Weighting for Coverage Bias in Internet Surveys

SDSU REUT 2007

Zachary Boboth
Alisha Zimmer
Mark Scott
Eloise Hilarides
Donovan Gromet
Craig Massey

Advisor: Dr. Kristin Duncan

July 27, 2007

Abstract

Over the past decade, Internet surveys have become a popular method for collecting data about the general population. In 2005, the Harris Poll published findings which claimed that 74% of the United States Population had access to the Internet access somewhere. While this number has steadily risen over recent years, bias still may be introduced if the population without Internet access is different from the Internet population in regards to the variables of interest. We study whether Internet users that only have access to the Internet outside their home can be useful in reducing bias by assuming that they are more similar to those without Internet access than the Internet population as a whole.

This paper outlines several weighting adjustment schemes aimed at reducing coverage bias. Data for this study was taken from the Computer and Internet Use Supplement of October 2003 administered by the Current Population Survey. We evaluate the schemes based on overall accuracy by considering the reduction in bias for ten variables of interest and the variability of estimates from the schemes. We find that several of our proposed schemes are successful in improving accuracy.

KEY WORDS: Coverage Bias, Weight Adjustments, Internet Surveys, Propensity Scores

Contents

1 Introduction **5**

2 Internet Access in the United States **9**
2.1 History 9
2.2 Comparison of Populations with and without Internet Access . 12

3 Internet Surveys **17**
3.1 History 17
3.2 Advantages and Disadvantages 20

4 The Current Population Survey **23**
4.1 Background 23
4.2 Design and Administration 24
4.3 Weighting 26
4.3.1 Raking 27
4.4 Variance Estimation 28
4.5 The Computer and Internet Use Supplement 29
4.6 Missing Data 29

5 Propensity Scores **31**
5.1 Logistic Regression 31
5.1.1 ROC Curve 34
5.2 Propensity Scores in Surveys 35

6 Weighting Schemes for Coverage Bias **37**
6.1 Base Scheme and Target 37
6.2 Raking by Internet Access Status 39
6.3 Propensity Scores 40

6.3.1 Modeling Variables 40
6.3.2 Propensity Scores 40
6.3.3 Missing Values 41
6.3.4 Modified Internet Access 42
6.3.5 Full Sample 44
6.4 Frequency of Internet Use Scheme 45

7 Results **47**
7.1 Variance Estimation Via Replication 48
7.2 Assessing Accuracy 49
7.3 Comparisons of Schemes 51
7.3.1 Effect of Raking by Internet Access Status 51
7.3.2 The Effect of Propensity Scores 53
7.3.3 Frequency of Use Scheme 59

8 Conclusions **61**

Appendices **65**

A **67**

B **69**

C **73**

D **75**

Bibliography **83**

Chapter 1

Introduction

The overall challenge in any survey, regardless of the mode of administration, is to maximize the efficiency of the interview process without jeopardizing the integrity or representativeness of the information collected. In recent decades, telephone interviews have been the primary means of collecting data from the general public. Phone surveys are relatively inexpensive, a large sample size can be obtained fairly easily, and methods to ensure a representative sample have been studied extensively [4]. In the past few years, Internet surveys have become an attractive mode for data collection [15]. Whether as simple as a pop-up that asks a specific user to take part in a brief opinion poll, or as complex as Web sites devoted to survey research, the Internet is a convenient, cost effective method for gathering a substantial amount of information in a short period of time. In addition, Internet surveys significantly reduce interviewer effects that can be problematic with in-person, as well as phone interviews. As of July 2005, Nielsen/Netratings ranked the United States as the country with the sixth highest Internet penetration rate at 69.6% [22]. While this information indicates a steady increase in accessibility to the Internet since the mid 1980's, it does not assure that Internet surveys obtain representative samples.

Coverage bias may occur in any study in which the sampling frame, the set of individuals from which the sample is drawn, does not match the target population [19]. Just as phone surveys exclude households without telephones, Internet surveys exclude individuals without access to the Internet. Thus, if an Internet survey polls the Internet population about a topic that may be correlated with whether or not the respondent has a computer or has access to the Internet, with the intent of generalizing the results to the full

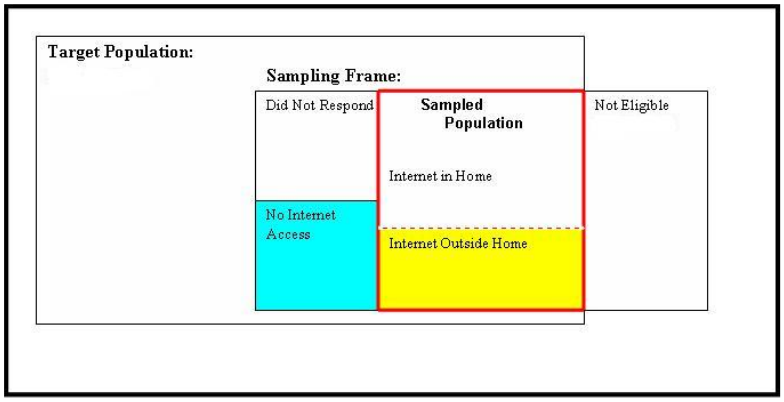


Figure 1.1: Internet Populations for Generalized Internet Surveys

population, the results of the study may be biased.

A fair amount of research has been done on minimizing coverage bias in telephone surveys. In a previous study, Duncan and Stasny [11] used propensity scores in conjunction with poststratification methods to adjust for the non-telephone population. In the study, individuals interviewed were classified as transients if they had any stoppage in phone service in the last year. These authors assumed, based on previous research (c.f. [3]), that transients were more representative of the population without telephone service than were individuals that have maintained service. Their findings illustrated that weighting schemes can be advantageous when propensity scores are used in conjunction with poststratification. However, without any benchmarks against which to compare their schemes, further research was warranted to verify this assumption.

This study attempts to build upon the previous research done on coverage bias and apply it to the realm of Internet surveys. We extend the previous assumption that transient status individuals are representative of the non-telephone population to Internet populations. As Figure 1.1 shows, Internet users can be broken into three groups; Internet access at home, Internet access outside of the home only, and no Internet access at all. Those with Internet access outside of the home only are labeled as transient, since they move in and out of the Internet population. If these respondents are more representative of the non-Internet population, weighting schemes that take advantage of this may reduce coverage bias.

Data for this study was obtained from the Current Population Survey, and more specifically the October 2003 Computer Use and School Enrollment Supplement. For a detailed description of this survey, refer to Chapter 4. The noteworthy aspect of this surveys is that it is conducted in-person. As a result, data is collected on all three distinct Internet populations. Thus, this study not only has a wealth of data to construct modeling schemes, but also provides a framework, the no Internet population, against which to test the proposed weighting schemes.

Chapter 2 provides a brief summary of Internet accessibility in the United States to give a framework for this research. Chapter 3 discusses the attractions of and problems with surveying via the Internet. Chapter 4 follows with an outline and its complex survey design. Chapter 5 examines the theory behind creating propensity scores through logistic regression. There is also a brief literature review of propensity scores in the realm of telephone surveys. In Chapter 6, the multiple weighting schemes from this study are introduced along with a discussion of creating the target values, against which schemes are compared. Chapter 7 provides an in-depth comparative analysis of the proposed schemes with attention to the bias/variance tradeoff associated with weighting. Chapter 8 summarizes our findings and provides our recommendations for future studies.

Chapter 2

Internet Access in the United States

Twenty-five years ago, the Internet was unheard of and today people can go on online at their local coffee shop. As technology has continued to advance the Internet seems to have become a necessity, but close to 30% of American adults still do not have access to the Internet. In this chapter we will discuss how Internet access has spread through the U.S., as well as looking at the differences among those who do and do not have Internet access.

2.1 History

The first multi-site computer network, called ARPAnet, was created by the Advanced Research Project Agency (ARPA) within the U.S. Department of Defense in 1969 and is considered to be the genesis of the Internet. In 1972 the first e-mail program was created by Ray Tomlinson of BBN. Twenty years later the World-Wide Web was founded after the National Science Foundation (NSF) lifted restrictions against the commercial use of the Internet. By the mid 1990's most Internet traffic was carried by independent Internet Service Providers, including MCI, AT&T, Sprint, UUnet, BBN Planet, and ANS.

The Internet did not start to become popular amongst the general public until the 1990's. The Harris Poll [16] reported in 1995 that nine percent of U.S. adults were online whether at home, work, or another location. This number has grown to 74 percent in 2005 [16]. Another study by Pew Internet and American Life Project reported the online population expanded from roughly 86 million Americans in March 2000 to 126 million in August 2003, an increase of 47% [20].

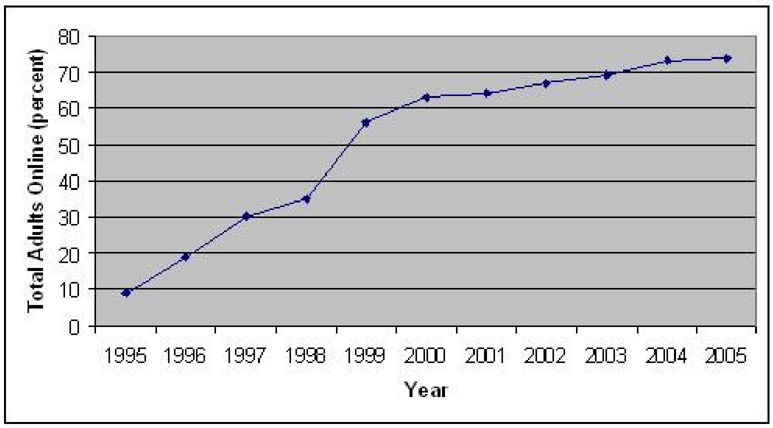


Figure 2.1: Adults Online (Harris Poll)

The increase in Internet penetration has come hand in hand with growth in computer access. Internet usage has been fueled by a variety of features and services such as games, online banking, news, and email. More than half of all U.S. adults used e-mail or instant messaging in 2003, compared with 12 percent of adults in 1997 [10]. Internet users discover more things to do online as they gain experience and as new applications become available [20]. Figure 2.2 displays the types of activities in which adults engage online.

As the utility of the Internet expands people use the Internet more often. In 2005 adults spent an average of nine hours per week online opposed to seven hours per week in 1999 [16]. Figure 2.3 shows that independent of the age category frequent Internet use is the norm.

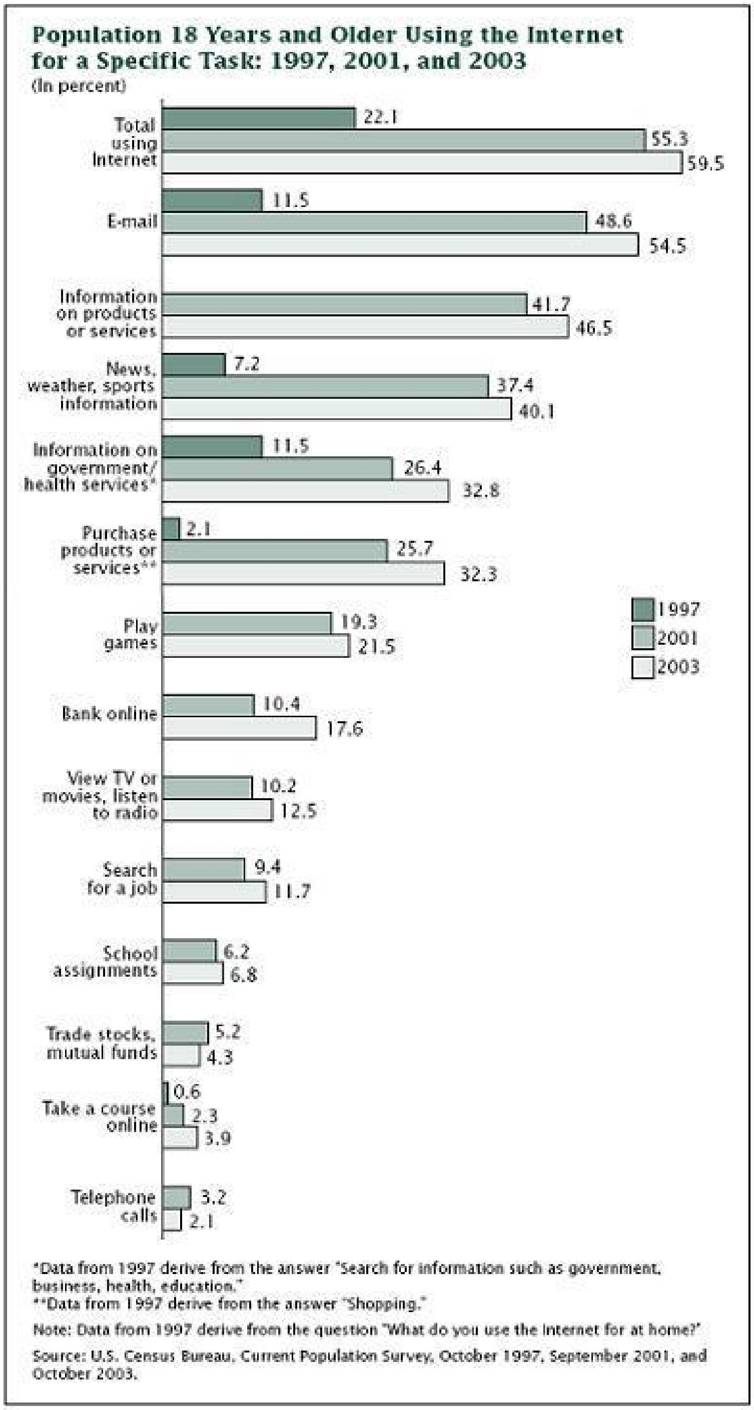


Figure 2.2: A Look at What Adults Do Online

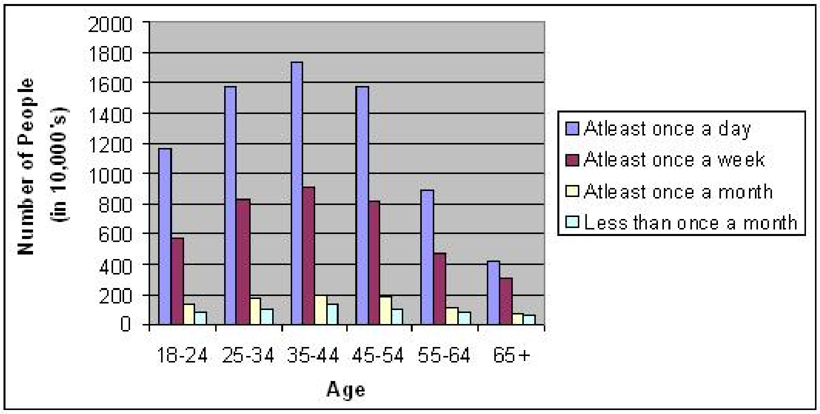


Figure 2.3: Internet Usage by Age Group

2.2 Comparison of Populations with and without Internet Access

In order to understand how coverage bias may affect an Internet only survey, we consider the differences between populations with access in the home, outside the home, and not at all. In particular, we are interested in knowing to what extent those with access outside the home only can be represented of those with no access.

Those who do not have Internet access are more likely to be black or Hispanic, poor, poorly educated, disabled, and elderly [26]. In addition, those with jobs are more likely than those without jobs to have access, parents of children under 18 living at home are more likely than non-parents to be online, and rural Americans lag behind suburban and urban Americans in the online population [20]. In Figure 2.4 below it is easy to see the inverse relationship between income level and Internet access. Similarly in Figure 2.5 we can see that employed people are more likely than unemployed, disabled, and retired persons to have Internet access.

In 2003, Madden [20] wrote that about a quarter of Americans live lives that are quite distant from the Internet. They have never been online, and don't know many others who use the Internet. At the same time, many Americans who do not use the Internet now were either users in the past or they live in homes with Internet connections.

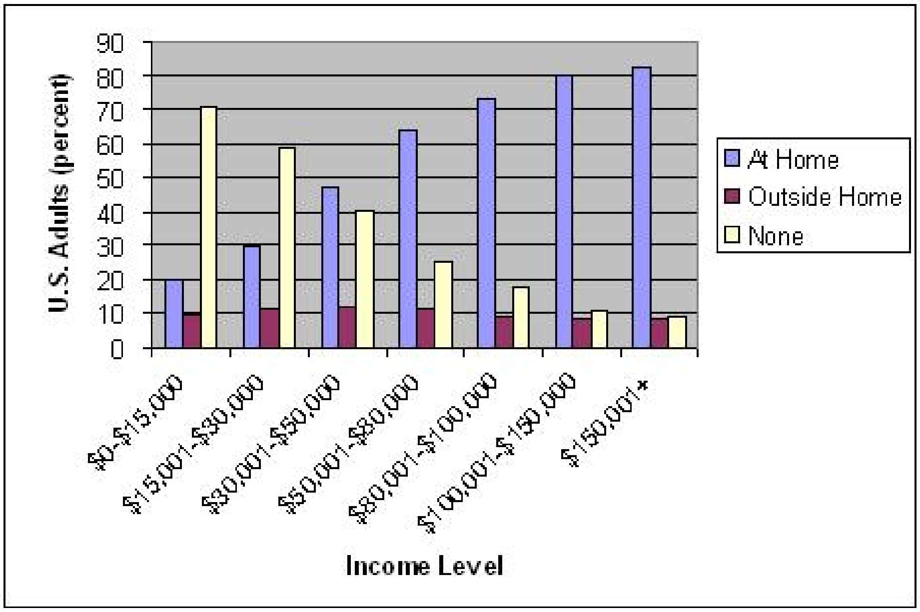


Figure 2.4: Internet Access by Income (2003)

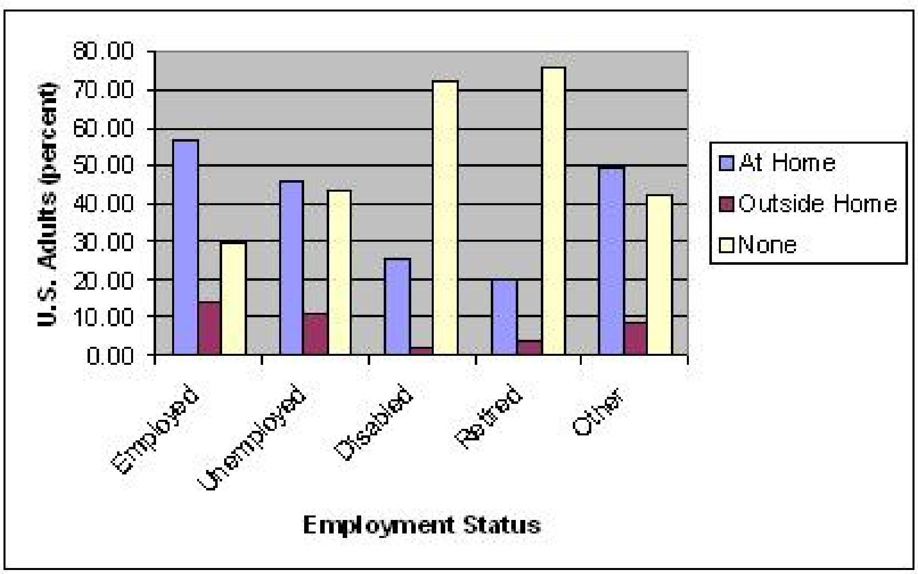


Figure 2.5: Internet Access by Employment (2003)

Using the Computer and Internet Use supplement from the Current Population Survey 2003, which will be discussed in the next chapter, we were able to make comparisons between those with access in the home, outside the home only, and no access at all. When looking at race the Internet access outside the home only was much closer to those with no Internet access at all, as seen in Figure 2.6. This is important since 55% of blacks have no Internet access, opposed to only 39% of whites. Figure 2.6 also shows how whites are over-represented in the Internet population while minorities are under-represented.

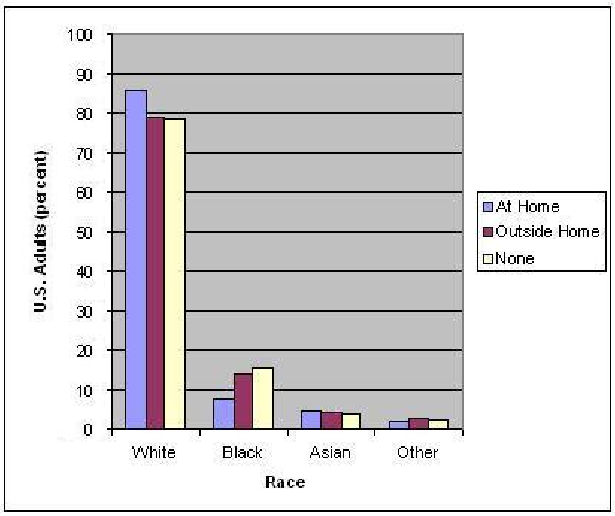


Figure 2.6: Internet Access by Race (2003)

However race is not the only category of interest, the same result is found in level of school completed, as seen in Figure 2.7. And looking at Figure 2.8 those with Internet access outside the home only are closer to those with no Internet access, aside from the income bracket $30,001 to $50,000. This makes sense since 47% of those in that income bracket have Internet access at home and 40% have no Internet access. In addition, referring back to Figure 2.4 we see low income adults are under-represented in the Internet population with 70% of people in the income bracket $0 to $15,000 have no Internet access. On the other hand we see that high income adults are over-represented in the Internet population with 82% in the $150,000+ income bracket.

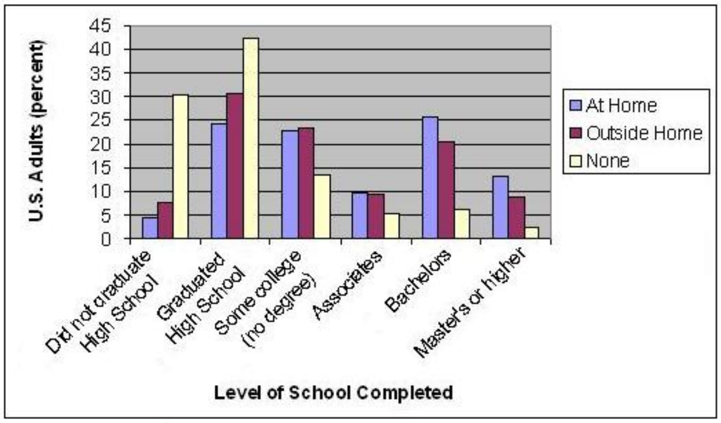


Figure 2.7: Internet Access by Level of School Completed (2003)

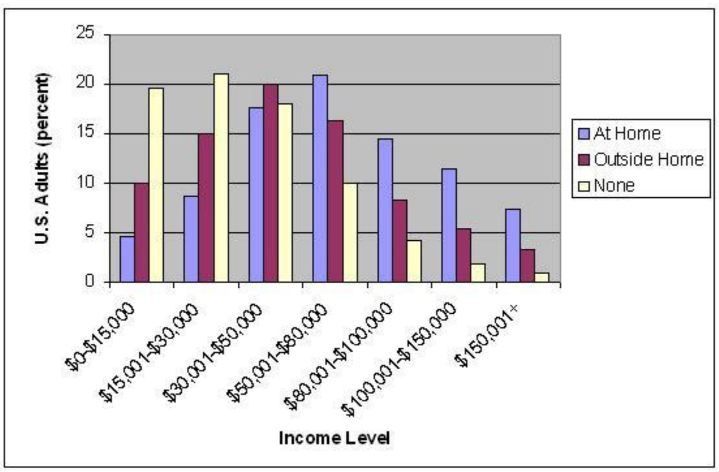


Figure 2.8: Internet Access by Income (2003)

Those with Internet access outside the home provide a bridge between those with Internet access at home and no Internet access at all. This is key to our study, as we can now use those with Internet access outside the home only to represent those with no Internet access.

Chapter 3

Internet Surveys

In this chapter we provide background information on Internet surveys, their classification, and the pros and cons of this survey mode.

3.1 History

In the late 1980's and early 1990's, prior to the widespread use of the Web, researchers began exploring email as a survey mode [24]. Originally email surveys were text-based and tended to resemble paper surveys structurally. Subjects were required to type their responses to the questions in an email reply as opposed to clicking on radio boxes or buttons. At this point in time, the only advantages e-mail surveys offered over traditional survey methods were reduced cost, delivery time, and response time.

As the Web became more widely available in the early to mid 1990's, it quickly supplanted email as the Internet survey medium of choice [24]. The Web provided a more interactive interface capable of incorporating audio and video into surveys. Also, Web surveys offered a way around the need for a sampling frame of email addresses. In recent years, technology has closed the gap in capabilities between Web and email surveys.

The Internet has opened new doors for the world of surveying, taking many survey methodologists by surprise. Leadership in Web survey development has come from people with a background in technology rather than the survey methodology professionals because of the computational effort that was required to construct Web surveys early on [25]. Now, software capable of producing survey forms is available to the general public at an affordable cost, enabling anyone with access to the Internet to conduct a survey with little difficulty [15].

Much of the focus of research on Internet surveys has been on Web design and layout of the questionnaires; whether to use radio boxes or drop down menus, for example. Couper [9] explained how inaccuracies in computer programming, which produced text boxes of different sizes, affected survey results in a University of Michigan survey. In recent years, researchers' focus has shifted from questionnaire formatting to investigating the validity of Internet surveys. The major sources of error of any survey include sampling, coverage, nonresponse, and measurement error, all of which are particularly worrisome for Web surveys [8]. Coverage, especially, has become a big concern [15].

There are many different ways to conduct an Internet survey. Couper [8] classified eight types of web surveys under the categories of nonprobability methods and probability-based methods. The difference between the two categories is that in nonprobability surveys, members of the target population have unknown nonzero probabilities of selection.

The nonprobability methods are as follows:

1. Web surveys as entertainment. These surveys are intended for entertainment purposes and are not surveys in a scientific sense. Examples include the "Question of the day" type polls found on many media Websites,

2. Self-selected Web surveys. This type of Web survey uses open invitations on portals, frequently visited Web sites, or dedicated "survey" sites. An example is National Geographic Society's "Survey 2000." This is probably the most prevalent form of Web surveys today and potentially one of the most threatening to legitimate survey enterprises because publishers or users often mistakenly claim the results to be scientifically valid [8].

3. Volunteer panels of Internet users. The Harris Poll Online is an example of such a survey. This survey approach creates a volunteer panel by wide appeals on well-traveled sites and Internet portals. The panel creates a large database of potential respondents for later surveys. These later surveys are typically by invitation only and controlled through email identifiers and passwords.

The probability-based methods are as follows:

1. Intercept surveys. This survey approach generally uses systematic sampling to invite every nth visitor to a site to participate in a survey. This type of Web survey is very useful for customer satisfaction surveys and site evaluations.

2. List-based samples of high-coverage populations. This type of survey approach is used when all or most of a subset of the population has Internet access. An example would be surveys of college students.

3. Mixed-mode designs with choice of completion method. This is popular in panel survey establishments, where repeated contacts with respondents over a long period of time are likely. This type of mixed mode survey offers the Web as a form of survey completion.

4. Pre-recruited panels of Internet users. This is similar to the volunteer panels of Internet users with the main difference being that panelists are recruited instead of volunteering. Recruitment takes place using probability methods similar to RDD for telephone surveys.

5. Probability samples of full population. This method has the potential for obtaining a probability sample of the full population, not just those who have Web access [8]. Here non-Internet approaches are used to elicit initial cooperation from a probability sample of the target population.

These are just some of the methods used, and not an all inclusive list, as there are other types of Web surveys and many variations on the ones presented here. Probability-based sampling designs yield more statistically sound results. The applications of this study focus on probability-based methods such as pre-recruited panels of Internet users and mixed mode surveys.

Internet surveys are a relatively new form of data collection and are rapidly increasing in popularity. As access to the Internet penetrates society further, the use of Internet surveys will continue to grow too.

3.2 Advantages and Disadvantages

As we have noted already, the Internet is quickly becoming a more popular platform for administering surveys due to the many advantages it offers over other conventional methods. When compared to more traditional methods, such as phone surveys and face-to-face interviews, Web surveys are much more efficient in terms of cost and time. A study by Schonlau [24] showed that their Internet survey cost $10 per completed case versus $51 for each completed case via the telephone. The same study also showed that the average time for a case completed on the Internet was 3.5 weeks, versus 3 months for the telephone cases. While this is strong evidence for the cost effectiveness of Internet surveys, there may not always be such a discrepancy between costs. It has been suggested that there is a minimum threshold of responses, between 2,000 and 4,000, for which Internet surveys ultimately prove to be much more affordable than more traditional methods [24]. Being able to get a large number of surveys out to the general public in a relatively short time span eases the burden of the recruitment process for convenience samples [12].

Another advantage of administering surveys on the Internet is that people tend to be more honest when completing a survey online when compared to other traditional modes [13]. The Internet can provide a sense of anonymity which other traditional methods of administering surveys lack. This allows a respondent to answer sensitive questions more honestly and admit to more embarrassing behavior. In sum, Internet surveys relieve the pressure on the respondent to give socially desirable answers, which can lead to more accurate research results.

Internet surveys also give respondents more motivation, less distraction, and greater flexibility than more traditional methods [14]. Internet surveys, unlike any other method of surveying, have the capability to "provide a unique advantage for motivating participants to respond seriously: appealing to peoples' desire for self-insight by providing interesting, immediate feedback. Participants are motivated to answer seriously to receive accurate feedback about their personality" [14]. While not all surveys offer this immediate feedback, Internet surveys are the only type of surveys that are capable of doing so. Thus, if a researcher is looking to motivate possible respondents by providing instant feedback, an Internet survey has the unique advantage of offering this motivation.

In addition, surveys administered online, unlike phone surveys, eliminate the need for the respondent to call upon his or her short term memory. A participant who is completing a survey online may re-read the questions several times, leading to more accurate answers [13]. This would have the most impact in surveys with open-ended questions, when the possibility for complex answers is heightened.

Internet surveys also provide the respondent more flexibility. Instead of having to complete the survey at a particular place or time, they have access to the survey twenty-four hours a day. They can also complete the survey at their own pace. The overall convenience of Internet surveys can lead to more accurate results [14]. Finally, Internet surveys provide an interface that is friendlier for the participants as well as the researchers. Internet and email surveys can mesh the use of text, sound, video, and live interaction unlike any other method for administering a survey. Also, an Internet survey has the advantage of easily being able to adjust the questions being presented based on how the respondent has answered previous questions [27]. This is clearly an advantage over paper questionnaires, which may have confusing skip patterns. In addition, Internet surveys lighten the burden of researchers when it comes to data entry. Web-based surveys are able to export data collected from a survey directly into analysis software packages, bringing about data that is free from key-in error by human data processors.

While the advantages of administering a survey via the Internet are numerous, there are also some serious pitfalls that need to be taken into consideration. The biggest of these issues is assessing the integrity of the data provided by a Web survey. The issue of data integrity is based on problems of coverage, sampling and nonresponse in Internet surveys.

As previously mentioned in Chapter 2, approximately 75 percent of the United States population had Internet access in their household as of 2005. This presents a problem when it comes to coverage bias involving Internet surveys. We have already noted that persons who do not have the Internet are more likely to be minorities, elderly, and of lower income. Thus, any results derived from an Internet survey will be hard to generalize to the public as a whole because these groups will not be properly represented. It is the goal of the current study to provide a solution to this coverage bias problem through weighting schemes.

Creating an appropriate sample for an Internet survey can also cause some major problems. There is not an exhaustive or comprehensive list of Internet users, and thus there is no method for sampling Internet users that is similar to the random digit dialing technique for creating samples in phone surveys [13]. There are also concerns when using convenience samples for Internet surveys. Self selection for certain types of Internet surveys, such as pop-ups and banners, may only be appealing to those who feel strongly about the issue of the survey, thus leading to polarizing results which may not be applicable to the general population [27].

There is also the issue of nonresponse in regards to Internet surveys. If the survey is done through a self selection process, then there is no way to account for nonresponse bias since we have no method of finding out who the sample members were to serve as our base [13]. If nonrespose rates can be calculated, the rates are usually very high compared to other traditional methods [13]. Also the variability of nonresponse within Internet surveys is very high. The range of someone not receiving an invitation to a survey via email varies from 1 to 20 percent. This variation is due to the quality of the target email list which the researchers use to contact possible respondents. The variability of someone actually completing the survey given the invitation ranges from 1 to 96 percent [21]. This high variability is mainly due to the presentation of the survey and the credibility of the organization conducting the survey. Also, there is an issue of respondents not taking an Internet survey as seriously as other types of more traditional surveys, which may cause them to rush through the questions and leave some unanswered.

Finally, psychometric properties like test-retest reliability can be difficult to account for in Internet surveys. A survey is considered reliable if it produces consistent results. One way to see if a survey is reliable is to administer the survey to a respondent, and then re-administer the same survey to the same respondent at a latter time and see if the results are similar. If an Internet survey is distributed through a pop-up or banner, then it is likely that the researcher will be unable to contact the participant for a follow up survey [13]. Because psychometric properties are difficult to test in Internet surveys, the integrity of the data may be questionable.

Chapter 4

The Current Population Survey

In order for us to test our proposed weighting schemes to compensate coverage bias in Internet surveys, we need survey data. The data resource we use in our study is the Current Population Survey (CPS). Data from the CPS is available from the Census Bureau through a program called *Data Ferrett*. Before we begin applying and testing our weighting schemes, however, we must understand how the survey was designed and administered and the estimation procedures used.

4.1 Background

The CPS is a monthly survey that reaches approximately 50,000 households. It is conducted by the *Bureau of the Census* for the *Bureau of Labor Statistics* (BLS). The survey's primary goal is to obtain information about the labor force characteristics of the U.S. population (both nationwide as well as statewide). In addition, there are many questions referencing population demographics, and thus the survey is an appropriate means for gathering summary data about the U.S. population as a whole between decennial censuses.

The survey is designed to be as close as possible to a probability sample, meaning that each household has the same probability of being selected to participate, and to target the whole non-institutionalized population aged 16 years and older. For a more in-depth analysis of the methods, reference CPS Technical Paper 63 [17]. The following is a summary of the survey design, implementation and estimation procedures.

4.2 Design and Administration

The BLS uses a two-stage stratified design for the survey in order to reduce variance and collection costs while staying as close as possible to a probability sample. A simple example of these stages is shown in Figure 4.1. In the first stage, the United States is broken down into contiguous, non-overlapping primary sampling units (PSUs). Each PSU is fully contained within a single state and is chosen to be as heterogeneous as possible with respect to demographic characteristics but small enough for a field representative to feasibly traverse. PSUs are typically composed of several neighboring counties or metropolitan areas (see Figure 4.1 (1)). Next, similar PSUs are grouped into strata based on population size and homogeneity. An algorithm is used to satisfy these characteristics as best as possible. These strata do not need to be geographically contiguous, but do need to fall within state borders (shown in Figure 4.1 (2)). Some PSUs, usually those containing large metropolitan areas, are self-representing because they are large enough to become their own stratum. Of the strata that consist of more than one PSU, the Maximum Overlap Algorithm is used to choose one from each to represent the whole stratum (shown in Figure 4.1 (3)). This algorithm minimizes costs by choosing a previously surveyed PSU so new field representatives do not have to be trained. At this point, all households are either a part of a self-representing PSU or are represented by or are representing other PSUs in their strata.

The next step is to choose ultimate sampling units (USUs) within the representing PSUs. USUs are made up of clusters of four expected housing units or housing unit equivalents (shown as number (4) in Figure 4.1). This clustering may increase the within PSU variance, but is necessary to meet time and cost constraints. A list of households is acquired from the Master Address File of the 1990 Decennial Census of Population and from the Building Permit Survey (also conducted by the BLS). Living quarters are classified as housing units or group quarters and then grouped into four types of frames. The unit frame consists of housing units with a high proportion of compete addresses; the group quarter frame consists of group quarters with a high proportion of complete addresses; the area frame includes housing areas with significant incomplete addresses; the permit frame includes addresses of houses that have permits but may or may not be built. Addresses in the group quarter, area, and permit frame are equivalent to four housing units. USUs are then sorted based on demographic variables using the previous decennial census. An algorithm is then used to choose USUs from each PSU in a way that minimizes variance and cost. The houses in these USUs are the ones that actually get chosen for the surveys.

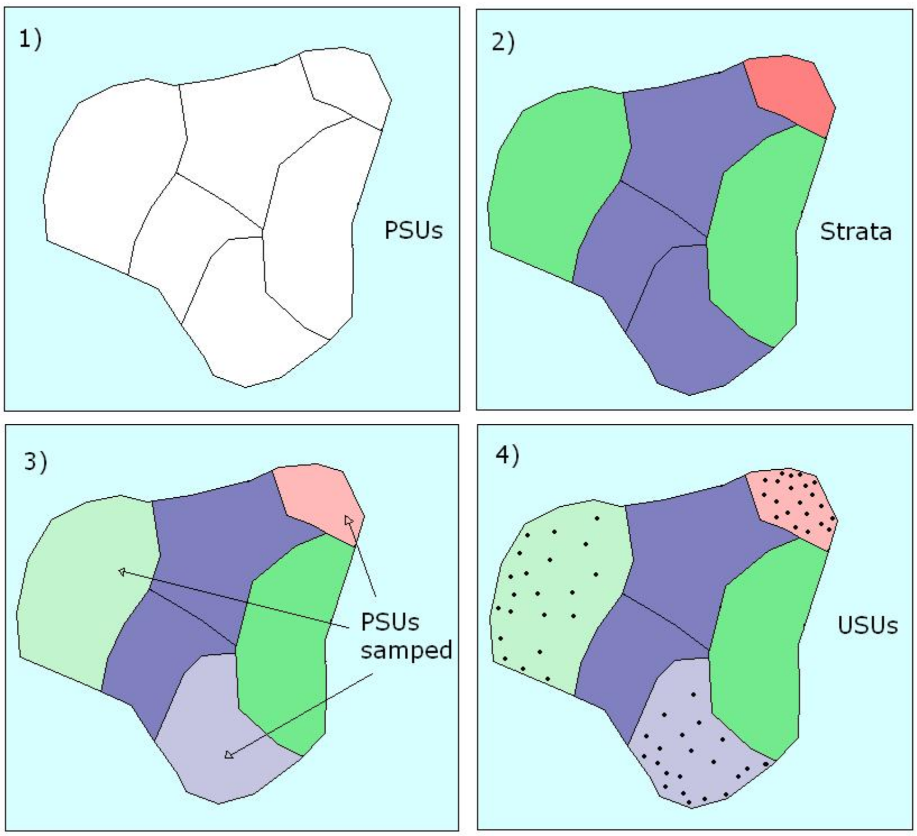


Figure 4.1: The two-stage design of the CPS

When the surveys are administered, housing units are interviewed in a 4-8-4 pattern. This means one interview a month for four months, eight months off, then four final interviews during the next months. This is done to minimize the following due to sampling mechanism: variance of month-to-month change ($\frac{3}{4}$ of the sample will be the same in consecutive months), variance of year-to-year change ($\frac{1}{2}$ of the sample is the same in the same month of consecutive years), and response burden (eight interviews dispersed over sixteen months).

The survey is administered with the intent to fulfill the following goals: implement the sampling procedures outlined in the design, produce complete coverage, discourage individuals being surveyed more than once within the decade, and be cost efficient. This process includes identifying addresses, listing living quarters, assigning field representatives, and conducting the interviews. The questionnaire used to conduct the survey remained unchanged from 1967 until 1994. The radical changes in 1994 were made to exploit the capabilities of the computer assisted personal interviewing (CAPI) and computer assisted telephone interviewing (CATI) programs as well as to adapt to economic and social changes that have happened over time. These changes include the growth in the number of service-sector jobs and the decline in the number of factory jobs, the more prominent role of women in the workforce, and the growing popularity of alternative work schedules. The redesign of the survey also attempted to reduce the potential for response error by making questions shorter and clearer.

4.3 Weighting

The information obtained from conducting the interview is transmitted to central locations for analysis. Weights must be applied to the raw counts to obtain an accurate and precise estimate for the entire population. First, the information for each sample unit is multiplied by the reciprocal of the probability with which that unit was selected. This creates unbiased estimates because it obtains probabilities through the sample design. These probabilities are state-specific and can be found in Table 3.1 of the technical report [17].

The next step is to account for nonresponse. Nonresponse can come in two forms: item nonresponse and complete (or unit) nonresponse. Item nonresponse edits are applied using one of three imputation methods. Relational imputations infer the missing value from other characteristics on the person's record or within the household. This technique is used exclusively in the demographic and industry and occupation variables. Longitudinal edits are used primarily in the labor force variables. These edits use the last month's data if it is available. If these methods cannot be used, 'hot deck' allocation assigns a missing value from a record with similar characteristics. For unit nonresponse (households in which members refuse, are absent, or are unavailable), households are grouped into clusters and the weights of interviewed sample units are increased to account for nonresponding units. The cells are determined by the Metropolitan Statistical Area (MSA) status and MSA size. This assumes that the households that do not respond are randomly distributed in relation to socioeconomic and demographic characteristics.

After adjustments for nonresponse, two stages of poststratification are applied. Information about demographics is obtained from outside sources to make comparisons of the population. The first stage adjusts weights for all cases in each selected NSR PSU for possible imbalance of black/non-black representation caused by PSU selection. The BLS admits that further research is needed to determine whether this adjustment is in fact meeting its purpose. The second stage adjustment is used to ensure that sample-based estimates of the population match independent population controls using a method called raking.

4.3.1 Raking

Poststratification is a technique used to reduce bias due to coverage error and raking is a specific type of poststratification. When the true population percentages are known for particular characteristics, they can be compared to the percentages found in the survey. Raking is necessary when the marginal distributions of poststratification variables are known but the joint distribution of these variables is unknown. If a particular subset of the population is underrepresented in the sample, the weights of individuals in this subset can be increased so their proportion matches the true proportion in the population. Simultaneously, the weights of units with the overrepresented characteristic are decreased so that the population total stays constant. The raking process becomes useful when there are multiple characteristics that need to converge to population proportions. For example, if true proportions are known about sex and race then adjusting the weights to the correct proportion of males and females disrupt the proportions in the race category. The weights must then be adjusted for race. This process goes back and forth through multiple iterations until the proportions in both categories converge to the true population proportions. In the CPS, a three way rake (by state, Hispanic/sex/age, and race/sex/age characteristics) is repeated through six iterations. Later we will use this raking process to recreate the weights in the CPS and also in concert with other weighting schemes.

4.4 Variance Estimation

Sampling variability is inherent in any sample survey, making it necessary for administrators and analysts to provide variance measures in addition to point estimates that are reported from the survey. Due to the substantial number of variables the CPS collects, it is not realistic for the Census Bureau to provide individual measures of error. For example, a data analyst looking for information regarding race and gender relationships would need 42 point estimates as well as 42 standard errors because there are 21 race categories and two gender categories coded within the CPS. It is easy to see that this quickly becomes unmanageable with a total of 374 variables in the CPS.

Instead, the CPS uses a generalized variance function (GVF) to provide these error estimates. This GVF is based on a modified half sample replication method [29]. Through experimentation, the U.S. Census Bureau has found that certain groups of estimates have consistent relationships between their point estimates and measures of variability [5]. As a result, the Census Bureau publishes a list of generalized variance parameters that can be used in conjunction with specific functions to elicit an estimated standard error. The generalized function for providing standard error estimates for percentages obtained from the CPS is given by

$$s_{x,p} = \sqrt{\frac{b}{x}p(100-p)}, \tag{4.1}$$

where b is the generalized variance parameter, x is the base population being analyzed, and p is the point estimate of the percentage. Other generalized equations for variance estimation exist if the point estimate is a total or a difference between two statistics. However, we are only concerned with estimating standard errors for percentages.

As with any generalized estimation procedure, caution should be exhibited in applying this function. For example, the CPS questions the validity of these estimates when smaller sample sizes are used, especially below 75,000 cases. In addition, the GVF is sensitive to adjustments within the data set. As a result, any weighting scheme, other than the CPS person weights, cannot apply these parameters legitimately.

4.5 The Computer and Internet Use Supplement

In addition to the general labor statistics and the demographic data collected on a monthly basis by the CPS, the BLS conducts specific supplemental inquiries throughout the calendar year. Of interest to this study was the School Enrollment and Computer Use Supplement designed by the National Telecommunications and Information Administration (NTIA), part of the United States Department of Commerce. This particular supplement is administered approximately every two years, most recently in October 2003, and targets the United States Population three years of age and older.[5]

The Census Bureau staff conducts the supplement in conjunction with the CPS to obtain information on the use of computers, the Internet and other emerging technologies by American people. The survey includes questions encompassing computer accessibility, specific uses of computers (school, finances, business, gaming), Internet and email use, and overall comfort regarding Internet security. The NTIA's major publication, *A Nation Online* [26], summarizes these findings as new information is made available.

As with most surveys, nonsampling error was inherent in the NTIA supplement. The overall nonresponse rate for the October 2003 supplement increased to 13.1% from 7.3% in the general CPS. In addition, item nonresponse was also an issue. Item nonresponse occurs when specific respondents do not provide answers to specific questions and/or portions of the survey. Unlike in the general CPS methodology, the NTIA did not impute values for missing responses, nor did they provide a supplement specific weighting scheme[26]. Thus, the person weight from the October 2003 monthly CPS can be applied to the Computer Use portion of the Supplement.

For a more detailed analysis of the supplement with an emphasis on the questionnaire design refer to the Supplement File for the Current Population Survey which can be found either at the CPS or NTIA Website. For access to the CPS Website, visit www.census.gov/cps and for access to the NTIA Website, visit www.ntia.doc.gov.

4.6 Missing Data

The prevalence of missing item responses in the supplement posed a problem for our study. Variables that had to do with income, employment, school and Internet access had significant amounts of missing data. These variables are critical because of the differences among those who do and do not have Internet access for categories such as income level, employment status and level of schooling demonstrated in Section 1.2.

The missing item responses arose from the difference in target populations of the monthly CPS and the NTIA supplement. The CPS concentrates on labor statistics with a target population of people fifteen years of age and older, while the NTIA interviewed anyone in a household three years of age or older. The major source of missing data was children. In addition, everyone with a missing value for level of school completed also had a missing value for employment status.

Table 4.1: Percentage of Missing Values Using Person Weights

	Everyone	17 & under	Adults
Employment Status	21.19%	82.19%	0.18%
School Completed	21.19%	82.19%	0.18%
Internet Access	4.05%	15.64%	0.06%
Income	18.25%	15.08%	19.35%

As seen above, 82.19% of those seventeen years of age and under had missing values for employment status and level of school completed. After removing them for the sample the percentage of people with missing data became substantially lower; income was the only variable where this did not help. We felt the removal of those seventeen years of age and under would not harm the outcome of the study since many surveys, regardless of the survey mode, target the U.S. adult population. An example of such a survey is the USA Today Gallup Poll, which is a phone survey designed to represent the general population with a sample of 1,000 national adults. However, since many technology initiatives are aimed at children, further research should encompass this group.

Chapter 5

Propensity Scores

Now that we have examined the source and structure of the data, we will look at a tool to reduce coverage bias. Through logistic regression, we are able to re-weight units allowing them to represent missing or under-represented populations. In this chapter, we discuss binary logistic regression and how this tool can be used to account for coverage bias. We also give a review of the existing research in this area.

5.1 Logistic Regression

Logistic regression is a tool used to model the likelihood of an event and to assess the effects of independent variables on this likelihood. This is achieved through a transformation of the dependent varaible, y, which is most often binary. Logistic regression can be described by the equation:

$$logit(p) = \left(\ln \frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{5.1}$$

where p is the probability of the event conditional on the x-values. That is, $p = P(y = 1|x)$.

The probability can thus be computed as

$$p = \frac{1}{1 + e^{-(\alpha + B_1 x_1 + \ldots + B_k x_k)}} \tag{5.2}$$

In order to obtain the β coefficients, logistic regression uses maximum likelihood estimation to maximize the likelihood that the observed values of the dependent variable are predicted from the observed values of the independent variables. For a more general overview of logistic regression see Agresti's *An Introduction to Categorical Data Analysis*[1].

Logistic regression relies on certain basic assumptions.

- The dependent variable must be coded in a way that is meaningful.
- All relevant independent variables must be included in the model.
- All irrelevant independent variables must be excluded from the model.
- The error tems $e_i = y_i - p_i$ must be independent.
- No significant measurement error may be present in the independent variables.
- The log odds of the dependent variable must be linearly related to the independent variables.
- No multicollinearity exists between the independent variable variables (Remedial measures are available to reduce the problem of multicollinearity, c.f. [1]).
- Outliers do not exist or have been removed.
- Data is composed of a sufficiently large sample.

For a simple example of logistic regression, consider Table 5.1 which compares number of pairs of shoes owned and gender (male = 1; female = 0).

A binary logistic regression with gender as the dependent variable yields $logit(p) = ln(\frac{p}{1-p}) = -.212(x) + 3.392$, where p is the likelihood of being male and x is the number of shoes owned by the individual.

Logistic regression models are assessed primarily by two statistics, Hosmer and Lemeshow Chi-square and Negelkerke's R-Square. The Hosmer Lemeshow statistic is given by

$$G^2_{HL} = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - \frac{E_j}{n_j})}, \tag{5.3}$$

Shoes	Gender	Shoes	Gender
3	1	10	0
4	1	10	1
4	1	10	1
4	1	12	0
5	1	17	0
5	1	21	0
6	1	25	0
6	1	25	0
7	1	26	0
7	1	30	0
7	1	30	1
8	1	40	0

Table 5.1: Logistic Regression Example

where n_j is the number of observation in the j^{th} group, $O_j = \sum_i y_{ij}$ is the observed number of cases in the j^{th} group, and $E_j = \sum_i \hat{p}_{ij}$ is the expected number of cases in the j^{th} group. Nagelkerke's R-Square is given by

$$R^2 = \frac{1 - \left(\frac{-2LL_{null}}{-2LL_k}\right)^{2/n}}{1 - (-2LL_{null})^{2/n}}, \tag{5.4}$$

where $-2LL_{null}$ represents the log likelihood of a logistic model with solely the constant as a predictor and $-2LL_k$ represents the log likelihood of a logistic model with the addition of k predictors.

The Hosmer and Lemeshow Chi-squared statistic tests the model to determine if the model does not fit the data. Therefore, lack of significance indicates that the model fits the data. Nagelkerke's R-Square ranges from zero to one. The closer it is to one the better the model. Nagelkerke's R-Square will usually be lower then an equally significant R-Square value in linear regression. With very large samples however, both statistics will tend to indicate that the model does not fit the data. Therefore, when working with large sets of data, as is the case in this study, it is prudent to compare these statistics to those of other models rather than to absolute values. The

ROC curve is another means for assessing logistic regression models which we describe in the next subsection.

5.1.1 ROC Curve

The area under the Receiver Operating Characteristic (ROC) curve is a measure of how well a logistic regression model predicts responses. The height of the ROC curve is the ratio of true positives to false positives as predicted using the variable of interest. In our application of logistic regression to come, true positives occur when the model predicts a unit as not having the Internet when the unit does not have the Internet. False positives occur when the model predicts someone as not having the Internet, when they actually do have the Internet. An ROC curve is shown in Figure 5.1.

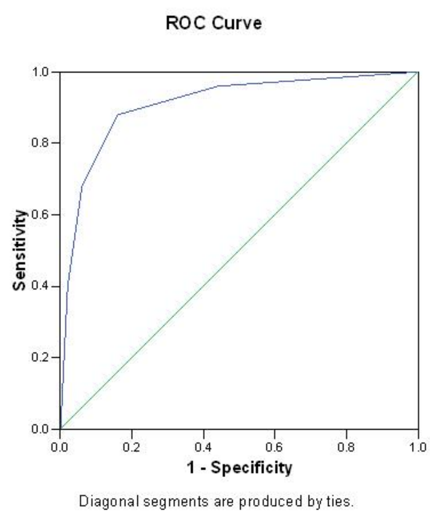


Figure 5.1: ROC Curve

A high area under the curve is desirable as this shows a high number of true positives and a low amount of false positives. The area under the curve will be 0.5 when the model has no predictive ability.

5.2 Propensity Scores in Surveys

A propensity score is the probability that a unit in the sample takes on a specified value of the dependent variable conditioned on a collection of covariate characteristics. The equation for this is given by

$$P(Y = y|X = x) \tag{5.5}$$

where Y is the dependent variable, y is a specified value of Y, X is the vector of covariates, and x is their characteristic values.

Propensity scores are used in telephone surveys to reduce selection bias, which is the tendency for units in the population to be over- or under-selected for the sample. Selection bias can occur due to nonresponse or non-coverage. Propensity scores reduce this bias via a weighting scheme or as an aid in post-stratification.

A weighting scheme utilizing propensity scores is implemented by giving each unit in the sample a weight adjustment equal to the inverse of the propensity score. This formula is shown below.

$$\frac{1}{1 - P(Transient)} \tag{5.6}$$

This method was used by Duncan and Stasny [11] in their study on telephone surveys. These authors observed that this weighting scheme did not provide significant reduction in bias when used alone. When combined with a raking scheme, however, this was found to provide a significant reduction in bias. Bethlehem and Cobben [6] conducted research to investigate the bias due to non coverage in phone surveys, and found similar results regarding the inability of the propensity score alone to reduce bias.

Bethlehem and Cobben also studied the use of propensity scores as a poststratification variable. This technique was introduced by Rosenbaum and Rubin [23]. Generally, five strata are formed with similar propensity scores, and these are grouped together without regard to the value of their dependent variable. Five strata are used because this number has been found to reduce as much as 90% of selection bias [7]. This results in every stratum containing units that have both values of the dependent variable, each with very similar propensity scores. The number of population units in each stratum is determined and used to obtain poststratification weights. This requires knowing population cell counts for many combinations of covariates.

Recent studies regarding Internet surveys have found that while some of the response results align with those from identical questions on telephone surveys, responses differ greatly from telesurveys for other questions. In his review, Couper [8] discusses studies which have attempted to reduce this difference by using propensity scores to weight the responses of the Internet surveys to better correspond with those of a telephone survey. While this did succeed in decreasing the inconsistency of some covariate responses, other covariate discrepancies remained unresolved. Couper goes on to cite Fleming and Sonner, who found no discernable patterns in which covariate inconsistencies occurred, nor any patterns in those covariates which were resolved or unresolved, between Internet and telephone survey. They go on to state that "even after weighting, there were a number of substantial differences between the online poll results and the telephone results." In their work, Bandilla, Bosnjak, and Altdorfer [2] found that 67% of the responses "differed significantly" from an Internet survey to those of a mail survey, even after non propensity score based poststratification. They also found that this difference could be decreased when certain categories of covariates were analyzed. For example, when the surveyors grouped those respondents with a relatively high education level, the answers of the Internet survey closely resembled those of the telephone survey. It should be noted that while previous research has incorporated both a telesurvey and an Internet survey to calculate propensity scores, some of our proposed methods will only be applicable to Internet-only surveys.

Internet surveys are currently a highly active area of research, and our review of the literature indicates that a portion of the data from Internet surveys still conflicts with data collected via phone or personal surveys. Due to the success of propensity scores in telesurveys, and the partial success they have achieved in previous studies on Internet surveys, we study whether the weighting of Internet survey data with propensity scores will successfully reduce coverage bias.

Chapter 6

Weighting Schemes for Coverage Bias

Weighting schemes incorporate two previously developed concepts: propensity scores and raking procedures. A weighting scheme is defined by the variables used to derive a propensity score and the raking procedure implemented. This chapter presents twelve weighting schemes, which use different combinations of raking procedures and propensity models. We also give a glimpse of how the schemes perform using point estimates for a few CPS variables. The point estimates are percentages of the weighted sample that fall into a specific category of a variable of interest. The following ten variables are used in the point estimates: own/rent living quarters, telephone in household, computer in household, cable television in household, own a cell phone, military status, marital status, live in a metropolitan area, hourly worker, and if Spanish is the primary language in household. We have chosen a broad range of variables, from technological to demographic, to see how our weighting schemes perform overall. A full table of the variables used for point estimates and the results from each scheme is given in Appendix C.

6.1 Base Scheme and Target

The target population of this study is the United States population age 18 and older. The focus of the present study is on the adult population for two reasons: first, many Internet studies are designed for the adult population, and secondly, most missing data issues were solved by eliminating people under the age of 18 from our analysis (as discussed in 4.6).

The target values will be based on data from 103,891 of the original 140,037 respondents to the CPS. The sum of person weights for the this group, which represents the number of people in the United States age 18 and older, is 213,426,278. It is the goal of all subsequent schemes to predict the characteristics listed in Appendix C of the target population using just the respondents of the CPS who have access to the Internet, whether it be at home or outside the home. There are 62,326 respondents to the CPS, weighted to 126,936,726, who have Internet access and are age 18 or older.

The base is the simplest scheme used to predict the target values. This scheme does not employ any remedial measure to reduce the effects of coverage bias when the no-Internet population is excluded. To properly weight the base population to the number of adults in the United States, we follow the exact raking procedure implemented by the CPS and perform a 3-way rake by state, age/sex/race, and age/sex/Hispanic. This procedure makes up for the individuals lost when excluding people who do not have Internet access from the base. Table 6.1 gives target values for four variables and point estimates of these values using the base weighting scheme.

	Base	Target		Base	Target
Cable Television			Owns Cell Phone		
Yes	58.2	54.6	Yes	65.2	54.2
No	41.8	45.4	No	34.8	45.8
Home Occupancy			Marital Status		
Own	77.1	72.3	Married	61.3	56.2
Rent	21.9	26.5	Not Married	38.9	42.4

Table 6.1: Target and Base

As expected, predictions using the base scheme were off substantially from their targets. However, the base scheme most likely accounted for some coverage bias, since age and race variables were included in our raking. As noted in Section 2.2, people who do not have Internet access are more likely to be non-white and elderly.

6.2 Raking by Internet Access Status

The goal of this study is to find weighting schemes that will bring our estimates closer to the target than the base scheme. A central variable in creating these weighting schemes is coded access, a variable that divides Internet access into three categories: Internet service at home, Internet service outside the home only (transients), and no Internet access at all. The first method we explored is to weight those with Internet access outside the home only (coded access 1) so that they represent themselves and all of those with no Internet access at all (coded access 2). This is accomplished by adding coded access as a raking variable along with the original three variables used for raking by the CPS to create a 4-way raking procedure.

Thus, the end result of a 4-way raking procedure in regards to people who have Internet access outside the home is given by

$$\sum_{i\,:\,CA_i\,=\,1} W_{ri} = N_n + N_o,$$

where CA_i is coded access for unit i, and W_{ri} is the raked weight for unit i. N_n and N_o are the sums of original person weights of respondents who have no Internet access and Internet access outside the home only, respectively.

	Base	4-way rake	Target
Cable TV (Yes)	58.2	54.2	54.6
Housing (Own)	77.1	73.6	72.3
Cell Phone (Yes)	65.2	59.7	54.2
Marital Status (Married)	61.3	55.9	56.2

Table 6.2: Base and 4-way Raking Procedures

Some point estimates of the 4-way raking scheme are found in Table 6.2. One can conclude with ease that the 4-way raking scheme tremendously reduces the bias of the point estimates when compared to the base scheme. The fact that the 4-way raking scheme gives significantly better results than the 3-way raking scheme supports our idea that weighting transients higher helps account for the bias of excluding people who do not have Internet access. Because of this result, most of our subsequent schemes will implement the 4-way procedure outlined here.

6.3 Propensity Scores

Propensity scores as described in Chapter 5, are conditional probabilities for whether or not you have a selected characteristic of the dependent variable. Here we describe the models we used to calculate propensity scores for our weighting schemes.

6.3.1 Modeling Variables

The variables chosen from the CPS which we believed would be useful for predicting Internet access are income, age, race, employment status, highest school completed, metropolitan, geographic region, citizenship, and computer in the household.

One goal of the research was to create a general and very applicable scheme that could be used in any Internet survey. To accomplish this goal, one of our schemes, the most widely tested, was our primary scheme. It was constructed as a compromise between the more simplistic, and most likely less accurate, schemes constructed and the more complicated schemes which will most likely be more accurate, but less applicable. The primary scheme consisted of the following five variables: income, age, race, employment status, and highest school completed.

These were chosen for their commonality in that these are basic demographic variables included on many surveys and are projected to adequately predict Internet access.

6.3.2 Propensity Scores

After completing the logistic regression in SPSS for each model, propensity scores were computed for the sample of 62326 units. With the data available to us in the CPS, we computed two propensity scores; a propensity for transience and a propensity for no Internet access. If those without Internet access are not included, so that the two groups considered are those with Internet inside the home and those with Internet outside the home, the propensity scores are the probability that you have the Internet outside the home. If those without Internet access are included in the study, so that the two groups considered are those with any Internet access at all, and those with no Internet access whatsoever, the propensity scores are the probability that you have no Internet access whatsoever.

After obtaining the propensity scores, the complimented inverse of this was then multiplied by the original CPS person weights. The equation for the final un-raked weights is given as

$$W_i^* = \frac{1}{1 - p_i} * w_i' \tag{6.1}$$

where p_i is the propensity score of the i^{th} unit, and w_i is the original CPS person weight of the i^{th} unit.

As mentioned in this previous section, this value was then raked both by the CPS defined 3-way rake, and by our own 4-way rake which included raking by coded access. The primary scheme is the only scheme for which we employ the 3-way rake as well as the 4-way rake (besides the non-weighting schemes), since the performance of the 4-way rake is superior.

6.3.3 Missing Values

Those sampled under the age of 18 were initially removed from the sample to reduce the amount of missing values in coded access. While this removed a large portion of cases with missing data, roughly 10% of the remaining cases had missing values for a combination of the following three variables: income, employment status, and highest level of school completed. To compensate for this nonresponse, the specific scheme was applied without the variable in question (income, employment, school), and then the resulting propensity scores from this reduced model was substituted as the propensity scores for the missing cases of the variable in question. This method was implemented for all three variables in every scheme we proposed to handle the missing data in the sample.

Binary Schemes

Our first approach to constructing a model for transience stemmed from previous research in telephone surveys in which independent variables were coded to binary. These consist of models in which the variables are divided into only two categories, such as either above or below $30,000 for income. The first model tested was a binary version of our primary scheme, while the second used only two independent variables. The two-variables model consisted of the two best predictive variables, income and highest level of school completed. This scheme was created to investigate the predictive power of a very basic model. Therefore if a simple model performs almost as well as a more complex model, the simple model we be preferred.

The binary primary model produced an ROC Curve value of 0.593, while the two-variables model yielded a 0.601 ROC value. The point estimates are listed in Table 6.3 below. Point estimates are very similar for the binary primary and two-variable schemes.

	Binary Primary	Binary 2 Var.	Target
Cable TV (Yes)	54.0	54.1	54.6
Housing (Own)	73.3	73.2	72.3
Cell Phone (Yes)	59.4	59.5	54.2
Marital Status (Married)	55.6	55.6	56.2

Table 6.3: Binary Scheme Point Estimates

Multi-Category Models

Propensity scores were then calculated using the primary model and a more specific version, the GMC scheme. The primary model consisted of a multi-categorical coding of each of the primary variables, ranging from three to seven categories. A detailed list of these categories for each variable can be found in Appendix B. The GMC model included the five primary variables, as well as geographic region, metropolitan, and citizenship. These were included to test whether a more complicated model, with higher Chi-Square and R^2 values, would better predict the target estimates than the more general, and more applicable, primary model.

The primary model had a relatively high ROC value of 0.660, while the GMC scheme registered a value of 0.645. The point estimates for these two schemes are shown in Table 6.4. Again, the two models yielded similar point estimates.

6.3.4 Modified Internet Access

One variable the CPS provides that proved to be extremely useful in our study was frequency of Internet use. The options for this item are: at least once a day, at least once a week, but not everyday, at least once a month

	GMC	Primary	Target
Cable TV (Yes)	53.9	53.9	54.6
Housing (Own)	73.1	73.1	72.3
Cell Phone (Yes)	59.3	59.3	54.2
Marital Status (Married)	55.5	55.5	56.2

Table 6.4: Primary and GMC Scheme Point Estimates

but not every week, and less than once a month. The modified coded access scheme uses this information to change some respondents' coded access status. Under this scheme, a respondent who has Internet access in their home, but accesses it less than once a week is considered to have Internet access outside the home. The idea behind this transformation is that those who access the Internet rarely are more like people who do not have Internet access at all rather than people who have Internet access at home. Thus, viewing those who use the Internet rarely as having Internet access outside the home only will give their responses more weight through the 4-way raking procedure presented in Section 6.2. In the end, this should improve the results if these people truly are similar to those who do not have Internet access at home. To see how this model performs, we use the variables from the primary model and look for any improvements. Some point estimates derived from this scheme are presented in Table 6.5.

	Modified Coded Access	Target
Cable TV (Yes)	54.9	54.6
Housing (Own)	74.2	72.3
Cell Phone (Yes)	60.4	54.2
Marital Status (Married)	57.3	56.2

Table 6.5: Modified Coded Access

Looking at point estimates only, this scheme does not seem to be an improvement over the original primary scheme. However, we will withhold judgment until Section 7.3, when we conduct a more thorough analysis of the schemes.

6.3.5 Full Sample

All models presented thus far have been derived using only adult respondents with Internet access. The reason for this is that Internet surveys will only have access to this subsection of the population. Thus, creating models using only this information optimizes their application to future Internet surveys.

Since the study does have information on people without Internet service through the CPS, it is possible to create weights using this information in hopes of generating better schemes. It is true that these schemes may not be as applicable to researchers if they are conducting their survey solely online, but they will be applicable if performing a mixed-mode survey or a survey of a population for which an outside source with data about Internet access exists.

To use the information in the CPS about people who do not have Internet access, we first redefines the coded access variable. The new coded access now distinguishes whether a person has Internet access anywhere, or whether they do not have Internet access at all. In previous schemes, we would first remove people who did not have Internet access and perform a logistic regression to get propensity scores. Now, we will run a logistic regression first on the entire adult sample, using the new coded access as the dependent variable, and then remove the people who do not have Internet access from our analysis. This change in procedure allows for propensity scores using the entire target population, which should increase the accuracy of these models.

The study uses the new coded access in conjunction with two different schemes already discussed: the primary scheme, and the GMC scheme. A third scheme that uses the new coded access is the GMC scheme plus one additional variable, computer in the household. A natural question is why this variable has not been utilized before. In previous schemes, the analysis was done only on those people who had Internet access, and whether they had it inside or outside the home. Including computer in the household as a variable in logistic regression for these previous schemes would have been too strong an indicator of where one accessed the Internet. Thus, if one did not have a computer in their household, but still accessed the Internet, then it can be concluded with certainty that they accessed the Internet outside the home. Therefore, the ensuing propensity score for this type of respondent would be too large for practical use.

Table 6.6 shows a sample of point estimates from the GMC scheme with computer in the household as an added variable. While the point estimates of these models are not impressive, we will again withhold judgment for the more complete analysis in Section 7.3.

	GMC w/CH	Target
Cable TV (Yes)	51.6	54.6
Housing (Own)	70.7	72.3
Cell Phone (Yes)	54.3	54.2
Marital Status (Married)	53.9	56.2

Table 6.6: Full Sample Coded Access

6.4 Frequency of Internet Use Scheme

Duncan and Stasny [11], proposed a weighting scheme based on how often a sampled person has telephone service. Those who have telephone service less are weighted more, and those who have telephone service more often are weighted less. A similar method was applied in our study in regards to a specific variable which stated how often a person used the Internet. This variable, which we used above for modified coded access, was found in the Internet and Computer Use Supplement.

A simple equation was determined to give a higher weight to those who used the Internet less. This was accomplished by taking a value akin to the inverse of the unit's probability of being selected. This probability is determined by taking the average number of days the unit is on the Internet, as given by their response to the Internet use variable, and dividing this by 365, as shown by the following equation:

$$\frac{\text{number of days in the year}}{\text{avg. number of days on the Internet}} \tag{6.2}$$

The point estimates are shown in Table 6.7. This scheme gave point estimates very close to the target.

	FIUS	Target
Cable TV (Yes)	53.9	54.6
Housing (Own)	72.9	72.3
Cell Phone (Yes)	58.0	54.2
Marital Status (Married)	57.5	56.2

Table 6.7: FIUS Scheme Point Estimates

Chapter 7

Results

Our goal in modeling is to produce accurate estimates of natural or experimental phenomena. To do this we need to minimize bias the bias of our estimates while maintaining precision. Bias refers to an estimator's proximity to the true value it seeks to estimate and an estimator is considered to be unbiased when the difference between these two values is zero [19]. High precision occurs when estimates of the same target value based on different sample data are located in close proximity to one another.

Weighting schemes, like those presented in Chapter 6, can have a significant influence on the variance of an estimate. As weighting schemes become increasingly unequal, variability increases. This can be shown with a simple example. Suppose X_1 and X_2 are independent and identically distributed random variables with known variance, σ^2 and w_1 and w_2 are known weights:

$$Var(w_1X_1 + w_2X_2) = w_1^2Var(X_1) + w_2^2Var(X_2) = (w_1^2 + w_2^2)\sigma^2 \tag{7.1}$$

If we constrain the sum of the weights to equal 2, it can be shown that the variance is minimized when $w_1 = w_2 = 1$. Table 7.1 illustrates that as the

w_1	w_2	$Var(w_1X_1 + w_2X_2)$
0.1	1.9	$3.62\sigma^2$
0.25	1.75	$3.125\sigma^2$
0.5	1.5	$2.5\sigma^2$
1	1	σ^2

Table 7.1: Effects of Unequal Weighting Schemes on Variance

weights diverge from one another, the variance increases. This is significant because some the weighting schemes presented in Chapter 6 do have a large range of weights and may have substantially increased variance over the base scheme. In fact, the CPS actually safeguards against this by collapsing cells if data collection leads to a weight adjustment of more than two or less than 0.6 [17]. The challenge in choosing a good scheme is to maintain a balance between precision and bias.

The mean square error quantifies the bias/variance trade off and gives an overall measure of accuracy.

$$MSE = bias^2 + Variance \tag{7.2}$$

In this chapter, we will propose a variance estimation procedure that will be applied to all weighting schemes introduced in Chapter 6. In addition, we present three scheme specific summary statistics that are used in comparative analysis of our modeling schemes.

7.1 Variance Estimation Via Replication

As was noted in Section 4.4, variance estimation using the CPS's generalized variance functions is not reliable in the context of our study because because these functions apply to only to units weighted by the CPS's person and household weights. In our schemes, units are reweighted to account for coverage bias. Therefore, another method is needed. We choose to obtain the variance estimates for variables of interest in each of our schemes by using a random groups method which takes small, representative subsamples of the data. These samples are then compared to the to the whole sample after both have been processed through one of the reweighting schemes. Through the following procedure variance estimates can be obtained for all variables of interest and all models:

1. The data was processed through one of the schemes described in Chapter 6. A post-scheme weight is obtained.

2. Point estimates, $\hat{\theta}$ for each of the ten variables of interest listed at the beginning of Chapter 6 were obtained by weighting the data by the post-scheme weight.

3. The data was partitioned into ten random, mutually exclusive groups with members of the same household assigned to the same group. Partions each contained approxametely 6,300 units.

4. Each group was run through a procedure identical to the procedure used to produce the weights in step 2. A post-scheme weight is obtained for each partition.

5. Point estimates, $\hat{\theta}_r$, for the variables of interest were obtained for each partition using the post-scheme weight for the given partition.

6. Variances are produced for each variable of interest from the following equation:

$$\hat{Var}(\hat{\theta}) = \frac{1}{k}\sum_{r=1}^{k}\frac{(\hat{\theta}_r - \hat{\theta})^2}{k-1}, \tag{7.3}$$

   where k is the number of partitions.

Through this process the mean squared errors (MSE) for each variable of interest can be obtained from the equation:

$$MSE(\hat{\theta}) = \hat{Var}(\hat{\theta}) + (\hat{\theta} - \theta_{target})^2 \tag{7.4}$$

For a broader survey of variance estimation procedures, see Wolter's *Introduction to Variance Estimation*[28].

7.2 Assessing Accuracy

Variance estimation via replication produces a point estimate and accompanying standard error for each of the ten analysis variables. In this section, we introduce summary measures of accuracy which we will use to compare weighting schemes. However, caution should be used when making overall comparisons because the summary statistics are sensitive to the variables of interest. Some of the variables chosen for this study where chosen deliberately because they are thought to be significantly affected by coverage bias. Variables that are not suspected to be affected by coverage bias may show the base scheme to perform better.

The first summary statistic provided is the average standard error of the scheme. This can be calculated by taking the arithmetic mean of the standard errors for the ten analytical variables. The average standard error for scheme j is given by

$$\bar{SE}_j = \frac{1}{10}\sum_{i=1}^{10} SE_{ij}, \tag{7.5}$$

where $SE_{ij} = \sqrt{\hat{Var}_j(\hat{\theta}_i)}$, the standard error of the point estimate for variable i using scheme j. This statistic provides a measure of the overall variability of each modeling scheme.

Average z-scores for each scheme were also calculated to incorporate bias and variance into a single measure. A z-score gives the distance between the central value and an estimate. For example, a z-score of 1.0 indicates that an estimate is exactly 1 standard error away from the true value. Thus, z-scores that are close to zero indicate an estimate that is closer to the true value and thus is a better predictor of the true parameter. As alluded to earlier (refer to Chapter 1), the strength of this analysis lies in the fact that the CPS data set has information about the non-internet population and as a result, the target population is representative of the entire U.S. population, not just those that have access to the internet. As a result, each of the scheme estimates can be compared to the corresponding target percentages (see Chapter 6), which we use as the "true value" in our computation of the z-scores. In order to provide an average z-score for scheme j, a z-value was calculated for each of the ten variables and then the arithmetic mean of the absolute values of these z-scores was obtained.

$$\bar{z}_j = \frac{1}{10}\sum_{i=1}^{10} \left| \frac{\hat{\theta}_i - \theta_{(target)i}}{SE_{ij}} \right| \tag{7.6}$$

The last piece of analysis provided in this study involves a comparison of prediction intervals for each scheme. Since the study deals with a sample, albeit a large sample, confidence intervals are a more reliable statistical measurement than point estimates because they account for sampling variability [18]. We compared 95% confidence intervals for each analytical variable in each weighting scheme. Then, we tallied the resulting number of target values, out of a total of ten variables, that fell within each schemes' confidence interval. We expect this measure to be more resistant to outliers than the average z-score. Indeed, we are really counting how many z-scores out of the ten have absolute value smaller than 1.96. We advise using this summary statistic only in conjunction with other evidence to draw conclusions.

7.3 Comparisons of Schemes

The following analysis will consider these summary statistics with our comparisons following the development of the schemes presented in Chapter 6.

7.3.1 Effect of Raking by Internet Access Status

Through a process described in Section 4.3.1 and implemented on our data in Section 6.2, schemes were created by adding Internet access status as a raking variable with the intention of reducing the sample's coverage bias. The two schemes were not raked by Internet access will provided a basis for comparison.

The most natural comparison to make is to compare the base scheme, which is raked only by the CPS raking categories, with the no propensity 4-way rake scheme, which is raked additionally by Internet access status. In this case, the average z-score fell substantially from 22 to 2.8 with the addition of Internet access status as a raking variable indicating a much more accurate scheme. It should also be noted that the standard error conversely rose substantially from 0.003 to 0.015.

The next direct comparison we can make is between the 3-way raked primary scheme and the primary scheme which is raked 4-ways. Both of these schemes are weighted with propensity scores. For more information on propensity scores and their effects, see Sections 6.3 and 7.3.2. Results of this comparison are similar to the comparison of the schemes without propensity scores but less dramatic. The average z-score decreased from 22 to 6.6 while standard error increased from .006 to .007. The primary 4-way raked scheme also correctly produced a confidence interval that captured three more target variables.

As we discussed in Section 7.3.2, the addition of propensity weights produces small and variable results. The addition of Internet access status as a raking variable, in contrast, produces substantial positive results in terms of accuracy. Therefore, even though many of our schemes do not have a 3-way raked counterpart for direct comparison, we can conclude that most of their increase in accuracy is do to raking to Internet access status.

Another interesting result in this comparison is that the standard error increased significantly less with the presence of a propensity weight, rising only .001 between the 3-way raked primary and the primary schemes verses .012 between the base and no propensity 4-way rake scheme.

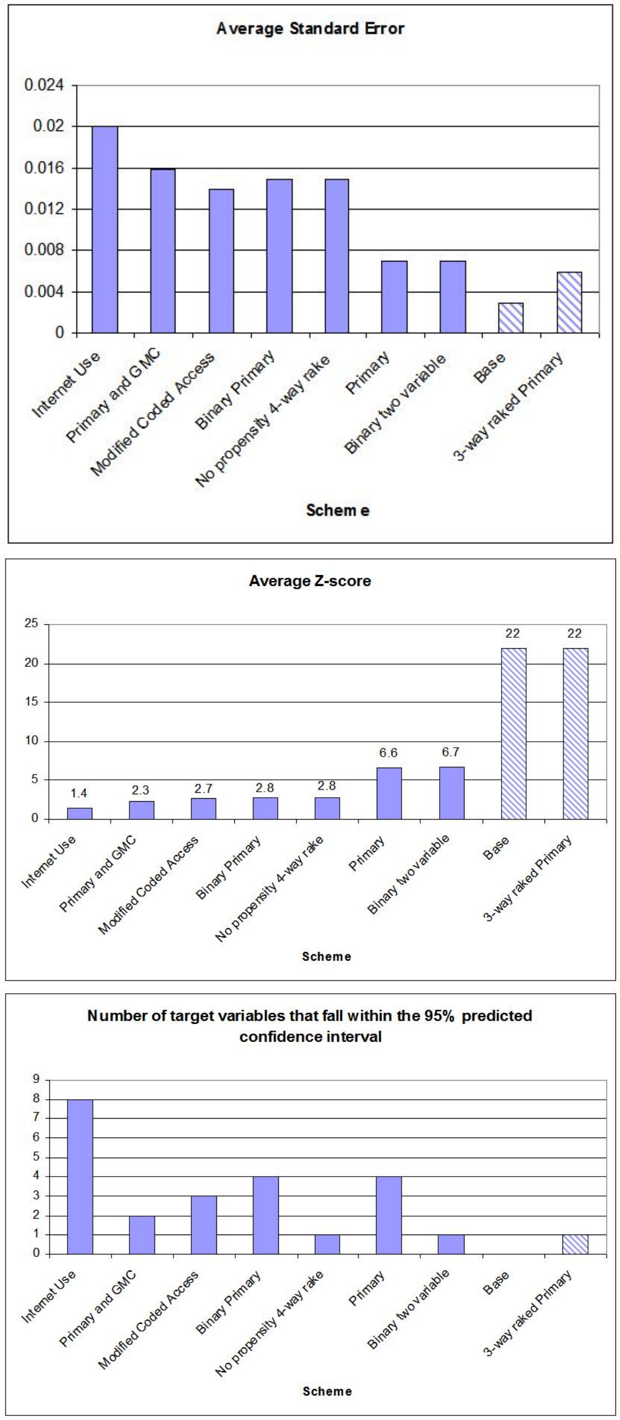


Figure 7.1: Comparison of 3-way Rake to 4-way Rake

7.3.2 The Effect of Propensity Scores

As discussed in Section 5.1, through logistic regression propensity scores can be calculated for all units in a given dataset. The models used to produce these scores are discussed in Section 6.3.

We begin our evaluation of schemes by comparing those with propensity weighting to those without it.

Of the schemes presented in Section 7.3.2, two groups can be directly compared as the only difference is the presences or absence of a propensity weight. The first two schemes that can be compared are the base scheme, which has no propensity weight, and the 3-way raked primary scheme, which is modeled with the five primary variables. It is clear in this case that the addition of a propensity weighting adjustment caused little improvement in the accuracy of the model while doubling the standard error.

Results of the remaining schemes, of which only the no propensity 4-way rake scheme is not adjusted, are mixed but generally more favorable to the concept of adjusting with propensity scores. While the propensity score decreased the z-scores of the primary and binary two variable schemes, the propensity score increased the accuracy of the binary and GMC scheme with only a slight increase in standard error. Overall, models with propensity scores did better forming confidence intervals which contained the target point estimate.

Thus, one important note to draw from this analysis is that propensity weighting is most effective when the scores are calculated with the largest number of relevant covariates. Experimentation may be needed to determine which set of covariates produces a model that best reduces bias for the variables of interest.

Effect of Modifying Coded Access

One of the alluring aspects of research with categorical variables is the room for manipulation within variables. Some categories are relatively fixed (ie. gender), whereas others like age can be coded a multitude of ways. In this portion of the study, a brief analysis is conducted to address the tradeoff between complexity of a model and additional information gained. The majority of the time in mathematical modeling, the more variables present and/or the more specific variables are, the better predictive power a model will have [18]. However, analysts must decide when a model becomes too complex and

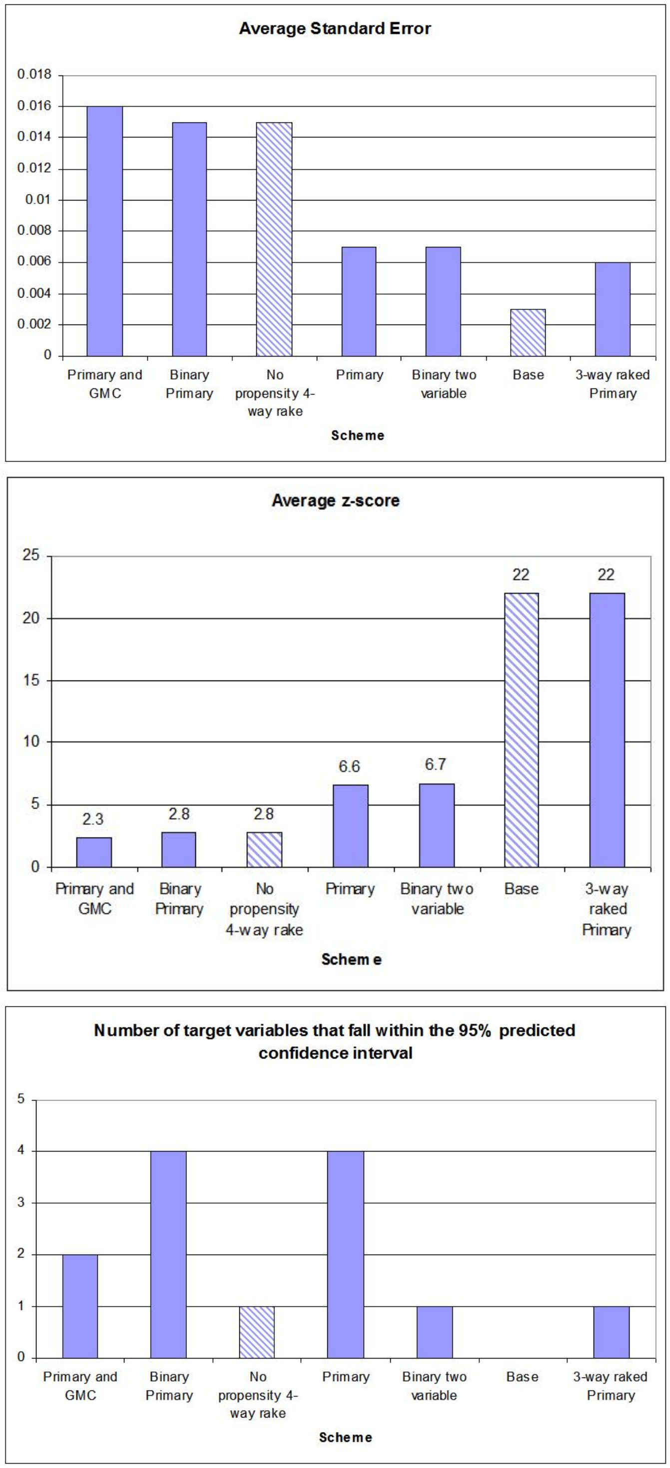


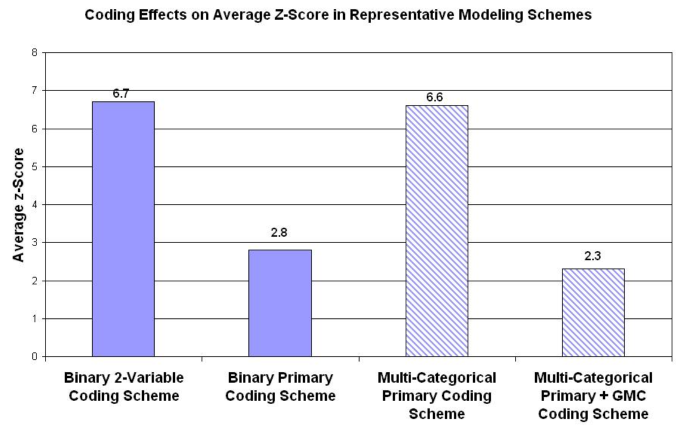
Figure 7.2: Comparison of Propensity Schemes to Raking Only Schemes

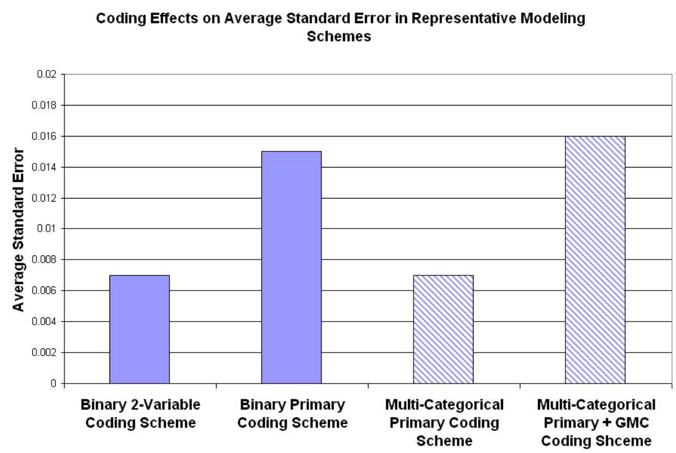
becomes less applicable.

First, consider binary and multi-categorical schemes separately. In the binary case, increasing the number of variables from two to five decreases the average z-score by more than three from 6.7 to 2.8 while maintaining a comparable standard error. In addition, the number of target variables included in the 95% prediction intervals increase by 3. As a result, it is clear that increasing the number of variables in the binary case drastically improves the model without making the model overly complex.

The same may not be said of the multi-categorical schemes. While similar improvements with regard to average z-score and average standard error estimates were observed, increasing the variables from the five primary variables to encompass geographic region, metropolitan status and citizenship did not result in an increase in the number of target variables within the 95% confidence interval. While the average z-score dropped from 6.6 to 2.3, individual prediction of the analysis variables did not change at all. This may be due to the kind of variables being predicted by the modeling schemes. Since the variables are sensitive to coverage bias, and the average standard error is so low, the confidence interval may be artificially small. In any case, more research is warranted to decide whether the more complex model is necessary in the multi-categorical schemes.

In addition to comparing the binary and multi-categorical cases separately, an analysis comparing binary to multi-categorical schemes is justified. In this case, it is natural to compare the binary and multi-categorical variables that have the same primary modeling scheme. In this case, the binary scheme has a much smaller average z-score; 2.8 versus 6.6 (see Figure 7.3. In fact, the primary binary scheme has a comparable average z-score to the multi-categorical that contains three additional variables. Also, the two primary models have comparable average standard errors. The interesting aspect of this comparison is that that the binary scheme actually does a significantly better job including target values in its confidence intervals than any of the models that claim a multi-categorical scheme. As a result, both the binary primary scheme and the multi-categorical primary scheme with three additional variables do an adequate job predicting variables that may be influenced by coverage bias. In this case, the minimal gains may not be worth the added complexity.





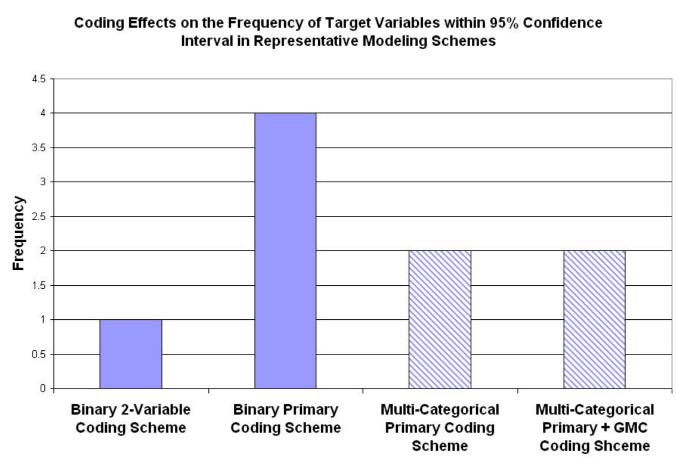


Figure 7.3: Comparison of Propensity Schemes

Effect of Including the Non-Internet Population

In creating our study, we wished to broaden the methods presented beyond strictly Internet surveys by using the additional data from respondents without Internet access. We created three schemes, called "new coded access" schemes, with the idea that the method could be applied to mixed mode surveys and supplement strictly Internet surveys. To read more about these schemes, see section 6.3.5. Four schemes, those that produced the best results of the alternative methods, are used for comparison.

These schemes can be compared in two ways. First, the new coded access schemes are compared to the schemes using only the Internet population, and second, new coded access schemes are compared amongst themselves. It should be noted that new coded access schemes and Internet population only schemes could only be compared in the context of information being available on the non-Internet population.

The most direct comparison that can be made between new coded access schemes and Internet population only schemes is that of the new coded access and GMC and primary and GMC schemes. These are the same model run on different data sets one with the non-Internet population and one without it. As can be seen in the average z-score chart (Table 7.4), the new coded access GMC model did significantly better then the Primary and GMC model. Again, however, we see the bias to variance trade off as the standard error increases by 0.009. Although other models are not directly comparable, we can observe that all new coded access schemes had lower average z-scores and produced confidence intervals that contained more target variables on average.

When comparing between new coded access schemes, we see that the new coded access and GMC scheme is the most accurate both in terms of z-score and number of target variables in the scheme's 95% confidence intervals with this increased accuracy comes increased standard error however.

In general, the more data that is available the more accurately that data can be assessed. We can see this when comparing new coded access schemes to Internet only population schemes. With the addition of the non-Internet population to the computation of propensity, predictions of the target still using only those with Internet acess become more accurate. This comparison is also a good illustration of the bias to variance trade off. With the exception of the Internet use model, new coded access schemes where less biased but more variable.

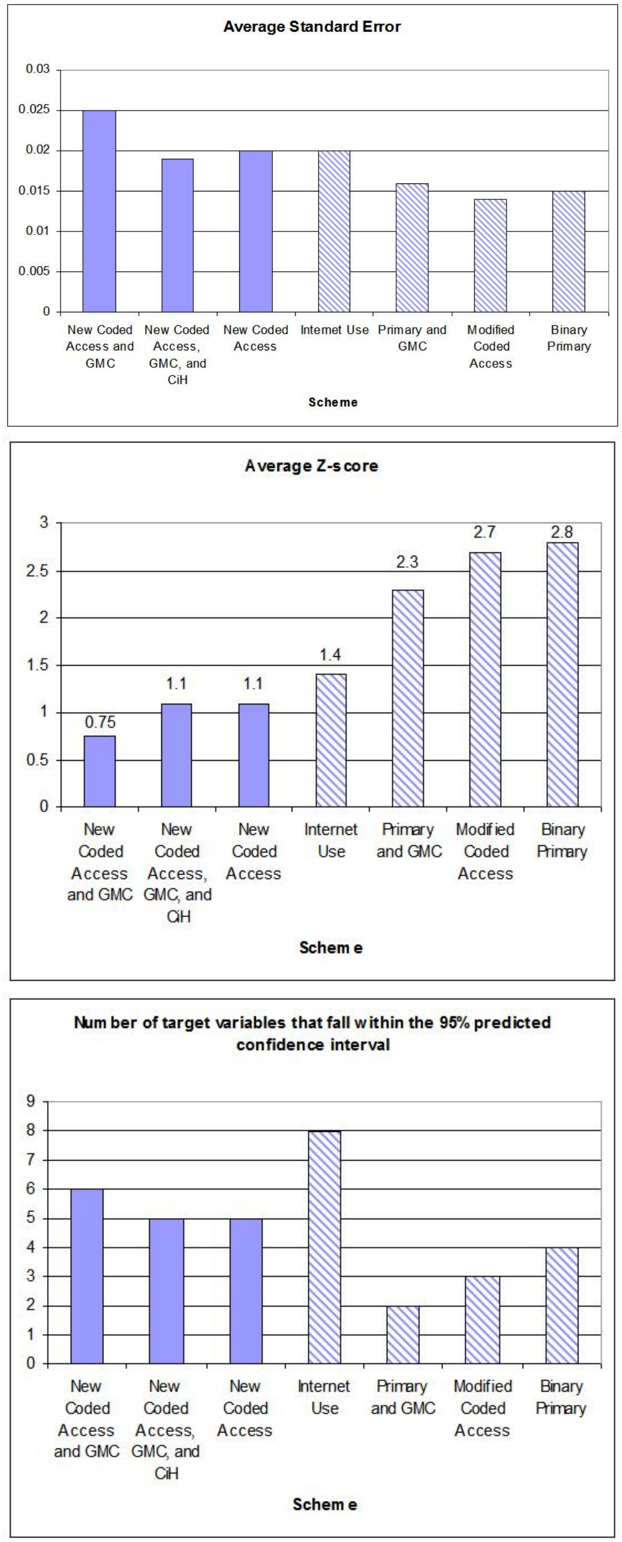


Figure 7.4: Comparison of New Coded Access Schemes

7.3.3 Frequency of Use Scheme

Lastly, this study investigates whether or not having access to information regarding frequency of Internet use is of any merit. In order to assess the two proposed frequency models, modified coded access and Internet use, comparisons are made between the frequency models as well as against two other previously established powerful schemes, the binary primary scheme and the multi-categorical scheme including geographic region, metropolitan status, and citizenship.

From Figure 7.5, we see that the Internet use scheme, where propensity is based solely on usage not availability of the Internet has the lowest z-score of all original coded access schemes. In contrast, the modified coded access scheme is comparable to the previously mentioned schemes. In addition to maintaining a relatively low average standard error, within 0.004 of the other schemes, the Internet use model contains nine out of a possible ten target variables in its 95% confident intervals.

As explained earlier, the modified coded access scheme gives individuals transient status if they rarely use the Internet. This model does not outperform any of the previous models and thus it may be concluded that this transient characterization does not provide additional predictive information. This may be due to the fact that individuals are being redistributed based solely on one characteristic that may or may not warrant a shift to transient status.

On the contrary, the Internet use scheme provides promising evidence that frequency of Internet usage is an advantageous variable with regards to weighting to account for non-Internet households. With only a small increase in average standard error, the average z-score and frequency of target variables within the 95% confident intervals give excellent results. More research should focus on this variable and its potential for minimizing coverage bias in Internet-only surveys.

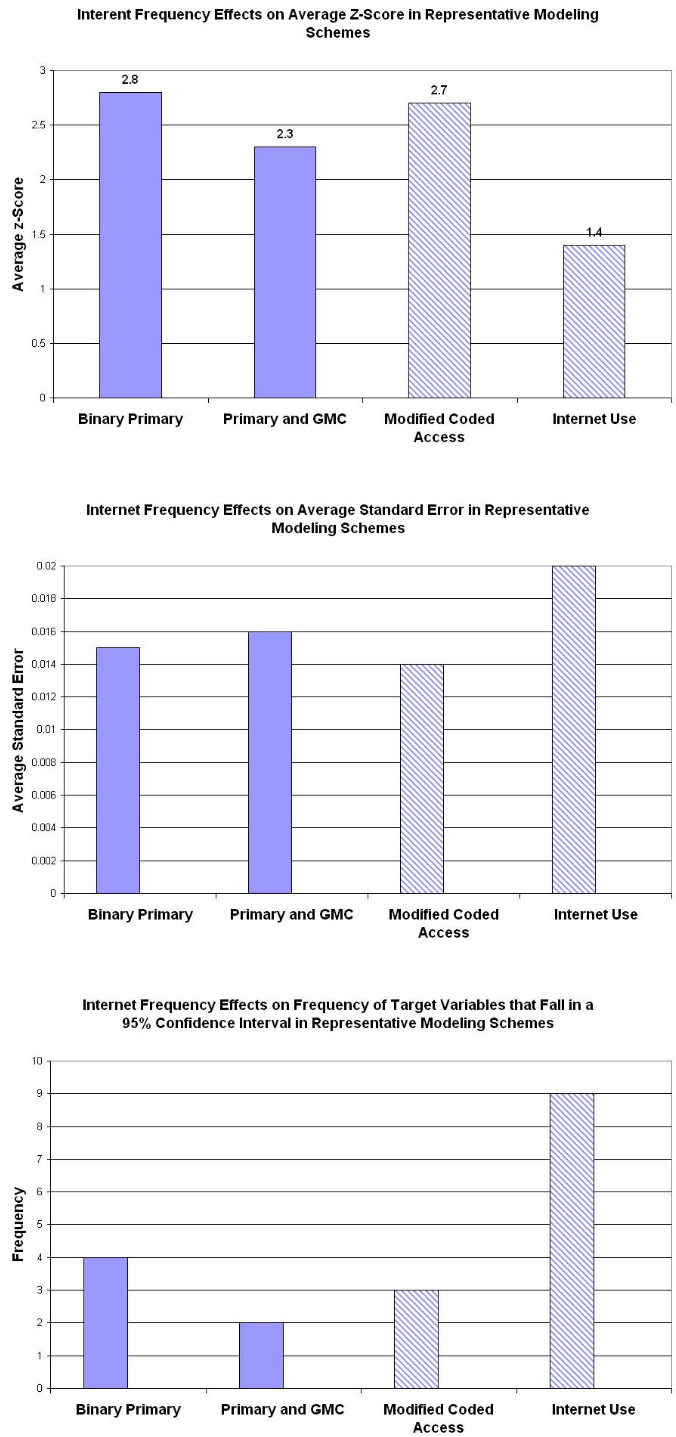


Figure 7.5: Comparison of Schemes Incorporating Frequency of Internet Use

Chapter 8

Conclusions

There is substantial evidence that computer use as well as Internet use will become more commonplace as the 21^{st} century progresses. However, in the realm of Internet surveys, coverage bias will continue to be a concern because the sampling frame will never encompass the entire general population. It is important to explore alternative methods to manage coverage bias because the Internet is such an efficient medium for collecting data.

This study provides an in-depth analysis of a number of methods for reducing coverage bias. Before considering methods to reduce coverage bias, there should be evidence that coverage bias is present in the dataset at hand. We observed this in the CPS by comparing the base predictions to the target parameters. The base predictions, which included only information from the Internet population and no coverage bias adjustments, were off by an average of 6.2% in the ten variables we examined.

The overall theme throughout this research was to weight individuals that only have access to the Internet outside their homes more heavily as to allow them to represent the population that do not have Internet access at all. The first clear delineation that can be made involves poststratification via raking. Two methods were proposed in this study; the basic 3-way CPS rake and a modified 4-way rake involving Internet access. Any modeling scheme that was executed with both poststratification methods provided significantly better results, even when increased variance was considered, when a 4-way rake was used. Thus, we recommend including Internet access as a raking variable when possible.

Another dichotomous relationship that was investigated in this study revolved around the inclusion of non-Internet population data to construct propensity scores. Table D.1 illustrates the modeling schemes that included the new coded access criteria resulted in the smallest average z-scores, without drastically altering the variance. As a result, we recommend that if external sources provide information about Internet access in the population of interest, this extra information should be including in the modeling schemes.

Regardless of the modeling framework, the degree of complexity of a model is always an issue. It is important to consider how complex a model is versus how much additional information the model gleans. The models in this study vary in complexity. Our results did not lead us to any decisive conclusion as to the amount of complexity that is best for controlling coverage bias. While the simple binary primary scheme did an adequate job predicting, the multi-categorical scheme including only three additional variables, geographic region, metropolitan status, and citizenship, lowered the average z-score of the model by 0.5 standard deviations. Some may see this as a worthwhile improvement while others may not. These additional variables also are not typically found in general Internet surveys; including them would increase response burden.

The last set of schemes analyzed in this study revolved around having data regarding frequency of Internet use. The modeling scheme labeled Internet use could be considered the best overall model not including any information about the non-Internet population. It resulted in the lowest average z-score of the original coded access schemes and captured nine out of the ten possible predictor variables in the scheme's 95% confidence interval. These results should spur continued research along these lines.

Our results should be taken into account by survey methodologists as they design questionnaires. By including some extremely simple demographic variables in an Internet survey, weighting adjustments can and should be made to help minimize inherent coverage bias. In addition, more research should be conducted to examine whether frequency of Internet usage should be a staple question in all Internet surveys. Preliminary evidence from this study indicates this variable can be very successful with respect to minimizing coverage bias.

With such a large dataset as the CPS, future studies warrant subsampling the raw data in an attempt to mimic the results that may be seen in a typical Internet survey. It would be interesting to subsample the CPS data based on probabilities created from the frequency of Internet use. For example, a subsample of approximately 40,000 cases, cut down from 63,000 cases, could be obtained by giving individuals who are on the Internet more frequently, a higher probability of responding. This would result in a sample that would be representative of what a survey methodologist might find if he or she conducted an Internet survey. Worthwhile research would result in running the aforementioned models on this sample and testing them against the benchmark target population. We expect that this would further exacerbate the imbalance in weights and require rethinking the bias/variance trade-off. This analysis would give further insight into how effective our proposed models are in minimizing coverage bias.

Appendices

Appendix A

Flow Chart of Weighting Schemes

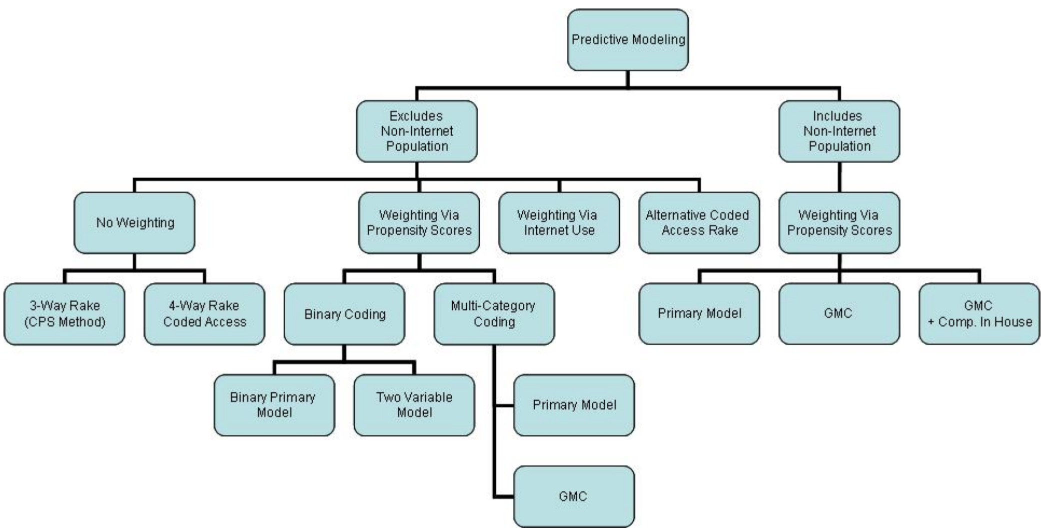


Figure A.1: Flow Chart

Appendix B

Variables Used in Modeling

Binary Coding	*Coded Value*	*Actual Value*
Income		
	0	\$0-\$30,000
	1	\$30,000+
School		
	0	Did not graduate High School
	1	Graduated High School
Employment Status		
	0	Employed
	1	Unemployed
Age		
	0	Under 65
	1	Over 65
Race		
	0	White
	1	Other

Multi-Category Coding	*Coded Value*	*Actual Value*
Income		
	1	\$0 - \$14,999
	2	\$15,000 - \$29,999
	3	\$30,000-\$49,999
	4	\$50,000-\$74,999
	5	\$75,000-\$99,000
	6	\$100,000-\$150,000
	7	\$150,000+
School		
	1	Less than 1st grade through 12th grade no diploma
	2	High School Grad-Diploma Or Equiv (ged)
	3	Some College But No Degree
	4	Associate Degree
	5	Bachelor's Degree
	6	Masters or Above
Employment Status		
	1	Employed
	2	Uneemployed
	3	Retired
	4	Disabled
	5	Other
Age		
	1	18-25
	2	26-39
	3	40-64
	4	Over 65
Race		
	1	White
	2	Black
	3	Asian
	4	Other

Multi-Category Cont.	*Coded Value*	*Actual Value*
Computer In Household		
	1	Yes
	2	No
Internet Use		
	1	Every day
	2	Every week, but not every day
	3	Every month, but not every week
	4	Less than once a month
Metropolitan		
	1	Yes
	2	No
Geographic Region		
	1	Northeast
	2	Midwest
	3	South
	4	West
Citizenship		
	1	Native (Born in the US)
	2	Native (Born in PR or outlying US area)
	3	Native, Born Abroad Of US Parent(s)
	4	Foreign Born, US Cit By Naturalization
	5	Foreign Born, Not a US Citizen

Appendix C

Point Estimates by Scheme

	Base +3-way Rake	Base +4-way Rake	Binary Primary	Binary Two Vari.
Cable TV (Yes)	58.2	54.2	54.0	54.1
Housing (Own)	77.1	73.6	73.3	73.2
Cell Phone (Yes)	65.2	59.7	59.4	59.5
Marital Status (Married)	61.3	55.9	55.6	55.6
Telephone (Yes)	97.6	96.7	96.6	96.6
Computer (Yes)	89.4	71.2	71.1	71.2
Military Service (Yes)	11.9	11.6	11.7	11.7
Spanish Speaking (Yes)	98.5	98.3	98.3	98.3
Hourly Worker (Yes)	52.9	54.3	54.9	54.9
Metropolitan (Yes)	84.1	83.6	83.3	83.3

Table C.1: Point Estimates by Scheme

	Primary 3-Way Rake	Primary 4-Way Rake	GMC	FUIS	Target
Cable TV (Yes)	57.9	53.9	53.9	53.9	54.6
Housing (Own)	76.6	73.1	73.1	72.9	72.3
Cell Phone (Yes)	64.7	59.3	59.3	58.0	54.2
Marital Status (Married)	59.9	55.5	55.5	57.5	56.2
Telephone (Yes)	97.6	96.6	96.5	96.7	95.7
Computer (Yes)	89.1	71.0	71.0	71.5	66.1
Military Service (Yes)	11.9	11.6	11.6	11.9	10.9
Spanish Speaking (Yes)	98.5	98.3	98.3	98.1	96.3
Hourly Worker (Yes)	54.1	55.1	55.1	65.3	59.2
Metropolitan (Yes)	83.9	83.3	N/A	83.6	81.5

Table C.2: Point Estimates by Scheme

Appendix D

Tables of Accuracy Assessment

Scheme	Standard Error	Z-scores
New Coded Access and GMC	0.025	0.75
New Coded Access, GMC, and CiH	0.019	1.1
New Coded Access	0.020	1.1
Internet Use	0.020	1.4
Primary and GMC	0.016	2.3
Modified Coded Access	0.014	2.7
Binary Primary	0.015	2.8
No propensity 4-way rake	0.015	2.8
Primary	0.007	6.6
Binary two variable	0.007	6.7
Base	0.003	22
3-way raked Primary	0.006	22

Table D.1: Average Standard Errors and Z-scores

Internet Use scheme	Point Estimate	SE	Bias
Rent Own	0.739	0.0137	0.0387
Telephone	0.967	0.0046	0.0130
Computer	0.716	0.0607	0.1719
Military	0.120	0.0094	0.0268
Metro	0.812	0.0099	0.0280
Cable TV	0.539	0.0147	0.0416
Cell Phone	0.580	0.0323	0.0914
Hourly Worker	0.653	0.0437	0.1238
Marital Status	0.588	0.0128	0.0362
Spanish Speaking	0.981	0.0040	0.0113
Primary scheme	Point Estimate	SE	Bias
Rent Own	0.740	0.0056	0.0161
Telephone	0.966	0.0013	0.0038
Computer	0.710	0.0022	0.0063
Military	0.116	0.0025	0.0071
Metro	0.836	0.0220	0.0624
Cable TV	0.539	0.0054	0.0153
Cell Phone	0.593	0.0057	0.0162
Hourly Worker	0.511	0.0171	0.0485
Marital Status	0.567	0.0029	0.0084
Spanish Speaking	0.983	0.0011	0.0032
New Coded Access scheme	Point Estimate	SE	Bias
Rent Own	0.718	0.0163	0.0461
Telephone	0.962	0.0055	0.0165
Computer	0.704	0.0637	0.1804
Military	0.112	0.0058	0.0165
Metro	0.827	0.0076	0.0217
Cable TV	0.523	0.0203	0.0576
Cell Phone	0.566	0.0366	0.1035
Hourly Worker	0.578	0.0224	0.0634
Marital Status	0.561	0.0146	0.0415
Spanish Speaking	0.981	0.0046	0.0131

Table D.2: Results By Scheme and Variable

Modified Coded Access scheme	Point Estimate	SE	Bias
Rent Own	0.750	0.0081	0.0231
Telephone	0.969	0.0032	0.0090
Computer	0.777	0.0414	0.1171
Military	0.119	0.0074	0.0209
Metro	0.834	0.0040	0.0114
Cable TV	0.549	0.0130	0.0367
Cell Phone	0.604	0.0256	0.0724
Hourly Worker	0.569	0.0201	0.0570
Marital Status	0.584	0.0093	0.0265
Spanish Speaking	0.983	0.0025	0.0071
Binary Primary scheme	Point Estimate	SE	Bias
Rent own	0.741	0.0117	0.0330
Telephone	0.966	0.0043	0.0121
Computer	0.711	0.0627	0.1773
Military	0.117	0.0067	0.0190
Metro	0.836	0.0035	0.0098
Cable TV	0.540	0.0154	0.0435
Cell phone	0.594	0.0282	0.0798
Hourly worker	0.550	0.0157	0.0444
Marital	0.567	0.0023	0.0064
Spanish speaking	0.983	0.0024	0.0066
New Coded Access and GMC scheme	Point Estimate	SE	Bias
rent own	0.725	0.0191	0.0541
telephone	0.961	0.0066	0.0187
computer	0.705	0.1019	0.2882
military	0.112	0.0064	0.0181
metro	0.820	0.0073	0.0205
cabletv	0.520	0.0212	0.0600
cellphone	0.561	0.0410	0.1159
hourly worker	0.579	0.0222	0.0629
marital	0.566	0.0185	0.0522
spanish speaking	0.975	0.0048	0.0136

Table D.3: Results By Scheme and Variable

New Coded Access, GMC, and CiH scheme	Point Estimate	SE	Bias
rent own	0.715	0.0153	0.0432
telephone	0.958	0.0054	0.0153
computer	0.578	0.0640	0.1809
military	0.114	0.0067	0.0190
metro	0.821	0.0057	0.0160
cabletv	0.516	0.0193	0.0545
cellphone	0.543	0.0353	0.0997
hourly worker	0.573	0.0225	0.0636
marital	0.553	0.0158	0.0447
spanish speaking	0.974	0.0032	0.0091
3-way raked Primary scheme	Point Estimate	SE	Bias
Rent Own	0.789	0.0057	0.0161
Telephone	0.958	0.0011	0.0031
Computer	0.578	0.0020	0.0059
Military	0.114	0.0025	0.0072
Metro	0.821	0.0218	0.0619
Cable Tv	0.516	0.0029	0.0084
Cell Phone	0.543	0.0046	0.0131
Hourly Worker	0.573	0.0103	0.0292
Marital Status	0.553	0.0039	0.0112
Spanish Speaking	0.974	0.0012	0.0036
Primary and GMC scheme	Point Estimate	SE	Bias
Rent Own	0.789	0.0112	0.0317
Telephone	0.978	0.0042	0.0120
Computer	0.899	0.0627	0.1775
Military	0.103	0.0066	0.0187
Metro	0.780	0.0035	0.0100
Cable Tv	0.583	0.0159	0.0450
Cell Phone	0.664	0.0287	0.0861
Hourly Worker	0.540	0.0157	0.0445
Marital Status	0.617	0.0141	0.0398
Spanish Speaking	0.991	0.0023	0.0067

Table D.4: Results By Scheme and Variable

No propensity 4-way rake scheme	Point Estimate	SE	Bias
Rent Own	0.745	0.0100	0.0284
Telephone	0.967	0.0038	0.0116
Computer	0.712	0.0618	0.1750
Military	0.116	0.0066	0.0188
Metro	0.838	0.0029	0.0083
Cable Tv	0.542	0.0136	0.0386
Cell Phone	0.597	0.0256	0.0725
Hourly Worker	0.541	0.0132	0.0374
Marital Status	0.57	0.0131	0.0371
Spanish Speaking	0.983	0.0022	0.0064
Base scheme	Point Estimate	SE	Bias
Rent own	0.779	0.0032	0.0091
Telephone	0.976	0.0009	0.0027
Computer	0.894	0.0025	0.0072
Military	0.119	0.0004	0.0207
Metro	0.841	0.0015	0.0042
Cabletv	0.582	0.0028	0.0080
Cellphone	0.652	0.0077	0.0219
Hourly worker	0.529	0.0093	0.0264
Marital Status	0.613	0.0029	0.0089
Spanish speaking	0.985	0.0016	0.0047
Binary two variable scheme	Point Estimates	SE	Bias
Rent own	0.741	0.0056	0.0160
Telephone	0.966	0.0013	0.0038
Computer	0.712	0.0022	0.0063
Military	0.116	0.0026	0.0074
Metro	0.836	0.0221	0.0627
Cabletv	0.541	0.0055	0.0155
Cellphone	0.595	0.0057	0.0161
Hourly worker	0.458	0.0169	0.0478
Marital Status	0.567	0.0030	0.0085
Spanish speaking	0.983	0.0011	0.0032

Table D.5: Results By Scheme and Variable

Bibliography

[1] A. Agresti. *An introduction to categorical data analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ, second edition, 2007.

[2] W. Bandilla, M. Bosnjak, and P. Altdorfer. Survey administration effects?: A comparison of web-based and traditional written self-administered surveys using the *issp* environment module. *Social Science Computer Review*, 21(2):235–243, 2003.

[3] J.M. Brick, I. Flores Cervantes, K. Wang, and T. Hankins. Evaluation of the use of data on interruptions in telephone service. *Proceedings of the American Statistical Association Section on Survey Research Methods*, pages 376–381, 1999.

[4] J.M. Brick, J. Waksberg, and S. Keeter. Using data on interruptions in telephone service as coverage adjustments. *Survey Methodology*, 22(2):185–197, 1996.

[5] US Census Bureau. Current population survey, computer use and school enrollment supplement: Technical documentation. Technical report, US Census Bureau.

[6] F. Cobben and J. Bethlehem. Adjusting undercoverage and non-response bias in telephone surveys. Technical report, Statistics Netherlands, 2005.

[7] W.G. Cochran. The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24(2):295–313, 1968.

[8] M.P. Couper. Review: Web Surveys: A Review of Issues and Approaches. *The Public Opinion Quarterly*, 64(4):464–494, 2000.

[9] M.P. Couper. Web Surveys: the Questionnaire Design Challenge. *Proceedings of the 53rd session of the ISI*, 2001.

[10] J.C. Day, A. Janus, J. Davis, et al. *Computer and Internet Use in the United States, 2003*. US Dept. of Commerce, Economics and Statistics Administration, US Census Bureau, 2005.

[11] K. Duncan and E.A. Stasny. Using propensity scores to control coverage bias in telephone surveys. *Survey Methodology*, 27(2):121–130, 2001.

[12] B.L. Fortson, J.R. Scotti, K.S. Del Ben, and Y.C. Chen. Reliability and validity of an internet traumatic stress survey with a college student sample. *Journal of Traumatic Stress*, 15:709–720, 2006.

[13] S. Fricker, M. Galesic, R. Tourangeau, and T. Yan. An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, 69(3):370–392, 2005.

[14] SD. Gosling, S. Vazire, S. Srivastava, and O. John. Should we trust web-based studies? *American Psychologist*, 59:93–104, 2004.

[15] Holly Gunn. Web-based surveys: Changing the survey process. *First Monday*, 7(12), 2002.

[16] Harris Interactive. Almost three-quarters of all US adults–an estimated 163 million–go online, 2005.

[17] D. L. Kostanich and C.S. Dippo. Current population survey: Design and methodology. Technical Report 63RV, U.S. Census Bureau and Bureau of Labor Statistics, 2002.

[18] R.J. Larsen and M.L. Marx. *An introduction to mathematical statistics and its applications*. Prentice-Hall Englewood Cliffs, NJ, 1981.

[19] S.L. Lohr. *Sampling: Design and Analysis*. Duxbury Press, 1999.

[20] M. Madden. *America's Online Pursuits: The Changing Picture of Who's Online and what They Do*. Pew Internet & American Life Project, 2003.

[21] K. Manfreda and V. Vehovar. Survey design features influencing response rates in web surveys. *Paper presented at International Conference on Improving Surveys, Copenhagen, Denmark.*, 2002.

[22] Nielsen. Country internet rankings. http://www.nielsen-netratings.com/, 2005.

[23] P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[24] M. Schonlau, R.D. Fricker, and M.N. Elliot. Conducting research surveys via email and web. *The Rand Corporation Reports*, 2002.

[25] D.M. Shannon, T.E. Johnson, S. Searcy, and A. Lott. Using electronic surveys: Advice from survey professionals. *Practical Assessment, Research & Evaluation*, 8 (1), 2002.

[26] N.J. Victory and K.B. Cooper. A nation online: How americans are expanding their use of the internet. Technical report, Washington DC: National Telecommunication Economics and Statistics and Information Administration, US Department of Commerce, 2002.

[27] H.C. Wang and H. Doong. Validation in internet survey research: Reviews and future suggestions. *Proceedings of the 40th Hawaii International Conference on System Sciences*, 2007.

[28] K. M. Wolter. *Introduction to variance estimation*. Springer Series in Statistics. Springer-Verlag, New York, 1985.

[29] T.S. Zimmerman. Evaluation of national CPS labor force variables. Technical report, Bureau of Labor Statistics.