0.0512	51.8%	0.0054	0.0102	90.2%	0.0267	0.0502	125.3%
3.2	0.0270	0.0506	52.3%	0.0061	0.0115	88.9%	0.0263	0.0493	121.9%
4.n	0.0481	0.0997	15.1%	-0.0455	-0.0945	17.7%	0.0154	0.0320	29.9%
4.1	0.0291	0.0547	48.6%	0.0045	0.0084	91.9%	0.0288	0.0541	142.6%
4.2	0.0285	0.0541	49.7%	-0.0015	-0.0029	97.2%	0.0284	0.0541	139.7%
5.n	0.0468	0.0969	17.3%	-0.0444	-0.0919	19.8%	***0.0148***	***0.0305***	***24.4%***
5.1	0.0289	0.0543	48.9%	0.0049	0.0093	91.1%	0.0285	0.0535	140.2%
5.2	0.0280	0.0531	50.6%	***-0.0007***	***-0.0013***	***98.8%***	0.0279	0.0530	135.6%

Note: The figure for the best estimate (excluding *y.R* and *y.U*) is highlighted in bold/Italic in each column.

Table 7.11.C. Error Properties of Reference Sample and Web Sample Estimates for Proportion of People Who Do Vigorous Physical Activities

Adjustment Combination	*rmse*	*s.rmse*	*p.rmse*	*bias*	*s.bias*	*p.bias*	*se*	*s.se*	*p.se*
y_3 : **ACT**									
R.n	0.0220			0.0003			0.0220		
R.1	0.0211			0.0002			0.0211		
R.2	0.0207			0.0003			0.0207		
n.n	0.0777	0.1518	-	0.0768	0.1501	-	0.0116	0.0227	-
n.1	0.0478	0.1005	38.5%	0.0409	0.0859	46.8%	0.0248	0.0522	113.5%
n.2	0.0394	0.0846	49.3%	0.0304	0.0654	60.4%	0.0250	0.0537	114.8%
1.n	0.0443	0.0935	43.0%	0.0389	0.0821	49.3%	0.0211	0.0446	81.9%
1.1	0.0407	0.0872	47.7%	0.0311	0.0668	59.5%	0.0261	0.0561	124.8%
1.2	0.0348	0.0759	55.2%	0.0234	0.0510	69.6%	0.0258	0.0562	121.6%
2.n	0.0340	0.0738	56.3%	0.0255	0.0554	66.8%	0.0225	0.0488	93.1%
2.1	0.0326	0.0721	58.0%	0.0176	0.0389	77.1%	0.0275	0.0607	136.3%

Table 7.11.C (continued)

Adjustment Combination	*rmse*	*s.rmse*	*p.rmse*	*bias*	*s.bias*	*p.bias*	*se*	*s.se*	*p.se*
2.2	0.0310	0.0689	60.1%	0.0153	0.0341	80.0%	0.0270	0.0599	131.9%
3.n	0.0337	0.0731	56.6%	0.0263	0.0570	65.8%	0.0211	0.0458	81.7%
3.1	0.0315	0.0697	59.4%	0.0175	0.0386	77.2%	0.0263	0.0580	125.9%
3.2	0.0306	0.0678	60.6%	0.0164	0.0363	78.7%	0.0259	0.0573	122.3%
4.n	***0.0278***	***0.0608***	***64.2%***	0.0226	0.0494	70.6%	0.0163	0.0356	39.9%
4.1	0.0314	0.0744	59.6%	-0.0128	-0.0304	83.3%	0.0286	0.0678	146.3%
4.2	0.0301	0.0708	61.2%	-0.0098	-0.0229	87.3%	0.0285	0.0670	145.1%
5.n	0.0411	0.0868	47.1%	0.0384	0.0811	50.0%	***0.0147***	***0.0310***	***26.3%***
5.1	0.0285	0.0649	63.3%	***0.0053***	***0.0120***	***93.1%***	0.0281	0.0637	141.3%
5.2	0.0286	0.0647	63.2%	0.0070	0.0158	90.9%	0.0277	0.0627	138.4%

Note: The figure for the best estimate (excluding *y.R* and *y.U*) is highlighted in bold/Italic in each column.

The smaller degree of the *rmse* reduction than bias reduction occurs because adjustment weights add variability in the estimates as they attempt to decrease discrepancies between the Web sample covariate distributions and their desired population distributions. The base weight, when no adjustment is made, is the same for every unit in the Web sample. As adjustments are made, weights diverge from the base weight. The divergence becomes even larger when the adjustments correct for large discrepancies. Recall Table 7.7 which showed a sizable discrepancy in some of the covariates between the population and volunteer panel Web survey respondents. Therefore, it would not be surprising to see the variation in weights after applying the adjustments as shown in Table 7.12. The base weight starts at 13.34. Once the adjustment is applied, the upper and lower boundaries diverge from the base weight radically. The ratio of the largest and the smallest weight from the same adjustment ranges from 1 to 89.7.

Table 7.12. Distribution of Weights for All Adjustments over All Simulations

Adjustment Combinations	*Lower Bound*	*Upper Bound*	*Ratio (Upper/Lower)*
n.n	13.34	13.34	1.00
n.1	6.68	144.46	21.62
n.2	6.67	178.23	26.72
1.n	6.13	63.47	10.35
1.1	3.91	133.78	34.21
1.2	3.35	160.60	47.94
2.n	5.41	62.76	11.60
2.1	3.15	147.26	46.77
2.2	2.78	159.62	57.49
3.n	5.85	62.24	10.64
3.1	3.48	135.55	38.93
3.2	3.04	147.12	48.38
4.n	6.29	34.53	5.49
4.1	3.15	260.52	82.72
4.2	3.14	282.13	89.73
5.n	7.24	31.56	4.36
5.1	3.63	250.35	69.05
5.2	3.62	273.28	75.53

Figure 7.6 shows the relationship between the decrease in bias and the increase in variability when adjustment is applied to the Web survey estimates. As was true in simulation in Section 7.2.4, this figure shows that the bias reduction is generally achieved at the cost of the standard error increase. The correlation between the two is .61 (not shown in the figure). The linear regression also indicates a fairly strong relationship between the two statistics.

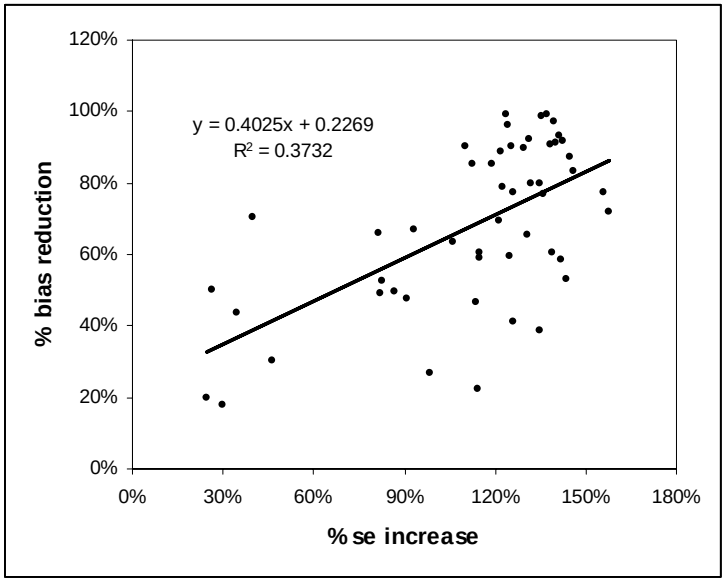


Figure 7.6. Relationship between Percent Bias Reduction and Percent Standard Error Increase in Adjusted Web Sample Estimates

We examine the effectiveness of propensity score adjustment and calibration adjustment using analyses of variance (ANOVA). ANOVA models are used to predict *p.rmse*, *p.bias*, and *p.se* as functions of two main effects (PSTATUS: propensity score adjustment status – whether or not propensity score adjustment is used; and CSTATUS: calibration adjustment status – whether or not calibration adjustment is used) and their interaction. All three ANOVA models are significant with $F = 11.89$, 20.66 and 65.34 ($df = 3/50$; $p < 0.0001$). Both propensity score adjustment status and calibration adjustment status have significant effects on *p.rmse*, *p.bias* and *p.se* with $p < 0.0001$. Their interactions are also significant in explaining the variances of all three error properties with p-values of 0.0045, 0.0317 and 0.0021, respectively.

Table 7.13. Least Square Mean of Percent Root Mean Square Error Reduction, Percent Bias Reduction and Percent Standard Error Increase by Propensity Score Adjustment Status, Calibration Adjustment Status and Their Interactions

	LS Mean		
Effect	***p.rmse***	***p.bias***	***p.se***
Propensity Score Adjustment (PSTATUS)			
PSA (P1)	**42.9%** [P2]	**64.5%** [P2]	102.5% [P2]
No PSA (P2)	20.1% [P1]	30.3% [P1]	**56.7%** [P1]
Calibration Adjustment (CSTATUS)			
CAL (C1)	**43.9%** [C2]	**69.7%** [C2]	123.9% [C2]
No CAL (C2)	19.1% [C1]	25.1% [C1]	**35.4%** [C1]
Interaction (PSTATUS*CSTATUS)			
P1*C1 (1)	**47.5%** [4]	**78.8%** [2,4]	134.4% [2,4]
P1*C2 (2)	38.2% [4]	50.2% [1,4]	70.7% [1,3,4]
P2*C1 (3)	40.3% [4]	60.7% [4]	113.4% [2,4]
P2*C2 (4)	0.0% [1,2,3]	0.0% [1,2,3]	**0.0%** [1,2,3]

Note: Superscripts indicate statistically different means at $p = 0.05$.

Next, least-squares means (LS Means), shown in Table 7.13, are computed for *p.rmse*, *p.bias* and *p.se* by each effect in the previous ANOVA. LS Means are predicted population margins – they estimate the marginal means over a balanced population (SAS Institute, 1999). For example, the ANOVA model for *p.rmse* is $\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, where α_i is the effect for level *i* of PSTATUS, β_j is the effect for level *j* of CSTATUS, and $(\alpha\beta)_{ij}$ is the interaction. The LS mean for the combination (PSTATUS/CSTATUS) for *p.rmse* is 47.5%, i.e. the use of both propensity score adjustment and calibration is predicted to reduce the *rmse* by 47.5% (averaged over the five propensity score adjustment models and two calibration methods). Pair differences are calculated using pairwise Tukey-Kramer adjusted differences. The results shown in Table 7.13 reveal that all three error statistics become large when either or both of the adjustments are applied. Among the four

possible combinations of the two adjustment status, using both adjustments is superior to any other adjustments. Its bias reduction (*p.bias*) can be as large as 78.8%. Although the standard error becomes 1.3 times larger, the root mean square error (*p.rmse*) size is smaller by 47.5% than that of the unadjusted estimate. Overall, one can say that the adjustment reduces the error in estimates.

7.3.4 Performance of Different Propensity Score Models and Calibration Models

How each propensity score model and calibration method affects all three error properties is examined in this section. As in the previous ANOVA, *p.rmse*, *p.bias* and *p.se* are fitted by the different types of propensity score models (PMODEL) and calibration adjustment methods (CMODEL). Note that the focus of examination here is on different models instead of adjustment application status. There are six different methods under PMODEL: no propensity score adjustment and propensity score Models *1* through *5*; and three CMODEL: no calibration and Calibration *1* and Calibration *2*.

All ANOVA models are significant in explaining the variance in the percent error statistics (See Table 7.14). In case of *se*, the model accounts for 97% of the variance in the percent *se* increase. Six different types of propensity score modeling and three calibration types have significantly different effects on the errors. However, their interactions are significant in explaining only *p.se*.

Table 7.14. Results of Analysis of Variance on Percent Root Mean Square Error Reduction, Percent Bias Reduction and Percent Standard Error Increase by Propensity Score Adjustment Models, Calibration Adjustment Models and Their Interactions

		p.rmse		*p.bias*		*p.se*	
Effect	***df***	***SS***	***F Value***	***SS***	***F Value***	***SS***	***F Value***
Model	17	0.7084	1.79*	2.5017	3.71**	9.3600	69.43**
Error	36	0.8363		1.4288		0.2855	
Total	53	1.5447		3.9304		9.6455	
		(R^2=0.4586)		(R^2=0.6365)		(R^2=0.9704)	
	df	***Type 3 SS***	***F Value***	***Type 3 SS***	***F Value***	***Type 3 SS***	***F Value***
Propensity Score Model (PMODEL)	5	0.2513	2.16*	0.7210	3.63**	1.2060	30.41**
Calibration Model (CMODEL)	2	0.2603	5.60**	1.3899	17.51**	6.2374	393.25**
Interaction (PMODEL* CMODEL)	10	0.1969	0.85	0.3908	0.98	1.9166	24.17**

Note: * $p < 0.1$, ** $p < 0.05$

Table 7.15 provides more detailed information on the performance of different propensity score models and calibration methods. It displays least square means of *p.rmse*, *p.bias* and *p.se* by effects of PMODEL and CMODEL included in ANOVA in the previous table. Contrary to expectations, the table does not convey clear-cut messages about the superiority of particular propensity score models and calibration methods. In general, propensity score Model *1* and *2* are preferable – although their statistical significance does not always hold, the direction is obvious. Propensity score Model *1* includes four stratifying variables (*age*, *educ*, *gender*, *race*); and model *2* includes all 30 covariates in Table 7.9. This importance of Model *1* is logical because Web samples are drawn based on the stratification on those variables but using the extremely imbalanced distributions of the Harris Interactive respondents. Model *2* ought to perform well, since it uses the full matrix of covariates in the

adjustment. There is no clear association between the *AIC* of the propensity models and their error properties. The role of *AIC* in model building discussed earlier cannot be verified from this result.

Table 7.15. Least Square Mean of Percent Root Mean Square Error Reduction, Percent Bias Reduction and Percent Standard Error Increase by Propensity Score Adjustment Models and Calibration Adjustment Models [a]

	LS Mean		
Effect	***p.rmse***	***p.bias***	***p.se***
Propensity Score Model (PMODEL)			
No Adjustment (P0)	26.8% P1, P2	40.4% P1, P2, P4, P5	**75.6%** P1, P2, P3, P4, P5
Model 1 (P1)	46.7% P0	**76.2%** P0	117.6% P0, P5
Model 2 (P2)	**47.0%** P0	71.6% P0	120.6% P0, P5
Model 3 (P3)	41.9%	62.2%	113.3% P0
Model 4 (P4)	42.7%	67.0% P0	110.1% P0
Model 5 (P5)	43.8%	69.2% P0	104.3% P0
Calibration Models (CMODEL)			
No Adjustment (C0)	31.8% C1, C2	41.8% C1, C2	**58.9%** C1, C2
Calibration 1 (C1)	44.8% C0	74.2% C0	133.4% C0
Calibration 2 (C2)	**47.8%** C0	**77.4%** C0	128.4% C0

a. LS Means by interactions are excluded from the table, since there is little difference across 18 different combinations.

Note: Superscripts indicate statistically different means at $p = 0.1$.

Among the two calibration methods, the second one using estimated population general health status as well as known population demographic characteristics seems to benefit the error structure to a larger degree than using only known values. One notable finding with calibration from the table above is that the Calibration *2* shows a larger decrease in bias and a smaller increase in standard error than Calibration *1*, although not statistically significant. This implies that a good calibration method may achieve bias reduction at a smaller level of variability increase.

7.3.5 Variance Estimation

7.3.5.1 Variance Estimation for Propensity Score Adjustment

There is no clear approach for deriving variance estimates when propensity score adjustment weights are applied. One method that commercial statistical software, such as SAS, may use would be the following estimator:

$$v_{naive}\left(y^{W.PSA}\right)=\left(1-f\right)\frac{n^{W}}{n^{W}-1}\sum_{c=1}^{C}\sum_{j\in\left(s^{W}\right)}\left[\left(d_{j}^{W.PSA}y_{j}-\frac{1}{n^{W}}\sum_{c=1}^{C}\sum_{j\in\left(s^{W}\right)}d_{j}^{W.PSA}y_{j}\right)\Big/N\right]^{2}, \quad (7.11)$$

where $d_j^{W.PSA}$ is the weight derived from propensity score adjustment in (6.2). However, this is a naïve approach, since the estimator does not account for the complexity of multiple weights in $d_j^{W.PSA} = f_c d_j^W$, where f_c is the PSA factor and d_j^W is the base design weight for unit *j*. If the weights reflect a nonresponse adjustment which has not beeen incorporated in this study, they will be even more complicated. Thus, naïvely applying (7.11) may give poor results.

Table 7.16 shows estimated and empirical standard errors for the estimators with propensity score adjustment but without calibration adjustment. It allows a comparison between the *se* estimates from (7.11) (*v.naive*) and the simulation *se* from (7.8) (*v.sim*) by the ratio of the two. The naïve estimator tends to overestimate the actual variability, although the degree of overestimation is not too extreme. This tendency is worse for y_2, where the naïve *se* estimates are at least 12% larger than the actual *se*. This echoes the finding in Valliant (2004) which showed an understatement of efficiency of employing the naïve estimator when calculating variances of estimates adjusted by multiple weights.

Table 7.16. Estimated Standard Error and Simulation Standard Error of Propensity Score Adjusted Web Sample Estimates

Propensity Score Model	y_1 : **HBP**			y_2 : **SMOKE**			y_3 : **ACT**		
	v.naive	***v.sim***	***Ratio (naive/sim)***	***v.naive***	***v.sim***	***Ratio (naive/sim)***	***v.naive***	***v.sim***	***Ratio (naive/sim)***
Model *1*	0.0231	0.0247	93.8%	0.0264	0.0226	116.8%	0.0239	0.0211	113.1%
Model *2*	0.0219	0.0215	101.9%	0.0262	0.0222	118.2%	0.0232	0.0225	103.5%
Model *3*	0.0207	0.0206	100.1%	0.0262	0.0217	120.7%	0.0230	0.0211	109.0%
Model *4*	0.0151	0.0152	99.3%	0.0176	0.0154	114.0%	0.0154	0.0163	94.7%
Model *5*	0.0146	0.0140	104.6%	0.0166	0.0148	112.5%	0.0150	0.0147	102.2%

7.3.5.2 Variance Estimation for Calibration Adjustment

Two variance estimation approaches are examined for cases when the calibration adjustment is added to the propensity score adjustment. The first follows the naïve approach in (7.11), where $d_j^{W.PSA}$ is replaced with the calibration weight $\tilde{w}_j$ from Section 6.3, such that

$$v_{naive}\left(y^{W.A}\right) = (1-f)\frac{n^W}{n^W-1}\sum_{j\in\left(s^W\right)}\left[\left(\tilde{w}_j y_j - \frac{1}{n^W}\sum_{j\in\left(s^W\right)}\tilde{w}_j y_j\right)\Big/ N\right]^2 . \qquad (7.12)$$

As mentioned above, this squared residual method is the same as what commercially available software packages typically utilize for variance estimation. The second method which originates from Deville and Särndal (1992) uses the following variance estimator modified from the asymptotic variance estimator for the GREG for the population total, t_y:

$$v_{ds}\left(t_y\right) = \sum\sum_{s}\left(\Delta_{ij}/\pi_{ij}\right)\left(\tilde{w}_i e_i\right)\left(\tilde{w}_j e_j\right), \qquad (7.13)$$

where *i* and *j* denote units in the sample; $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$; π_i and π_j are inclusion probabilities for unit *i* and *j* into the sample; π_{ij} is a joint inclusion probability of the two units; and e_i is the sample-based residual defined as $e_i = y_i - \mathbf{z}_i'\hat{\mathbf{B}}_{ws}$. $\hat{\mathbf{B}}_{ws}$ is the

regression slope estimate computed as in (6.15) and (6.16). Since we use Poisson sampling to draw Web samples, the variance estimator (7.13) becomes simplified as

$$v_{ds}\left(t_y\right) = \sum_{i\in\left(s^W\right)} \left(\tilde{w}_i e_i\right)^2 + \sum_{i\neq j} \left(\tilde{w}_i e_i\right)\left(\tilde{w}_j e_j\right), \tag{7.14}$$

As we are estimating the population mean and the samples are drawn with replacement, (7.14) is changed to obtain

$$v_{ds}\left(y^{W.A}\right) = \left(1-f\right)\frac{n^W}{n^W-1}\sum_{j\in\left(s^W\right)}\left[\left(\tilde{w}_j e_j\right)/N\right]^2 . \tag{7.15}$$

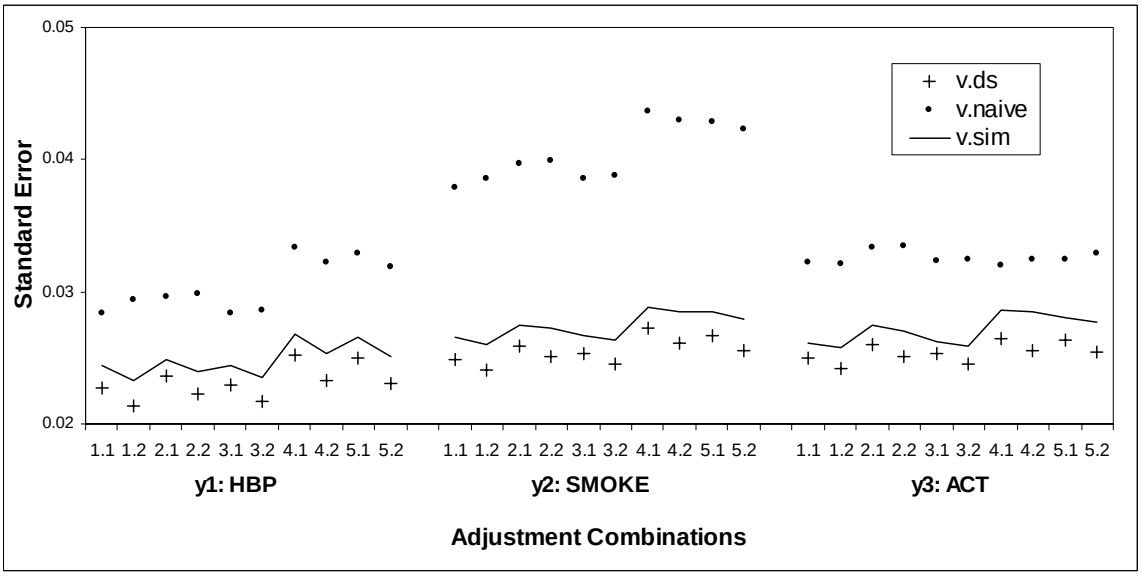


Figure 7.7. Standard Error of Adjusted Web Sample Estimates by Different Adjustment Method Combinations

Standard errors estimates using (7.12) and (7.15) are computed in simulations using *v.naive* and *v.ds* in **`newcal.fcn`** shown in Appendix 1.7. The resultant statistics are compared to the simulation standard error for the estimators using both propensity score and calibration adjustment in Figure 7.7. As shown in Valliant

(2004) and in the previous discussion, the naïve approach overestimates the variability of survey estimates, presenting the survey estimates as if they are far less efficient. The estimator suggested by Deville and Särndal (1992) appears to estimate the actual variance reasonably well. Although it tends to underestimate the variability, the degree of its underestimation is much smaller than the degree of the overestimation in the naïve approach.

The standard errors of adjusted Web sample estimates are plotted against the respective bias reduction in the estimated mean achieved in adjustments in Figure 7.8. Over the range of bias reductions shown, *v.naive* is always a substantial overestimate while *v.ds* is somewhat too small.

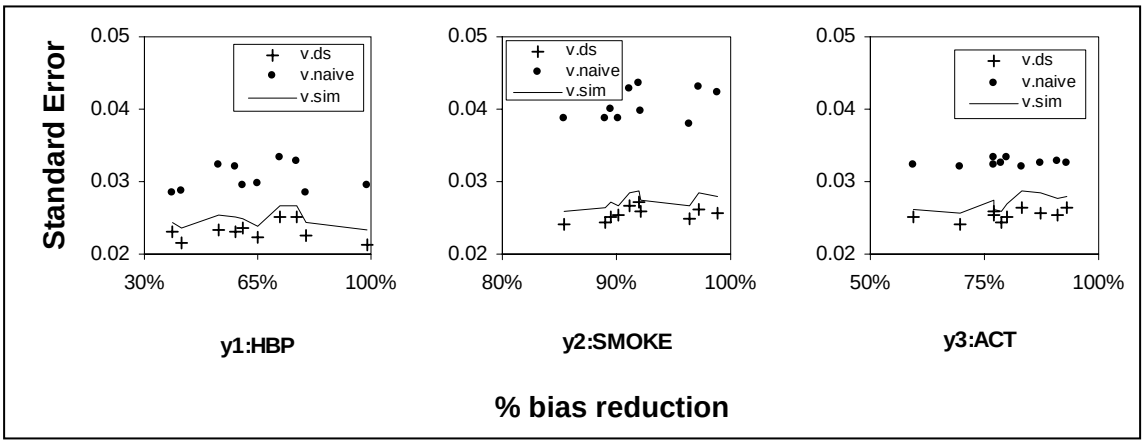


Figure 7.8. Relationship between Standard Error and Percent Bias Reduction of Adjusted Web Sample Estimates

The respective coverage rates for 95% confidence intervals in the simulation when *v.ds* and *v.naive* are used are presented in Table 7.17. It is striking that underestimation by *v.ds* leads to consistent undercoverage by its confidence intervals.

In contrast, the confidence intervals using *v.naive* have coverage rates that can be more or less than 95%. The cases with the poorest coverage tend to be ones where the estimated mean is biased so that the confidence intervals are not properly centered. For example, the standardized biases of the *3.1* and *3.2* estimates for HBP are about -9% from Table 7.11, and the coverage for *v.ds* are 75.2% and 74.4%, respectively.

Table 7.17. Coverage Rates of 95% Confidence Interval by Standard Error Estimated with *v.ds* and *v.naive*

Adjustment Combination	y_1 : **HBP**		y_2 : **SMOKE**		y_3 : **ACT**	
	v.ds	***v.naive***	***v.ds***	***v.naive***	***v.ds***	***v.naive***
1.1	90.7%	95.6%	92.5%	99.5%	74.5%	91.7%
1.2	92.7%	98.4%	90.8%	99.5%	81.6%	95.9%
2.1	85.8%	92.2%	92.4%	99.4%	88.3%	97.3%
2.2	86.3%	94.0%	91.6%	99.4%	88.6%	97.8%
3.1	75.2%	84.7%	92.4%	99.5%	88.7%	97.3%
3.2	74.4%	86.9%	91.7%	99.4%	88.7%	97.9%
4.1	89.5%	94.8%	92.1%	99.6%	88.5%	92.0%
4.2	81.8%	91.5%	91.8%	99.3%	89.1%	93.8%
5.1	89.8%	95.3%	92.3%	99.6%	92.6%	97.2%
5.2	83.8%	92.4%	92.1%	99.3%	91.5%	97.8%

One may argue that it is safe to use *v.naive* for calculating variance when calibration weights are applied, because it is conservative in stating the estimation efficiency. However, the degree of variance overestimation tends to be too large, especially for y_2. Confidence intervals obtained by standard errors using *v.naive* cover the population value over 99% of the time in Table 7.17. Recall that the bias reduction is achieved at the greatest degree for adjustments applied to y_2. For this variable, the coverage rates of *v.ds* are not as poor as for other variables. Therefore, it

seems sensible to examine the relationship between the bias reductions and the confidence interval coverage rates.

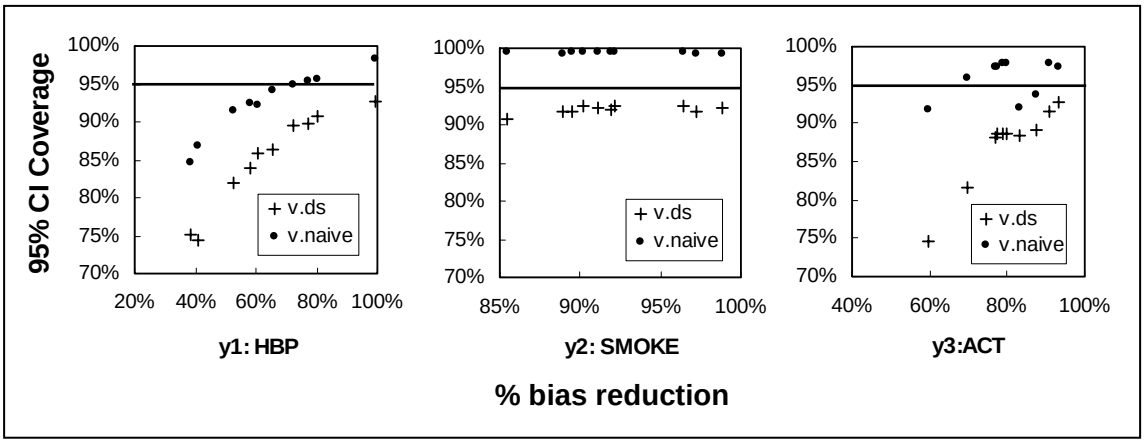


Figure 7.9. Relationship between 95% Confidence Interval Coverage and Percent Bias Reduction of Adjusted Web Sample Estimates

The relationship between the bias reductions and the confidence interval coverage rates is depicted in Figure 7.9. Although not remarkable, there is a noticeable relationship between the degree of bias reduction and that of confidence interval coverage. When the adjustment is poor in reducing the bias, the coverage rates of confidence intervals computed with variances from both (7.11) and (7.14) are far lower than the nominal rate, 95%. In contrast, the coverage rates tend to converge to 95%, as the biases are reduced to a larger degree. *V.naive* is not as good as *v.ds* where more than 85% biases are reduced.

7.3.6 Discussion

This case study examines the combination of propensity score adjustment and calibration adjustment for volunteer panel Web survey estimates. The results from

simulation show that using any of the two adjustments improves the accuracy of the sample estimates. The interaction of the two also has a significant effect on bias and *rmse* reduction. This confirms the effectiveness of each adjustment separately and the combined adjustment. At the same time, these adjustments increase the standard errors of estimates significantly. This reaffirms the trade-off between the bias reduction and the variability increase. Nonetheless, the adjustments decrease the magnitude of overall error substantially.

The examination of the separate propensity score models does not reveal clear implications. Models that include variables used in the sample selection and all auxiliary variables perform better. However, this does not bring in substantive understandings about the propensity score modeling strategies. As recommended by Rubin and Thomas (1996), it may be a sound approach to include all available covariates even if some are only remotely related to the study variables. The simulation on calibration adjustment suggests benefits from the inclusion of substantive variables whose population figures are estimated from another larger and more reliable survey. When the general health item is added in the calibration, the percent bias reduction and *rmse* reductions are larger but the percent *se* is smaller than when excluded. This item is asked in large-scale national health surveys, such as National Health Interview Survey, from which reliable population estimates are available. Therefore, the utilization of more substantive covariates is practical and effective in calibration adjustment.

The variance estimation methods tested in this study, unfortunately, do not provide conclusive guidelines. The naïve variance estimator that uses the multiply

adjusted weight as a simple single weight produces highly inflated figures. The estimator suggested by Deville and Särndal is a better approximation but tends to be too small. However, confidence intervals based on Deville and Särndal's standard errors cover the population values at lower rates than the targeted rate. This will deceptively portray the estimates as if they are more efficient than they actually are. The naïve approach, on one hand, can be better for some estimates since it is conservative. On the other hand, confidence intervals based on the naïve estimators can cover the population value close to 100% of the time, when the nominal rate is 95%. On a positive note, when the applied adjustment results in larger bias reduction, the Deville-Särndal estimator provides near nominal coverage. Replication is another variance estimation option that has the potential to be quite effective when complicated weighting methods are used like propensity score and calibration adjustment. Variations of the jackknife or bootstrap could be good choices for future investigation.

Chapter 8: Conclusion

With the advance in communication technology and the accompanying societal and cultural changes, Web surveys are here to stay. This research is carried out not to assert the scientific significance of Web surveys or to advocate the embracing of Web surveys, but to supplement what is lacking in current Web survey practice. Web surveys are popular but without any proven methodological value. In order to make potentially biased Web survey estimates usable, statistical adjustments may be employed in the estimation process. However, traditional adjustment techniques are found to be limited in compensating for the biases in Web survey estimates.

Based on that finding, this study attempts to adopt existing adjustment methods from the causal inference and survey statistics literature to volunteer panel Web survey data. First, protocols for recruiting volunteers for Web surveys are not guaranteed to produce random samples. This is viewed as a selection bias in this study. Propensity score adjustment in causal inference using observational data is a method that can remove or reduce the selection bias. We applied this to Web survey settings to derive an adjustment weight for selection bias. A second calibration adjustment is made to decrease the bias arising from the differences between the adjusted Web sample and the population. The study provides a mathematical presentation of processes in these adjustments which are absent in the existing research.

The performance of the adjustments is diagnosed in simulations. The two case studies carried out in this research convey the same clear implications about the adjustments: the propensity score adjustment and the calibration adjustment decrease bias and root mean square error in volunteer Web panel survey estimates; however, these reductions are realized with an increase in variance of estimates. It is also found that the error reduction becomes larger when the propensity score adjustment is used in conjunction with the calibration adjustment. The contention that nondemographic covariates are needed in propensity score models made by some survey organizations is not supported. The best method of covariate selection for propensity modeling appears to be the inclusion of all available variables in the adjustment. For calibration adjustment, utilization of substantive variables whose population estimates are obtainable from larger surveys does improve the quality of the adjustments.

The application scope of these adjustments may exceed volunteer panel Web surveys. When the quality of data collection is doubtful, one may adopt the adjustments examined in this research to make a better use out of the data. Imagine that one has a survey data set but fear that respondents' self-selection may have introduced bias but that there is a more reliable survey which has variables in common with one's survey. Propensity score adjustment can take advantage of the power of those overlapping covariates between the two surveys. This adjustment may be tuned to a finer degree by calibration using a smaller set of variables whose population figures are known or estimable from larger surveys. The survey estimates may become more usable after these adjustments.

When applying these adjustments to survey data, one should bear in mind the following. First, the adjustments are post-hoc in their nature. If feasible given the survey budget, it is important is to improve the survey procedures to collect better data. It would be unwise to intentionally collect suboptimal data, assuming that the adjustments will remove all biases. While the biases are reduced, they are not eliminated. It may not work under all circumstances, and as shown bias reduction depends on the model used. Second, when the covariates used in adjustments have missing data, propensity score adjustment becomes more difficult, because propensity scores cannot be assigned to the units in the merged data set with missing covariate information. This research uses hot-deck imputation to avoid this problem. One may consider following a recommendation of D'Agostino and Rubin (2000) to condition the propensity score on both observed values of covariates and the observed missing-data indicators. Third, this research uses the main effects of covariates in propensity models. One of the advantages of using propensity score adjustment weighting over the traditional weighting is the flexibility of the model formation. Propensity model refinement including higher order interactions among the covariates and using more covariates may provide a clearer insight about the variable selection. Fourth, the effectiveness of nondemographic covariates may not have been confirmed in this study, because the Web samples are drawn based on the distribution of demographic variables and these variables are also included in the adjustment. One may consider another way of drawing Web samples or conducting a series of Web surveys on substantive variables whose true values are either known or obtainable. Fifth, the sample size of the reference survey matters. When the size is small, the

subclassification based on propensity scores may not be possible. Instead of conducting a small reference survey for each Web survey, one possibility is to adopt a large-scale national survey as a reference survey. Lastly, the two variance estimation methods examined in this research did not perform well enough to be recommended for general use. Alternative variance estimators are needed as the final weights from the adjustments account for multiple steps of adjustment. Variance estimation methods with replication are alternatives. These remarks are hoped to provide directions for future research.

Appendices

1. R© Code Used in the Study

1.1 psa.fcn

```
function (dframe, form, pfit, prnk, qbin, bin, trmt)
    #   Propensity Score Adjustment Weight Calculation
    #    - Calcultates adjustment weights and throws
out
           the reference sample
    #
    #   dframe: data frame
    #   form: propesnity score model which needs to be
          defined beforehand
    #   pfit: fitted propensity scores
    #   prnk: propensity score rank
    #   qbin: propensity score bin number factor
created
          by PSdefine
    #   bins: number of bins to be formed
    #   trmt: treatment/control group variable

{
       if(missing(dframe)||!inherits(dframe,"data.fram
       e"))
            stop("First argument to PSdefine must be a
            Data Frame name.")
       if(missing(form)||class(form)!="formula")
            stop("Second argument to PSdefine must be
            a formula.")
       trtm <- deparse(form[[2]])

       if(!is.element(trtm,dimnames(dframe)[[2]]))
            stop("Response variable in the PSdefine
            formula must be an existing treatment
            factor.")
       dframe[,trtm] <- as.factor(dframe[,trtm])

       last.glm <- glm(form, family = binomial (link =
                       logit), data = dframe,
                            na.action = na.omit)


       df3 <- as.data.frame(fitted.values(last.glm))
       pfit <- deparse(substitute(pfit))
```

```
        dimnames(df3)[[2]] <- "pfit"
        prnk <- deparse(substitute(prnk))
        df3[,"prnk"] <- rank(df3[,"pfit"], na.last = T)
        qbin <- deparse(substitute(qbin))
        df3[,"qbin"] <-
                        factor(1+floor((bins*df3[,"prnk"
                  ])
                        /(1+length(df3[,"prnk"])))))

        newdframe <- merge(dframe, df3,
by.x="row.names",
                        by.y="row.names", all.x=T)

        if (any(ftable(newframe
             [,c("depend","qbin")])[]==0)==T)
           (newframe.c<-1)
        else {newframe.c<-0}

        if ((newframe.c >= 1)
        (skip<-TRUE)

        if (!skip)
        {
        nwc <- table(newdframe[newdframe[,var]==1,
bins])
        nrc <- table(newdframe[newdframe[,var]==0,
bins])
        nw <- length(newdframe[newdframe[,var]==1,
bins])
        nr <- length(newdframe[newdframe[,var]==0,
bins])

        wgt <- (nrc*nw)/(nwc*nr)
        wgt <- as.vector(wgt[newdframe[,bins]])
        allwgt <- data.matrix(cbind(newdframe, wgt))

        pwgt <- allwgt[,"basewgt"]*allwgt[,"wgt"]
        pwgt <- as.vector(pwgt)
        allwgt <- data.matrix(cbind(allwgt, pwgt))

        PSdframe <-
data.frame(allwgt[allwgt[,var]==2,])
                  #bc data.matrix makes trmt+1
        }
}
```

1.2 cal.fcn

```
function (pop, sam, sampx, knownx, estimx, L, U,

            conv.crit=0.01, max.steps=10., min.B=5)
{

  # Calculation - GLS wgts using restricted linear
                distance function
  #
  # pop: population
  # sam: sample
  # X: matrix of auxiliary vars; n x p
  # X.pop: matrix of pop controls; p x 1
  # X.hat: vector of HY estimates of X.pop
  # a: vector of base wgts (1/pi); n x 1
  # c: vector of model vars (usually set to 1); p x 1
  # L: lower bound on wgt ratio w/a
  # U: upper bound on wgt ratio w/a

 X <- sampx(sam)
 X.pop <- knownx(pop)
 X.hat <- estimx(sam, "pwgt")
 a <- as.matrix(sam[,"pwgt"])
 p <- ncol(X)
 c.vec <- rep(1., length(X[,1.]))


# convergence check on lambda
        lambda.old <- #
        rep(0.,p)

# iteration
        converged <- function(old, new, conv.crit)
        {
            check <- F
            D <- max(abs((old - new)/old))
            if (D < conv.crit) {
                 check <- T
            }
            check
        }
        step.num <- 0.


# compute weights
repeat{
```

```
    step.num <- step.num + 1
    max.steps.reached <- step.num > max.steps
    lambda.x <- lambda.old %*% t(X)/c.vec
    sA <- (lambda.x < (L - 1.))
    sB <- (lambda.x >= (L - 1.)) & (lambda.x <= (U -
1.))
    #
    sC <- (lambda.x > (U - 1.))

    if(sum(sB)< min.B)

         stop("Set sB too small, no. cases = ",
sum(sB),
                "No. of iteration steps used: ",
                step.num, "where:  ", sam, sampx,
           "\n")


    phi.sA <- phi.sB <- phi.sC <- 0.

    lambda.xsB <- lambda.old %*% t(X[sB, ])/c.vec[sB]

    Z.sB <- (a/c.vec)[sB] * X[sB, ]

    phi.prime <- t(Z.sB) %*% X[sB, ]

    if(sum(sA) != 0.) {
        if(length(a[sA])==1.)
            phi.sA <- (L - 1.) * a[sA] * X[sA, ]
        else phi.sA <- (L - 1.) * a[sA] %*% X[sA,]
    }

    phi.sB <- lambda.old %*% t(Z.sB) %*% X[sB, ]
    if(sum(sC) != 0.) {
        if(length(a[sC])==1.)
            phi.sC <- (U - 1.) * a[sC] * X[sC, ]
        else phi.sC <- (U - 1.) * a[sC] %*% X[sC,]
    }

    phi.sA <- as.vector(phi.sA)
    phi.sB <- as.vector(phi.sB)
    phi.sC <- as.vector(phi.sC)

    phi.s1 <- phi.sA + phi.sB + phi.sC
    phi.s2 <- as.matrix(phi.s1)
    phi.s3 <- t(phi.s2)
```

```
    lambda.new <- lambda.old - ginv(phi.prime) %*%
                  (t(phi.s3) +X.hat - X.pop)
    if(converged(lambda.old, lambda.new, conv.crit) |
       max.steps.reached) {
        cat("No. of iteration steps used:", step.num,
            "\n")
        break
    }

    lambda.old <- as.vector(lambda.new)
}

        g.fcn <- rep(0., length(X[, 1.]))
        lambda.x <- as.vector(lambda.new) %*%
t(X)/c.vec

        sA <- (lambda.x < (L - 1.))
        sB <- (lambda.x >= (L - 1.)) & (lambda.x <=
(U-
                  1.))
        sC <- (lambda.x > (U - 1.))

        g.fcn[sA] <- L
        g.fcn[sB] <- 1. + lambda.x[sB]
        g.fcn[sC] <- U

        calwgt <- a * g.fcn
        cwgt <- as.vector(calwgt)
        calwgt <- data.matrix(cbind(sam, cwgt))

        caldframe <- data.frame(calwgt)


}
```

1.3 ref.sam

```
function (pop, n)
{
  #   Select an srs as a reference sample
  #   pop: population
  #   n: sample size

   N <- nrow(pop)
     sam <- sample(1:N, n, replace = F)
     dat <- pop[sam, ]
     basewgt<-dim(pop)[[1]]/dim(dat)[[1]]
     dat<-cbind(dat, basewgt)
}
```

1.4 pois.sam

```
function(subpop, pop, ph, str, n)
{
  #   Select stratified Poisson sample from pop of
  size
       Nh
  #   subpop: subpopulation, e.g., web population
            #   pop: population, e.g., Entire GSS
  population
  #   ph: vector of proportions in strata that define
       rates of web usage
  #   str: column of pop for stratum (can be name or
       number)
  #   n: desired expected total sample size

        h <- subpop[,str]
        N <- nrow(subpop)
        Nh <- table(subpop[, str])
        H <- length(Nh)
        u <- runif(N, min=0, max=1)

        if (any(is.na(h))){
             stop("stratum vat str missing for some
             cases. Processing stopped.\n")
        }
        if (sum(ph)!=1){
             stop("sum(ph) != 1. Processing
stopped.\n")
        }

        if(H != length(ph)) {
             stop("\H != length(ph). Processing
             stopped.\n")
        }

        adjh <- n/ sum(Nh * ph)
        ph <- ph*adjh

        ph.pop <-ph[h]

        sam <- (u < ph.pop)
        sam <- subpop[sam,]

        basewgt<-dim(pop)[[1]]/dim(sam)[[1]]

        dat <- cbind(sam, basewgt)
```

```
}
```

1.5 psa.sim

```
function(pop, wpop, nr, nw, y, bw, pw,
         form1, form2, trmt, bin,
         seed, NoSams)
{

#################################################
# Propensity Score Adjustment Only
#################################################
# Estimation for "y"
# pop:  population data set
# wpop: web subpopulation
# nr:   reference sample size
# nw:   web sample size
# y:    variable of interest, e.g.,"vote"
# bw:   base weight
# pw:   PSA weight
# form1: PSA model 1 defined previously
# form2: PSA model 2 defined previously
# trmt: treatment variable for PSA, "depend"
# bin:  variable name for bins in PSA
# NoSams: Number of Simluated Samples
#################################################


set.seed(seed)
out.est <- array (0, dim=c(2,6,NoSams))
        cat ("Begin", date(), "\n")

for(s in 1:NoSams)
   {
   skip <- FALSE
   #skip_ct <- 0

   if (s%%1==0)
        cat("s=", s, date(), "\n")

#################################################
# sample draw

        ref<-ref.sam(pop, nr)
        strat<-pois.sam(wpop, pop, ph =
                        c(0.11441573, 0.06110840,
                                0.10417682,
0.03347960,
```

```
                              0.12099789, 0.02267187,
                                   0.09393792,
0.01227044,
                              0.07346010, 0.02754754,
                              0.06663416, 0.02007151,
                                   0.11173411,
0.01649602,
                              0.10498944, 0.01600845),
                              "str", nw)
        harris<-pois.sam(wpop, pop, ph =
                         c(0.0203, 0.0164, 0.0085,
                              0.0060, 0.0245, 0.0048,
                                   0.0170, 0.0024,
0.1328,                                          0.1337,
0.0758, 0.0909,
  0.1558, 0.0458, 0.2082,
                              0.0571), "str", nw)


  # basic estimates

        y.pop <- est(pop, y)
        y.pop <- rbind(y.pop, y.pop, y.pop)
        colnames(y.pop) <- "y.pop"

        y.wpop <- est(wpop, y)
        y.wpop <- rbind(y.wpop, y.wpop, y.wpop)
        colnames(y.wpop) <- "y.wpop"

        bp.wpop <- y.pop-y.wpop
        colnames(bp.wpop) <- "bp.wpop"

        y.R <- w.est(ref, y, bw)
        y.R <- rbind(y.R, y.R, y.R)
        colnames(y.R) <- "y.R"

        y.U.t <- w.est(strat, y, bw)
        y.U.h <- w.est(harris, y, bw)
        y.U <- rbind(y.U.t, y.U.h)
        colnames(y.U) <- "y.U"


#####################################################
# merge reference and web samples


        rt<-rbind(ref, strat)
```

```
        rh<-rbind(ref, harris)


#################################################
# propensity score adjustment

  psaform1.t <- psa.fcn(rt, form1, pfit, prnk, qbin,
                            bin, trmt)
  psaform1.h <- psa.fcn(rh, form1, pfit, prnk, qbin,
                                bin, trmt)
  psaform2.t <- psa.fcn(rt, form2, pfit, prnk, qbin,
                                bin, trmt)
  psaform2.h <- psa.fcn(rh, form2, pfit, prnk, qbin,
                                bin, trmt)


  # adjusted estimates

       y.pform1.t <- w.est(psaform1.t, y, pw)
       y.pform1.h <- w.est(psaform1.h, y, pw)
       y.pform1 <- rbind(y.pform1.t, y.pform1.h)
       colnames(y.pform1) <- "y.pform1 "

       y.pfporm2.t <- w.est(psa.form2.t, y, pw)
       y.pform2.h <- w.est(psa.form2.h, y, pw)
       y.pform2 <- rbind(y.pform2.t, y.pform2. h)
       colnames(y.pform2) <- "y.pform2"

#########################################################
###
#  bind all estimates into y.est

       y.est <- cbind (y.pop,
                       y.wpop,
                       y.R,
                       y.U,
                       y.pform1,
                       y.pform2)

       dimnames(y.est)[[1]][1]<-"strat"
       dimnames(y.est)[[1]][2]<-"harris"

out.est[ , , s] <- y.est
dimnames(out.est) <- list(dimnames(y.est)[[1]],
dimnames(y.est)[[2]], NULL)

  }   # end of s loop
```

```
cat("end", date(),"\n")
list("estimates"=out.est)

}
```

1.6 cal.sim

```
function(pop, wpop, nr, nw, y, bw, pw, cw,
         form1, form2,
         samp1, known1, estim1, samp2, known2, estim2,
         trmt, bin,
         seed, NoSams)
{

##################################################
# Estimation for "y"
# pop:  population data set
# wpop: web subpopulation
# nr:   reference sample size
# nw:   web sample size
# y:    variable of interest, e.g.,"vote"
# bw:   base weight
# pw:   PSA weight
# cw:   calibration weight
# form: PSA forms defined previously
# samp1: function for obtaining only calibration
             covariate matrix from Sample for
calibration 1
# known1: function for obtaining population figures of
             calibration covariates for calibration 1
# estim1: function for ontaining sample estimates of
             calibration covariates for calibration 1
# samp2: function for obtaining only calibration
             covariate matrix from Sample for
calibration 2
# known2: function for obtaining population figures of
             calibration covariates for calibration 2
# estim2: function for ontaining sample estimates of
             calibration covariates for calibration 2
# trmt: treatment variable for PSA, "depend"
# bin:  variable name for bins in PSA
# NoSams: Number of Simluated Samples
##################################################

set.seed(seed)
out.est <- array (0, dim=c(2,14,NoSams))

        cat ("Begin", date(), "\n")

for(s in 1:NoSams)
  {
  skip<-FALSE
```

```
  if (s%%1==0)
           cat("s=", s, date(), "\n")

################################################
# sample draw

        ref<-ref.sam(pop, nr)
                strat<-pois.sam(wpop, pop, ph =
                        c(0.11441573, 0.06110840,
                                0.10417682,
0.03347960,
                            0.12099789, 0.02267187,
                                0.09393792,
0.01227044,
                            0.07346010, 0.02754754,
                            0.06663416, 0.02007151,
                                0.11173411,
0.01649602,
                            0.10498944, 0.01600845),
                            "str", nw)
        harris<-pois.sam(wpop, pop, ph =
                        c(0.0203, 0.0164, 0.0085,
                            0.0060, 0.0245, 0.0048,
                                0.0170, 0.0024,
0.1328,                                          0.1337,
0.0758, 0.0909,
  0.1558, 0.0458, 0.2082,
                            0.0571), "str", nw)

       # basic estimates

        y.pop <- est(pop, y)
        y.pop <- rbind(y.pop, y.pop, y.pop)
        colnames(y.pop) <- "y.pop"

        y.wpop <- est(wpop, y)
        y.wpop <- rbind(y.wpop, y.wpop, y.wpop)
        colnames(y.wpop) <- "y.wpop"

        y.R.n <- w.est(ref, y, bw)
        y.R.n <- rbind(y.R.n, y.R.n, y.R.n)
        colnames(y.R.n) <- "y.R.n"

        y.U.n.t <- w.est(strat, y, bw)
        y.U.n.h <- w.est(harris, y, bw)
        y.U.n <- rbind(y.U.n.t, y.U.n.h)
        colnames(y.U.n) <- "y.U.n"
```

```
#####################################################
# merge reference and web sample

        rt<-rbind(ref, strat)
        rh<-rbind(ref, harris)



#####################################################
# propensity score adjustment only


   psaform1.t <- psa.fcn(rt, form1, pfit, prnk, qbin,
                         bin, trmt)
   psaform1.h <- psa.fcn(rh, form1, pfit, prnk, qbin,
                                  bin, trmt)

   psaform2.t <- psa.fcn(rt, form2, pfit, prnk, qbin,
                                  bin, trmt)
   psaform2.h <- psa.fcn(rh, form2, pfit, prnk, qbin,
                                  bin, trmt)



   # adjusted estimates

        y.pform1.n.t <- w.est(psaform1.t, y, pw)
        y.pform1.n.h <- w.est(psaform1.h, y, pw)
        y.p1.n <- rbind(y.pform1.n.t, y.pform1.n.h)
        colnames(y.p1.n) <- "y.p1.n"

        y.pform2.n.t <- w.est(psa.form2.t, y, pw)
        y.pform2.n.h <- w.est(psa.form2.h, y, pw)
        y.p2.n <- rbind(y.pform2.n.t, y.pform2.n.h)
        colnames(y.p2.n) <- "y.p2.n"



#####################################################
# calibration adjustment


        psaform1.cal1.t <- cal.fcn(pop, psaform1.t,
```

```
                          samp1, known1, estim1, L,
U)
       psaform1.cal1.h <- cal.fcn(pop, psaform1.h,
                          samp1, known1, estim1, L,
U)

       psaform1.cal2.t <- cal.fcn(pop, psaform1.t,
                          samp2, known2, estim2, L,
U)
       psaform1.cal2.h <- cal.fcn(pop, psaform1.h,
                          samp2, known2, estim2, L,
U)
       psaform2.cal1.t <- cal.fcn(pop, psaform2.t,
                          samp1, known1, estim1, L,
U)
       psaform2.cal1.h <- cal.fcn(pop, psaform2.h,
                          samp1, known1, estim1, L,
U)

       psaform2.cal2.t <- cal.fcn(pop, psaform2.t,
                          samp2, known2, estim2, L,
U)
       psaform2.cal2.h <- cal.fcn(pop, psaform2.h,
                          samp2, known2, estim2, L,
U)

       psano.cal1.t <- cal.fcn(pop, strat, samp1,
                          known1, estim1, L, U)
       psano.cal1.h <- cal.fcn(pop, harris, samp1,
                          known1, estim1, L, U)

       psano.cal2.t <- cal.fcn(pop, strat, samp2,
                               known2, estim2, L, U)
       psano.cal2.h <- cal.fcn(pop, harris, samp2,
                               known2, estim2, L, U)

       psano.cal1.R <- cal.fcn(pop, ref, samp1,
known1,
                          estim1, L, U)
       psano.cal2.R <- cal.fcn(pop, ref, samp2,
known2,
                          estim2, L, U)


   # adjusted estimates

       y.p1.c1.t <- w.est(psaform1.cal1.t, y, cw)
```

```
y.p1.c1.h <- w.est(psaform1.cal1.h, y, cw)

y.p1.c2.t <- w.est(psaform1.cal2.t, y, cw)
y.p1.c2.h <- w.est(psaform1.cal2.h, y, cw)

y.p2.c1.t <- w.est(psaform2.cal1.t, y, cw)
y.p2.c1.h <- w.est(psaform2.cal1.h, y, cw)

y.p2.c2.t <- w.est(psaform2.cal2.t, y, cw)
y.p2.c2.h <- w.est(psaform2.cal2.h, y, cw)

y.n.c1.t <- w.est(psano.cal1.t, y, cw)
y.n.c1.h <- w.est(psano.cal1.h, y, cw)

y.n.c2.t <- w.est(psano.cal2.t, y, cw)
y.n.c2.h <- w.est(psano.cal2.h, y, cw)

y.R.c1 <- w.est(psano.cal2.R, y, cw)
y.R.c2 <- w.est(psano.cal2.R, y, cw)

y.p1.c1 <- rbind(y.p1.c1.t, y.p1.c1.h)
y.p1.c2 <- rbind(y.p1.c2.t, y.p1.c2.h)

y.p2.c1 <- rbind(y.p2.c1.t, y.p2.c1.h)
y.p2.c2 <- rbind(y.p2.c2.t, y.p2.c2.h)

y.U.c1 <- rbind(y.n.c1.t, y.n.c1.h)
y.U.c2 <- rbind(y.n.c2.t, y.n.c2.h)

y.R.c1 <- rbind(y.R.c1, y.R.c1)
y.R.c2 <- rbind(y.R.c2, y.R.c2)


colnames(y.p1.c1) <- "y.p1.c1"
colnames(y.p1.c2) <- "y.p1.c2"

colnames(y.p2.c1) <- "y.p2.c1"
colnames(y.p2.c2) <- "y.p2.c2"

colnames(y.U.c1) <- "y.U.c1"
colnames(y.U.c2) <- "y.U.c2"

colnames(y.R.c1) <- "y.R.c1"
colnames(y.R.c2) <- "y.R.c2"
```

```
#########################################################
###
#  bind all estimates into y.est


        y.est <- cbind (y.pop,
                        y.wpop,
                        y.R.n,
                        y.R.c1,
                        y.R.c2,
                        y.U.n,
                        y.U.c1,
                        y.U.c2,
                        y.p1.n,
                        y.p1.c1,
                        y.p1.c2,
                        y.p2.n,
                        y.p2.c1,
                        y.p2.c2)


        dimnames(y.est)[[1]][1]<-"strat"
        dimnames(y.est)[[1]][2]<-"harris"


out.est[ , , s] <- y.est

dimnames(out.est) <- list(dimnames(y.est)[[1]],
dimnames(y.est)[[2]], NULL)


        }   # end of s loop

        cat("end", date(),"\n")
        list("estimates"=out.est)

}
```

1.7 newcal.fcn

```
function (pop, sam, sampx, knownx, estimx, L, U,
            conv.crit, max.steps, min.B, y)
{
#####################################################
# Calibration and Variance estimation
#####################################################
# pop: population
# sam: sample
# X: matrix of auxiliary vars; n x p
# X.pop: matrix of pop controls; p x 1
# X.hat: vector of HY estimates of X.pop
# a: vector of base wgts (1/pi); n x 1
# c: vector of model vars (usually set to 1); p x 1
# L: lower bound on wgt ratio w/a
# U: upper bound on wgt ratio w/a
# conv.crit: convergence criterion
# max.steps: maximum number of calibration iteration
# y: variable of interest, e.g. “HBP”
#####################################################

    X <- sampx(sam)
    X.pop <- knownx(pop)
    X.hat <- estimx(sam, "pwgt")
    a <- as.matrix(sam[,"pwgt"])
    p <- ncol(X)
    c.vec <- rep(1., length(X[,1.]))

# convergence check on lambda
   lambda.old <- #
   rep(0.,p)

# iteration
   converged <- function(old, new, conv.crit)
   {
       check <- F
       D <- max(abs((old - new)/old))
       if (D < conv.crit) {
              check <- T
       }
       check
   }
   step.num <- 0.


# compute weights
```

```
repeat{
    step.num <- step.num + 1
    max.steps.reached <- step.num > max.steps
    lambda.x <- lambda.old %*% t(X)/c.vec

    sA <- (lambda.x < (L - 1.))
    sB <- (lambda.x >= (L - 1.)) & (lambda.x <= (U -
1.))
    #
    sC <- (lambda.x > (U - 1.))

    if(sum(sB)< min.B)

   stop("Set sB too small, no. cases = ", sum(sB),
            " No. of iteration steps used: ",
step.num,
            "where:  ", sam, sampx, "\n")

    phi.sA <- phi.sB <- phi.sC <- 0.

    lambda.xsB <- lambda.old %*% t(X[sB, ])/c.vec[sB]

    Z.sB <- (a/c.vec)[sB] * X[sB, ]

    phi.prime <- t(Z.sB) %*% X[sB, ]

    if(sum(sA) != 0.) {
        if(length(a[sA])==1.)
            phi.sA <- (L - 1.) * a[sA] * X[sA, ]
        else phi.sA <- (L - 1.) * a[sA] %*% X[sA,]
    }

    phi.sB <- lambda.old %*% t(Z.sB) %*% X[sB, ]
    if(sum(sC) != 0.) {
        if(length(a[sC])==1.)
            phi.sC <- (U - 1.) * a[sC] * X[sC, ]
        else phi.sC <- (U - 1.) * a[sC] %*% X[sC,]
    }

    phi.sA <- as.vector(phi.sA)
    phi.sB <- as.vector(phi.sB)
    phi.sC <- as.vector(phi.sC)
    phi.s1 <- phi.sA + phi.sB + phi.sC
    phi.s2 <- as.matrix(phi.s1)
    phi.s3 <- t(phi.s2)
```

```
    lambda.new <- lambda.old - ginv(phi.prime) %*%
                  (t(phi.s3) +X.hat - X.pop)

    if(converged(lambda.old, lambda.new, conv.crit) |
       max.steps.reached) {
        cat("No. of iteration steps used:", step.num,
            "\n")
        break
    }
    lambda.old <- as.vector(lambda.new)
}

  cat("Max relative change in lambda at last step: ",#
  max(abs((lambda.old - lambda.new)/lambda.old)),
"\n")
  g.fcn <- rep(0., length(X[, 1.]))
  lambda.x <- as.vector(lambda.new) %*% t(X)/c.vec

  sA <- (lambda.x < (L - 1.))
  sB <- (lambda.x >= (L - 1.)) & (lambda.x <= (U-1.))
  sC <- (lambda.x > (U - 1.))

  g.fcn[sA] <- L
  g.fcn[sB] <- 1. + lambda.x[sB]
  g.fcn[sC] <- U

  calwgt <- a * g.fcn

  cwgt <- as.vector(calwgt)
  calwgt <- data.matrix(cbind(sam, cwgt))
  calwgt <- data.frame(calwgt)

#####################################################
# Variance estimation
#####################################################

  # Deville-Sarndal variance

  Y <- as.matrix(sam[, y])
  sampsize <- dim(sam)[[1]]
  popsize <- dim(pop)[[1]]

  A <- t(X*cwgt) %*% X
  B <- ginv(A) %*% t(X*cwgt) %*% Y
  e <- Y - X %*% B
  nwgt <- cwgt/sum(cwgt)
```

```
  v.ds <- (1-sampsize/popsize)*(sampsize/(sampsize-
            1))*sum((nwgt*e)^2)
  v.ds <- rbind(v.ds.y1, v.ds.y2, v.ds.y3)



  # Naive variance

  m.y <- mean (nwgt*Y)

  v.naive <- (1-sampsize/popsize)*(sampsize/(sampsize-
                1))*sum((nwgt*Y-m.y)^2)

  newcaldframe<- list("calwgt"=calwgt, "v.ds"=v.ds,
                "v.naive"=v.naive)

}
```

2. GSS Propensity Score Model Specification in R©

2.1 y_{blks} : Warm Feelings towards Blacks

D1

```
depend ~  age+educ+newsize+hhldsize+income+
          as.factor(race)+as.factor(gender)+
          as.factor(married)+as.factor(region)+
```

D2

```
depend ~  age+educ+as.factor(race)+as.factor(gender)+
          as.factor(region)
```

D3

```
depend ~  newsize+hhldsize+income+as.factor(married)
```

A1

```
depend ~  age+educ+newsize+hhldsize+income+
          as.factor(race)+as.factor(gender)+
          as.factor(married)+ as.factor(region)+
               class+as.factor(work)+as.factor(party)+
          as.factor(religion)+ethnofit
```

A2

```
depend ~  age+educ+as.factor(race)+as.factor(gender)+
          as.factor(region)+ethnofit
```

A3

```
depend ~  newsize+hhldsize+income+as.factor(married)+
          class+as.factor(work)+as.factor(party)+
          as.factor(religion)
```

N1

```
depend ~  class+as.factor(work)+as.factor(party)+
          as.factor(religion)+ethnofit
```

N2

```
depend ~  ethnofit
```

N3

```
depend ~  class+as.factor(work)+as.factor(party)+
          as.factor(religion)
```

4

```
depend ~  age+educ+newsize+hhldsize+income98+
          as.factor(race)+as.factor(gender)+
          as.factor(married)+as.factor(region)+ethnofit
```

2.2 y_{blks} : Voting Participation in 2000 Presidential Election

A1

```
depend[22] ~      age+educ+newsize+hhldsize+income+
          as.factor(race)+as.factor(gender)+
          as.factor(married)+ as.factor(region)+
               class+as.factor(work)+as.factor(party)+
          as.factor(religion)
```

A2

```
depend ~  age+educ+income+as.factor(race)+
          as.factor(married)+class+as.factor(party)
```

A3

```
depend ~  newsize+hhldsize+as.factor(gender)+
          as.factor(region)+as.factor(work)+
          as.factor(religion)
```

D1

```
depend ~  age+educ+newsize+hhldsize+income+
          as.factor(race)+as.factor(gender)+
          as.factor(married)+as.factor(region)+
```

D2

```
depend ~  age+educ+income+as.factor(race)+
          as.factor(married)
```

D3

```
depend ~  newsize+hhldsize+as.factor(gender)+
          as.factor(region)+
```

N1

```
depend ~  class+as.factor(work)+as.factor(party)+
          as.factor(religion)
```

N2

```
depend ~  class+as.factor(party)
```

N3

```
depend ~  as.factor(work)+as.factor(religion)
```

4

```
depend ~  age+educ+newsize+hhldsize+income+
          as.factor(race)+as.factor(gender)+
          as.factor(married)+ as.factor(region)+
               class+as.factor(party)
```

[22] depend: An indicator for the status of each unit whether included in the Web or reference sample; the same as *g* in Chapter 6.

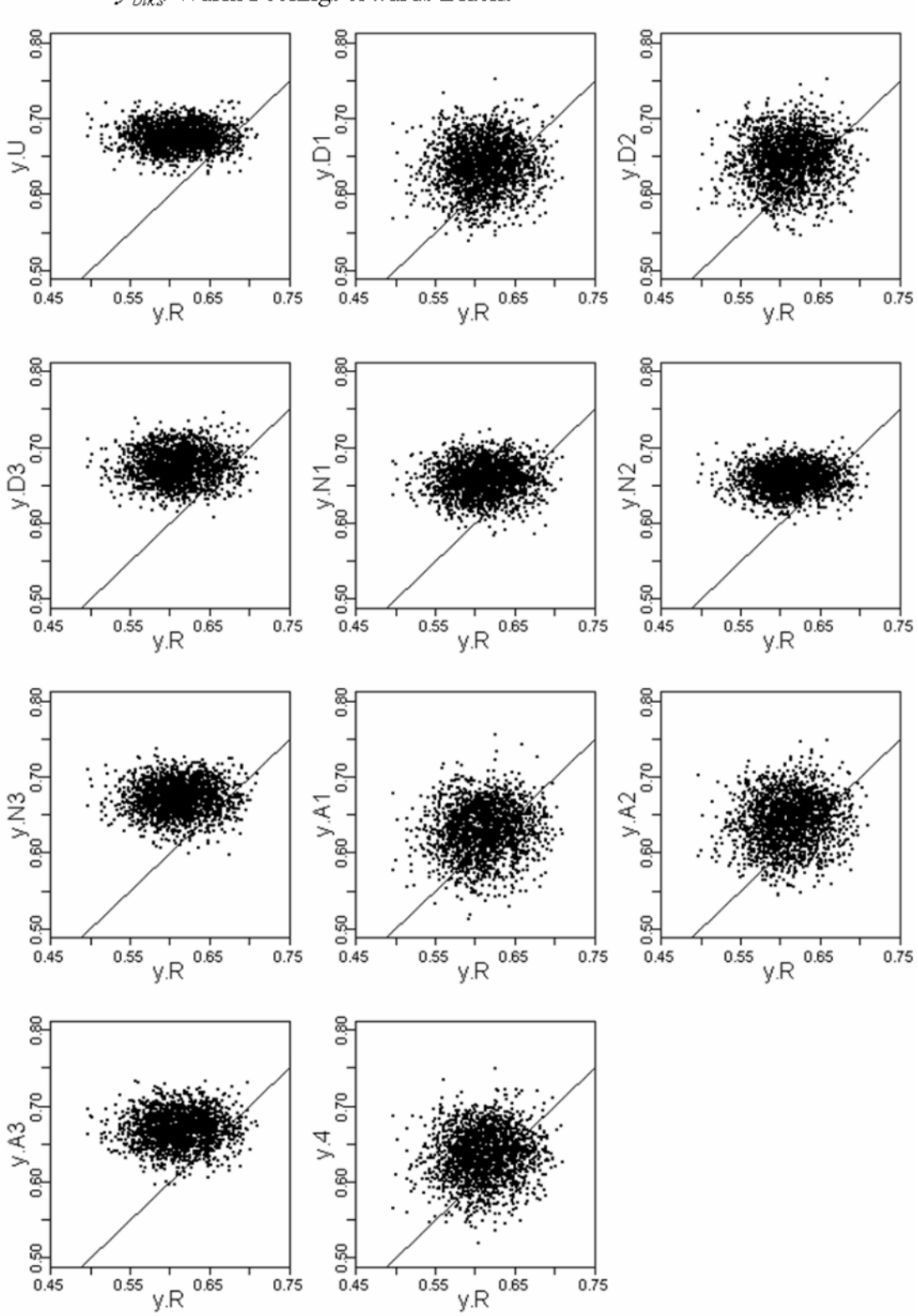
3. Reference Sample and Unadjusted and Propensity Score Adjusted Web Sample Estimates for y_{blks} and y_{vote}

	$s^{W.ST}$						$s^{W.HI}$					
	estimate	***bias***	***p.bias***	***rmsd***	***p.rmsd***	***se***	***estimate***	***bias***	***p.bias***	***rmsd***	***p.rmsd***	***se***
y_{blks}												
y.R	0.612					0.034	0.612					0.034
y.U	0.636	0.024		0.0448	0.0%	0.016	0.675	0.064		0.074	0.0%	0.016
y.D1	0.623	0.012	52.4%	0.0405	9.6%	0.022	0.638	0.026	58.6%	0.052	29.4%	0.032
y.D2	0.622	0.010	57.1%	0.0398	11.2%	0.021	0.645	0.034	47.0%	0.056	24.6%	0.031
y.D3	0.637	0.025	-4.7%	0.0457	-2.0%	0.018	0.675	0.063	0.4%	0.074	-0.8%	0.021
y.N1	0.620	0.008	65.7%	0.0388	13.5%	0.020	0.657	0.046	28.3%	0.060	18.5%	0.022
y.N2	0.622	0.010	58.6%	0.0386	13.9%	0.018	0.658	0.046	27.3%	0.059	19.5%	0.017
y.N3	0.632	0.020	17.5%	0.0430	4.1%	0.018	0.672	0.061	4.8%	0.072	2.3%	0.021
y.A1	0.616	0.004	82.0%	0.0390	13.1%	0.023	0.629	0.017	72.6%	0.048	35.5%	0.032
y.A2	0.617	0.005	79.4%	0.0387	13.6%	0.022	0.642	0.030	52.2%	0.054	27.5%	0.032
y.A3	0.636	0.024	1.7%	0.0451	-0.5%	0.019	0.669	0.057	10.0%	0.070	5.8%	0.021
y.4	0.619	0.007	71.3%	0.0392	12.5%	0.023	0.635	0.023	63.9%	0.050	31.8%	0.032

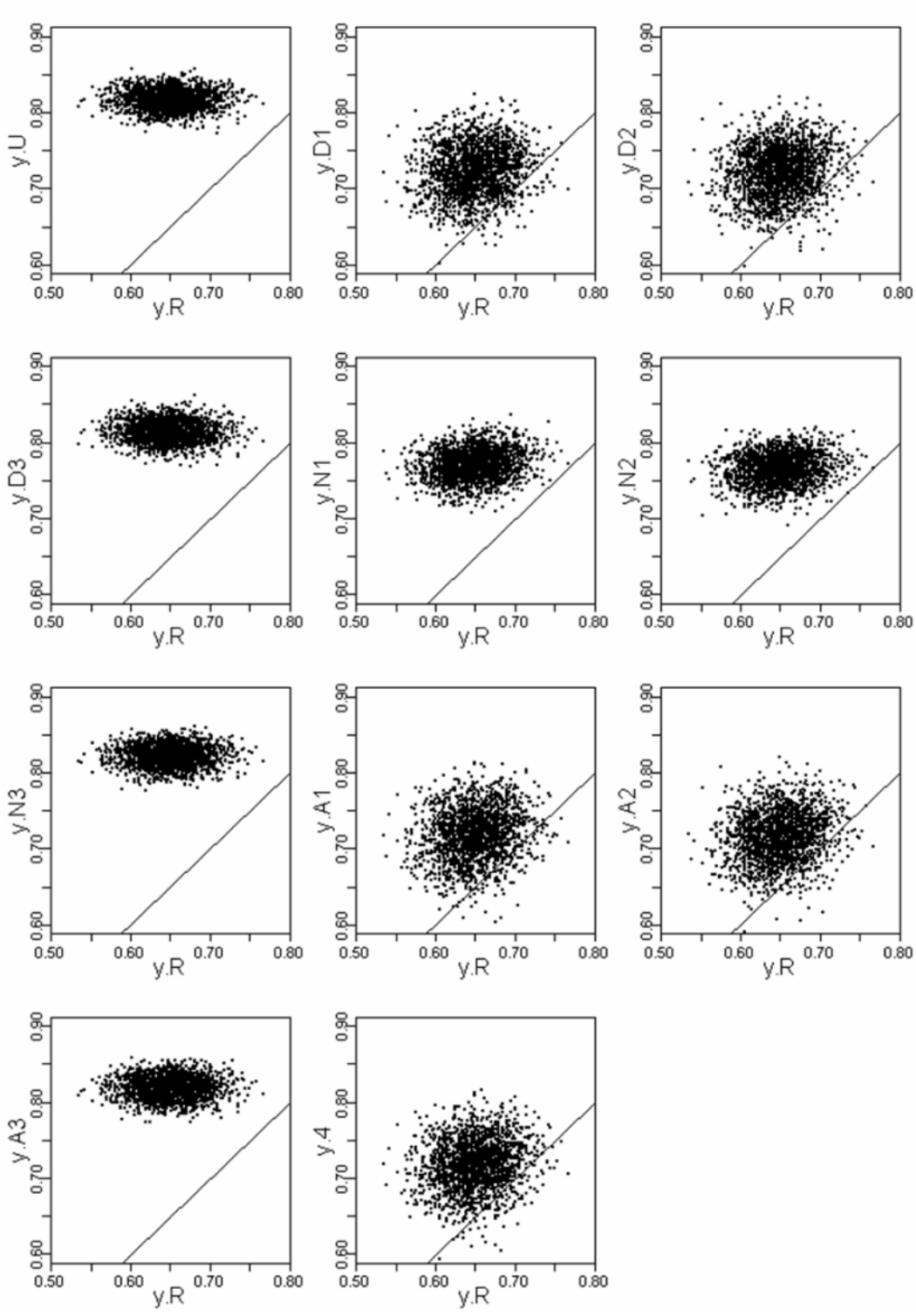
	$s^{W.ST}$						$s^{W.HI}$					
	estimate	***bias***	***p.bias***	***rmsd***	***p.rmsd***	***se***	***estimate***	***bias***	***p.bias***	***rmsd***	***p.rmsd***	***se***
y_{vote}												
y.R	0.650					0.034	0.650					0.034
y.U	0.715	0.065		0.075	0.0%	0.015	0.817	0.167		0.171	0.0%	0.013
y.D1	0.709	0.059	9.7%	0.069	8.3%	0.022	0.724	0.074	55.7%	0.086	50.0%	0.031
y.D2	0.711	0.062	5.4%	0.071	5.2%	0.021	0.721	0.072	57.2%	0.084	51.2%	0.032
y.D3	0.720	0.070	-7.1%	0.079	-5.7%	0.016	0.814	0.164	1.7%	0.169	1.6%	0.014
y.N1	0.695	0.045	30.5%	0.057	23.4%	0.019	0.771	0.121	27.5%	0.127	26.1%	0.020
y.N2	0.694	0.044	32.0%	0.057	24.4%	0.019	0.764	0.115	31.4%	0.121	29.6%	0.019
y.N3	0.719	0.069	-5.6%	0.078	-4.3%	0.016	0.821	0.172	-2.6%	0.175	-2.4%	0.013
y.A1	0.702	0.052	19.9%	0.063	16.2%	0.024	0.718	0.069	58.9%	0.081	52.7%	0.032
y.A2	0.706	0.057	13.5%	0.066	11.9%	0.023	0.716	0.066	60.4%	0.079	54.1%	0.032
y.A3	0.724	0.074	-13.4%	0.083	-10.5%	0.017	0.818	0.169	-0.7%	0.172	-0.6%	0.014
y.4	0.703	0.053	18.8%	0.063	15.5%	0.024	0.718	0.068	59.2%	0.080	53.1%	0.032

4. Relationship between the Distributions of the Different Web Sample Estimates and the Reference Sample Estimates for y_{blks} and y_{vote}

I. y_{blks}: Warm Feelings towards Blacks

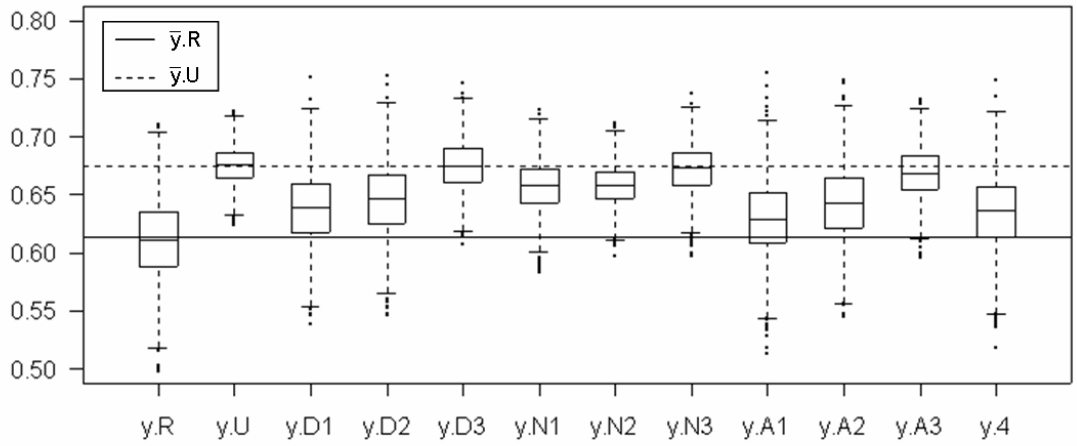


II. y_{vote}: Voting Participation

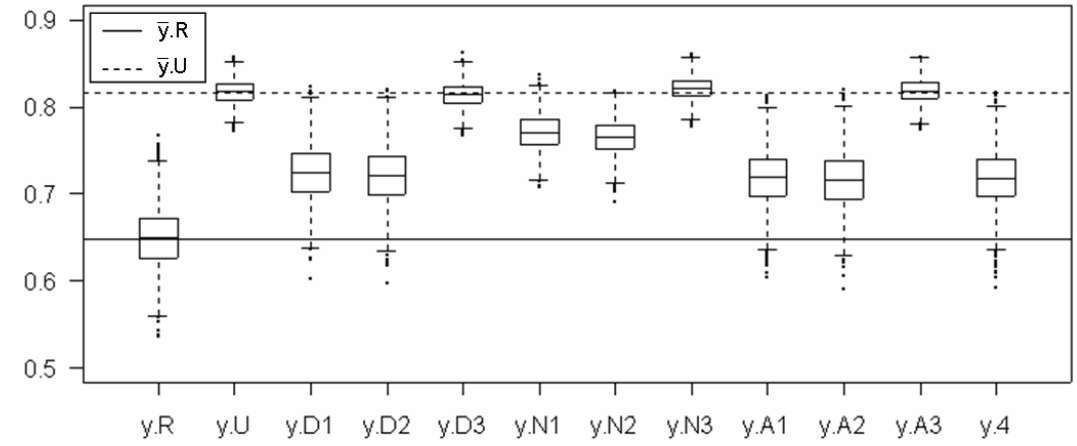


5. Distributions of the Web Estimates by Different Propensity Score Adjustments

I. y_{blks}: Warm Feelings towards Blacks



II. y_{vote}: Voting Participation



6. BRFSS Propensity Score Model Specification in R©

Model 1

```
depend ~  age+educ+as.factor(gender)+as.factor(race)
```

Model 2

```
depend ~  ghealth+as.factor(coverage)+as.factor(doctor)+
          as.factor(cprevent)+as.factor(phyact)+
          as.factor(diabete)+as.factor(cholest)+
          as.factor(losewgt)+ as.factor(wgtadv)+

          as.factor(asthma)+as.factor(flushot)+
          as.factor(pneumon)+as.factor(sunburn)+
          age+educ+income+weight+numphone+
          as.factor(gender)+as.factor(jointsym)+

          as.factor(limitact)+as.factor(modact)+
          as.factor(army)+as.factor(cellphon)+

     alcohol+hhsize+as.factor(work)+as.factor(marry)+
          as.factor(race)+veggie
```

Model 3

```
depend ~  ghealth+as.factor(doctor)+as.factor(cprevent)+
               as.factor(diabete)+as.factor(losewgt)+

     as.factor(sunburn)+educ+income+as.factor(gender)+
          as.factor(limitact)+as.factor(army)+
          as.factor(cellphon)+as.factor(race)
```

Model 4

```
depend ~  ghealth+as.factor(coverage)+as.factor(doctor)+
          as.factor(cprevent)+as.factor(phyact)+
          as.factor(diabete)+as.factor(cholest)+
          as.factor(losewgt)+ as.factor(wgtadv)+

          as.factor(asthma)+as.factor(flushot)+
          as.factor(pneumon)+as.factor(sunburn)+
          income+weight+numphone+as.factor(jointsym)+
               as.factor(limitact)+as.factor(modact)+
          as.factor(army)+as.factor(cellphon)+alcohol+
          hhsize+as.factor(work)+as.factor(marry)+veggie
```

Model 5

```
depend ~  ghealth+as.factor(doctor)+as.factor(cprevent)+
               as.factor(diabete)+as.factor(losewgt)+
          as.factor(sunburn)+ income+as.factor(limitact)+
               as.factor(army)+as.factor(cellphon)
```

Bibliography

Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91 (434), 444-472.

Benjamin, D.J. (2003). Does 401(k) Eligibility Increase Saving? Evidence From Propensity Score Subclassification. *Journal of Public Economics*, 87(5-6), 1259-90.

Berk, R.A., and Newton, P.J. (1985). Does Arrest Really Deter Wife Battery? An Effort to Replicate the Findings of the Minneapolis Spouse Abouse Experiment. *American Sociological Review*, 50, 253-262.

Burnett, R., and Marshall, P.D. (2003). *Web Theory. An Introduction.* New York, NY: Routledge.

Casady R.J., and Lepkowski, J.M. (1993). Stratified Telephone Survey Designs. *Survey Methodology,* 19(1),103-113.

Cochran, W.G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24, 295-313.

Cochran, W.G., Mosteller, F., and Tukey, J.W. (1954). *Statistical Problems of Kinsey Report (on Sexual Behavior in the Human Male).* Washington, D.C.: American Statistical Association.

Cook, E.F., and Goldman, L. (1989). Performance of Tests of Significance Based on Stratification by a Multivariate Confounder Score or by a Propensity Score. *Journal of Clinical Epidemiology*, 42, 317-324.

Couper, M.P. (2000). Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P. (2001). The Promises and Perils of Web Surveys. In *The Challenge of the Internet*, ed. A. Westlake et al. London, UK: Association for Survey Computing.

Couper, M.P. (2002). Web Survey Design. *Unpublished Coursenote.*

Couper, M.P., and Tourangeau, R. (2002), Web-Based Survey Applications: A Review of Opportunities and Issues for NCHS. Report submitted to the National Center for Health Statistics.

Crown, W.H. (2001). Antidepressant Selectionand Economic Outcome: A Review of Methods and Studies from Clinical Practice. *The British Journal of Psychiatry*, 179, s18-s22.

Curtin, R., Presser, S., and Singer, E. (2000). The Effects of Response Rate Changes on the Index of Consumer Sentiment. *Public Opinion Quarterly*, 64, 413-428.

Czajka, J.L., Hirabayashi, S.M., Little, R.J.A., and Rubin, D.B. (1992). Projecting from Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics. *Journal of Business and Economic Statistics*, 10(2), 117-132.

D'Agostino, R.B. Jr. (1998). Propensity Score Methods for Bias Reduction for the Comparison of a Treatment to a Non-randomized Control Group. *Statistics in Medicine*, 17, 2265-2281.

D'Agostino, R.B. Jr., and Rubin, D.B. (2000). Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association*, 95(451) 749-759.

Danielsson, S. (2002). The Propensity Score and Estimation in Nonrandom Surveys - An Overview. Accessed from http://www.statistics.su.se/modernsurveys/publ/11.pdf.

Deming, W.E. (1944). On Errors in Surveys. *American Sociological Review*, 9(4), 359-369.

Deming, W.E., and Stephan, F.F. (1940). On A Least Squares Adjustment of A Sampled Frequency Table When the Expected Marginal Totals Are Known. *Journal of the American Statistical Association*, 35, 615-630.

Deville, J.C., and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

Deville J.C., Särndal, C.E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling, *Journal of the American Statistical Association*, 88(423), 1013-1020.

Dillman, D.A. (2000). *Mail and Internet Surveys. The Tailored Design Method.* Second Edition. New York, NY: John Wiley & Sons.

Dillman, D.A. (2002). Navigating the Rapids of Change: Some Observations on Survey Methodology in the Early 21st Century. Draft of Presidential Address to American Association for Public Opinion Research Annual Meeting. Accessed from http://survey.sesrc.wsu.edu/dillman/papers.htm.

Drake, C. (1993). Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. *Biometrics*, 49(4), 1231-1236.

Duncan, K.B., and Stasny, E.A. (2001). Using Propensity Scores to Control Coverage Bias in Telephone Surveys. *Survey Methodology*, 27(2). 121-130.

Frigoletto, F.D., Lieberman, E., Lang, J.M., Cohen, A.P., Barss, V., Ringer, S.A. and Datta, S. (1995). A Clinical Trial of Active Management of Labor. *New England Journal of Medicine*, 333, 745-750.

Gattiker, U.E. (2001). *The Internet as A Diverse Community: Cultural, Organizational, and Political Issues.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gelman, A., Carlin, J., Stern, H., and Rubin, D.B. (1995). *Bayesian Data Analysis.* Boca Raton, FL: Chapman & Hall

Groves, R.M. (1989). *Survey Errors and Survey Costs.* New York, NY: John Wiley & Sons.

Groves, R.M., and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys.* New York, NY: John Wiley & Sons.

Groves, R.M., and Kahn, R.L. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews.* New York, NY: Academic Press.

Heckman, J.J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1), 153-162.

Heckman, J.J., and Smith, J.A. (1995). Assessing the Case for Social Experiments. *Journal of Economic Perspectives*, 9, 85-110.

Heckman, J.J. (1997). Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. *The Journal of Human Resources*, 32(3), 441-462.

Hoffer, T., Greeley, A.M., and Coleman, J.S. (1985). Acheivement Growth in Public and Catholic Schools. *Sociology of Education*, 58, 74-97.

Huggins, V., and Eyerman, J. (2001). Probability Based Internet Surveys: A Synopsis of Early Methods and Survey Research Results. Paper presented at the 2001 Research Conference for the Federal Committee on Statistical Methodology.

International Telecommunication Union. (2003). *Internet Indicators: Hosts, Users and Number of PCs by Country.* Accessed from http://www.itu.int/ITU-D/ict/statistics/.

Jayasuriya, B., and Valliant, R. (1996). An Application of Restricted Regression Estimation to Post-Stratification in a Household Survey. *Survey Methodology*, 22**,** 127-137.

Keeter, S., Kohut, A., Miller, C., Groves, R.M., and Presser, S. (2000). Consequences of Reducing Nonresponse in a Large National Telephone Survey. *Public Opinion Quarterly*, 64, 125-148.

Kish, L. (1965). *Survey Sampling.* New York, NY: John Wiley & Sons.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukhopadhyay, T., and Scherlis, W. (1998). Internet Paradox: A Social Technology That Reduces Social Involvement and Psychological Well-being? *American Psychologist*, 53 (9), 1017-31.

Lavori, P.W. (1992). Clinical Trials in Psychiatry: Should Protocol Deviation Censor Patient Data? *Neuropsychopharmacology*, 6(1), 39-48.

Lavori, P.W., and Keller, M.N. (1988). Improving the Aggregate Performance of Psychiatric Disgnostic Methods When Not All Subjects Receive the Standard Test. *Statistics in Medicine*, 7, 723-737

Lee, S. (2003). An Evaluation of Nonresponse and Coverage Errors in a Web Panel Survey. Paper presented at the annual Joint Statistical Meeting, American Statistical Association, San Francisco, CA.

Lee, S. (2004). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. Paper presented at the annual meeting of the American Association for Public Opinion Research, Phoenix, AZ.

Leiner, B.M., Cerf, V.G., Clark, D.D., Kahn, R.E., Kleinrock, L., Lynch, D.C., Postel, J., Roberts, L.G., and Wolff, S. (2000). *A Brief History of the Internet.* Accessed from http://www.isoc.org/internet/history/brief.shtml

Lieberman, E., Lang, J.M., Cohen, A.P., D'Agostino, Jr. R, Datta, S., and Frigoletto, Jr. F.D. (1996). Association of Epidural Analgesia with Caesareans in Nulliparous Women. *Obstetrics and Gynecology*, 88, 993-1000.

Lepkowski, J.M. (1988). Telephone Sampling Methods in the United States. In. *Telephone Survey Methodology,* ed. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, and J. Waksberg, New York, NY: John Wiley & Sons.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* Second Edition. Hoboken, NJ: John Wiley & Sons.

Manfreda, K.L., (2001). Web Survey Errors. *Unpublished Doctoral Dissertation.* University of Ljubljana (Slovenia), Faculty of Social Science.

Merkle, D. and Edelman, M. (2002). Nonresponse in Exit Polls: A Comprehensive Analysis. In *Survey Nonresponse*. ed. R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, New York, NY: John Wiley & Sons.

Moore, J. (2002). *The Internet Weather: Balancing Continuous Change and Constant Truths.* New York, NY: John Wiley & Sons.

Mitofsky, W.J. (1970). Sampling of Telephone Households. *Unpublished CBS News Memorandum.*

Mitofsky, W.J. (1999). Pollsters.com. *Public Perspective*, June/July, 24-26.

Nie, N.H., and Erbring, L. (2000). *Internet and Society: A Preliminary Report*. Palo Alto, CA: Stanford Institute for the Quantitative Study of Society. Accessed from http://www.stanford.edu/group/siqss

Obenchain, R.L. (1999). *Propensuty Score Binning and Smoothing in Splus, Version 9911.* Accessed from http://www.math.iupui.edu/~indyasa/bobodown.htm.

Obenchain, R.L., and Melfi, C.A. (1997). Propensity Score and Heckman Adjustments for treatment Selection Bias in Database Studies. *Proceedings of the Biopharmaceutical Section*, American Statistical Aassociation, 297-306.

Presser, S., Blair, J., and Triplett, T. (1992). Survey Sponsorship, Response Rates, and Response Effects. *Social Science Quarterly*, 73, 3, 699-702.

Reid, E. (1991). Electropolis: Communication and Community on Internet Relay Chat. *Unpublished Honors Thesis*, University of Melbourne. Accessed from ftp://ftp.parc.xerox.com/pub/MOO/papers/electropolis

Rosenbaum, P.R., and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1), 41-55.

Rosenbaum, P.R. (1984a). From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment. *Journal of the American Statistical Association*, 79(385), 41-48.

Rosenbaum, P.R. (1984b). The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment, *Journal of the Royal Statistical Society*, Series A (General), 147(5), 656-666.

Rosenbaum, P.R., and Rubin, D.B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387), 516-524.

Rosenbaum, P.R., and Rubin, D.B. (1985a). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, 39(1), 33-38.

Rosenbaum, P.R., and Rubin, D.B. (1985b). The Bias Due to Incomplete Matching. *Biometrics*, 41(1), 103-116.

Rubin, D.B. (1973). Matching to Remove Bias in Observational Studies. *Biometrics*, 29(1), 159-183.

Rubin, D.B. (1978). Multiple Imputation in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse. *Proceedings of the Section on Survey Research Methodology*, American Statistical Association, 20-34.

Rubin, D.B. (1979). Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association*, 74(366), 318-328.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York, NY: John Wiley & Sons.

Rubin, D.B. (1997). Estimation from Nonrandomized Treatment Comparisons Using Subclassification on Propensity Scores. *Annals of Internal Medicine*, 127, 8(2), 757-763.

Rubin, D.B., and Thomas, N. (1992). Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions. *Biometrika*, 79(4), 797-809.

Rubin, D.B., and Thomas, N. (1996). Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics*, 52, 254-268.

SAS Institute, Inc. (1999). *SAS/STAT® User's Guide*, *Version 8*. Cary, NC: SAS Institute, Inc.

Schonlau, M., Fricker, R.D., Jr., and Elliott, M.N. (2002). *Conducting Research Surveys via E-mail and the Web*. Santa Monica, CA: RAND.

Schonlau, M., Zapert, K., Simon L.P., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R., and Berry, S. (2004). A Comparison between a Propensity Weighted Web Survey and an Identical RDD Survey. *Social Science Computer Review*, 22(1).

Slevin, J. (2000). *The Internet and Society.* Cambridge, UK: Polity Press.

Smith, P.J., Rao, J.N.K., Battaglia, M.P., Daniels, D., and Ezzati-Rice, T. (2000). Compensating for Nonresponse Bias in the National Immunization Survey Using

Response Propensities. *Proceedings of the Section on Survey Research Methods*, American Statistical Asosication, 641-646.

Spiegelhalter, D.J., Thomas, A., and Best, N.G. (1999). *WinBUGS Version 1.2 User Manual.* MRC Biostatistics Unit.

Stone, R.A., Oborsky, S., Singer, D.E., Kapoor, W.N., and Fine, M.J. (1995). Propensity Score Adjustment for Pretreatment Differences between Hospitalized and Ambulatory Patients with Community-Acquired Pneumonia. *Medical Care*, 33, AS56-66.

Taylor, H., and Terhanian, G. (2003). The Evaluation of Online Research and Surveys over the Last Two Years. *Unpublished Manuscript.*

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J.W., and Terhanian, G. (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), 127-135.

Taylor, H. (2000). Does Internet Research Work? Comparing Online Survey Result with Telephone Survey. *International Journal of Market Research*, 42(1), 58-63.

Terhanian, G. (2000). How to Produce Credible, Trustworthy Information through Internet-Based Survey Research. Paper presented at the annual meeting of the American Association for the Public Opinion Research, Portland, OR.

Terhanian, G., and Bremer, J. (2000). Confronting the Selection-Bias and Learning Effects Problems Associated with Internet Research. Research Paper: Harris Interactive.

Terhanian, G., Bremer, J., Smith, R., and Thomas, R. (2000). Correcting Data from Online Survey for the Effects of Nonrandom Selection and Nonrandom Assignment. Research Paper: Harris Interactive.

Toffler, A. (1991). *Powershift: Knowledge, Wealth, and Violence at the Edge of the 21st Century.* New York, NY: Bantam Books.

Toffler, A. (1980). *The Third Wave.* New York, NY: William Morrow.

Toffler, A. (1970). *Future Shock.* New York, NY: Bantam Books.

Turkle, S. (1995). *Life on the Screen: Identity in the Age of the Internet*. New York, NY: Simon & Schuster.

U.S. Department of Commerce (2002). *A Nation Online: How Americans Are Expanding Their Use of the Internet*. Accessed from http://www.ntia.doc.gov/ntiahome/dn/.

Valliant, R. (2004). The Effect of Multiple Weighting Steps on Variance Estimation. To appear in *Journal of Official Statistics*.

Varedian, M., and Försman, G. (2002). Comparing Propensity Score Weighting with Other Weighting Methods: A Case Study on Web Data. Paper presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg Beach, FL.

Vativarian, S., and Little, R. (2003). On the Formation of Weighting Adjustment Cells for Unit Nonresponse. University of Michigan Department of Biostatistics Working Paper Series.

Vehovar, V., and Manfreda, K.L. (1999). Web Surveys: Can the Weighting Solve the Problem? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 962-967.

Venables, W.N., Smith, D.M., and the R Development Core Team. (2003). An Introduction to R©.

Waksberg, J. (1978). Sampling Methods for Random Digit Dialing. *Journal of the American Statisical Association*, 73, 40-46.

Westat (2000). *WesVar™ 4.0 User's Guide*. Rockville, MD: Westat.